LIZARDS AND LINEs: PHYLOGEOGRAPHY AND GENOME EVOLUTION OF THE GREEN ANOLE (*ANOLIS CAROLINENSIS*)

by

MARC ANTHONY TOLLIS

A dissertation submitted to the Graduate Faculty in Biology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2013

This manuscript has been read and accepted for the
Graduate Faculty in Biology in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Stéphane Boissinot

4/9/2013
Date

Stéphane Boissinot
Chair of Examining Committee

Laurel Eckhardt

4/9/2013
Date

Laurel Eckhardt
Executive Officer

Supervisory Committee:

Mike Hickerson

Jason Munshi-South

Susan Perkins

Anthony Furano

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Lizards and LINEs: Phylogeography and Genome Evolution of The Green Anole (*Anolis carolinensis*)

Dissertation by Marc Anthony Tollis

Advisor: Stéphane Boissinot

The sequencing of the green anole lizard (*Anolis carolinensis*) genome has already provided insights into how vertebrate genomes have evolved since the phylogenetic split between reptiles and mammals ~300 million years ago. For instance, the diversity and abundance transposable elements (TEs) in the *Anolis* genome, particularly the non-LTR retrotransposons (nLTR-LTs), shows more similarity to fish than to mammals, which suggests that mammals have significantly diverged from the amniote ancestor in terms of genome structure. The fate of TEs in a genome relies on the relative strengths of purifying selection against deleterious elements and genetic drift in host populations. Surprisingly, the geographic distribution and demographic history of populations within *A. carolinensis* has largely escaped scrutiny. We studied the patterns of mitochondrial and nuclear DNA sequences and found that there are five evolutionary lineages of green anoles, which diverged ~2 million years ago. Climatic shifts during the early Pleistocene may have driven their early diversification, particularly on ancient island refugia in what is now Florida, where a remarkable phylogeographic diversity of green anoles is found. More recently, the dispersal of green anoles onto the continental mainland led to a dramatic westward range expansion across the Gulf Coastal Plain. These insights into the evolutionary history of *A. carolinensis* allowed us to infer the population dynamics of nLTR-RTs in the *Anolis* genome. While nLTR-RTs are rare in *Anolis*, we find that they reach fixation in populations quite readily. We also find that full-length (FL) nLTR-RT insertions may be subjected to purifying selection as they are found at lower population frequencies than truncated (TR) insertions. This suggests deleterious effects of ectopic recombination or the process of retrotransposition are limiting the copy number of FL elements in *Anolis*. Finally, we find that FL elements are much more likely to be fixed in populations of small effective size, where purifying selection may not be acting as efficiently due to strong genetic drift. While FL elements are subjected to purifying selection, the fixation of TR elements suggests that another mechanism, such as DNA loss, may account for the relative paucity of nLTR-RTs in the *Anolis* genome.

ACKNOWLEDGMENTS

While there can only be one author of a dissertation, the list of individuals who helped complete this thesis is long. A great many people have inspired, challenged and supported me and I am forever thankful. Unfortunately, I cannot thank all of them in person. Throughout this process I have often asked myself, “Why am I doing this?” A degree is certainly a noble end. People who write dissertations have chosen a life of the mind. But I am still not sure what the true differences are between the people who do these kinds of things and those who do not. I have done it (you are reading it, after all), but most of the people from whom I inherited the wherewithal to get it done, either genetically or by example, have not done it. Yet the strength to finish a PhD did not simply arise endogenously. I have a robust family history of high-risk high-reward. No one struck it rich exactly (at least not yet), but I think it is safe to say that leaving a war-ravaged Europe to cross an ocean and start a family in America (as my grandparents did) is probably harder to do than grad school. And only as a parent now can I understand the sacrifice and uncertainty involved in straining to raise two kids who felt cared for and had the freedom to follow their interests in New York City in the 1970s and 80s (as my parents did). That requires a lot more guts than arranging a committee meeting. I can never be grateful enough for the risks those people have taken *for me*. In the grand scheme of things, they make grad school look easy. I think the common thread linking me to them is delayed gratification: a person should be able to withstand difficulty when there is a promise of better days ahead. I am proud to come from a line of folks who have that kind of integrity. Plus, like that famous child psychology experiment, we will all get marshmallows after it’s done.

Of course, it with great pleasure that I acknowledge Stéphane Boissinot, my thesis advisor, for providing the training I (badly) needed at the beginning and affording me the flexibility to move the project where it needed to go as things moved along. I came to grad school with a lot of knowledge but no way to apply it; Stéphane showed me how to use it to be a scientist. I am a relaxed person in general, yet grad school could break that down at times. One thing Stéphane has been good at (among so many others, he is really good in general) is consistently reminding me of the true nature of the process, from the field to the bench to the publication, and how simple it can be. He also has helped me become a better writer, and

not just in a technical way. I am lucky to have him as a mentor and a friend, and now as a colleague going forward.

To Liz: I read somewhere that people who are married are more likely to finish grad school. I am not going to bother to look for the reference (it's just the acknowledgements section), but I don't need one because I am living, breathing proof of that idea. The course of this degree has required a lot of heavy lifting from us both, and I could not have done it without her. She believed in me when I least believed in myself, which was probably at the very beginning, and whipped me into game shape when I had deadlines or other challenges, which became more complicated towards the end. Liz was with me and behind me the whole time. The remarkable thing was that she often had no idea what she was forcing me to go to work and do. Here's an acknowledgement that needs to be inscribed for the ages: it's hard being married to an early career scientist. Thank you, honey.

Our daughter Velma Jo was born during this degree process, giving me a renewed sense of purpose beyond the self. I think kids keep you on a clock and not a schedule. Schedules are flexible, movable, written in pencil. That's all well and good, but the problem is many grad students with schedules still take forever to graduate. In contrast, clocks tick forward whether you like it or not, reminding you that a certain someone is going to need to be dropped off somewhere, or picked up, or hugged by me before bedtime. I needed to do the PCRs, run the gels, and analyze the data before the clock ran out. This kind of pressure does amazing things for the productivity of a graduate student, perhaps more so than the fear of getting scooped. But most of all, Velma has just really grounded me with an appreciation for what is truly important outside of the ivory tower.

I must thank our parents: Sal and Jo for giving me life and believing in me, and Shirley and Alex for their support. And I cannot conclude this section without mentioning Phyllis and John, who let me wander in the woods. There is no doubt that the early years with them gave me the observation skills necessary to be an evolutionary biologist.

I also have many technical acknowledgements to make, including those for financial support. Funding for my research was provided by a Theodore Roosevelt Memorial Grant from the American Museum of Natural History and two Doctoral Student Research Competition Grants from the Graduate Center of the City University of New York. I also received funding from the Queens College Biology Department and the Seymour Fogel Endowment Fund for Research in Genomics. I would also like to thank the CUNY Graduate Center for awarding me their Dissertation Fellowship, the generous support of this award allowed me to sharply focus on completing this degree.

I would like to thank the following for their personal assistance in obtaining scientific collecting permits: Roger McCoy of the Tennessee Natural Heritage Program, Richard Kirk and Greg Wathen of the Tennessee Wildlife Resources Agency, Tereza Ancelet of the Department of Conservation and Natural Resources Wildlife and Freshwater Fisheries Division for the state of Alabama, and Stephen Bennett of the South Carolina Department of Natural Resources. I would also like to thank the North Carolina Wildlife Resources Commission, the Georgia Department of Natural Resources, the Arkansas Game and Fish Commission, and the Louisiana Department of Wildlife and Fisheries.

Special thanks are also in order to Xin Chen (College of Staten Island and Graduate Center, CUNY), Chaya Kazarnovsky-Shack, Drhuba Ghmire, and Sela Sherr (Queens College) for assistance in the field.

TABLE OF CONTENTS

LISTS OF TABLES

LISTS OF FIGURES

OVERVIEW: THE AGE OF REPTILES

The human genome was published in 2001, and although more genomes are being sequenced every year, the taxonomic distribution of fully sequenced vertebrates during the first part of the genomic era was heavily skewed towards mammals. For instance, by 2011 complete genomes had been released for two birds, four fish, one amphibian, and 33 placental mammals. Of course, this emphasis on the human side of the vertebrate family tree stems from applications to medicine and the desire to understand the origin of human genes. However, answering questions about the deeper origins of vertebrate genomes was always going to require a more robust sampling of non-mammals.

Within vertebrates, the fish and amphibians are restricted to aquatic or wet environments but amniotes, whose embryonic development occurs within a watertight egg, are well adapted to life on land. The Amniota is comprised of two great evolutionary lineages, which diverged more than 300 million years ago. On one branch are the synapsids, which includes mammals and their extinct relatives. On the other branch are the sauropsids, which includes the scaled reptiles and birds. Sauropsids are the sister taxon to mammals; thus, inferring ancestral states in the evolution of functional and non-functional aspects of the human genome requires extensive data from reptiles. The green anole lizard, *Anolis carolinensis*, was the first reptile to have its entire genome sequenced, and begins to fill this phylogenetic gap. Since I started my doctoral work, several other reptiles have been or are being sequenced, including a python, a turtle and three crocodilians. The first ten years of the genomic century belonged to mammals, but it is clear that we are now living in a genomic Age of Reptiles.

A major difference between mammalian and reptilian genomes is the diversity and abundance of the ubiquitous genomic parasites known as transposable elements (TEs). As I will cover in the chapters of this dissertation, TEs are DNA sequences that can move from one genomic region to another and can affect their hosts in neutral, deleterious, or adaptive ways. They are also the main driver of genome size differences between vertebrate groups. My dissertation uses the *Anolis* genome to understand not only the contrasting patterns of genome structure in reptiles versus mammals, but to also shed light on how population level forces explain the vastly different genomic profiles across evolutionary distant taxa.

CHAPTER 1: BACKGROUND

I will provide an overview of TEs, including their diversity, structure and mechanisms of transposition. I will then discuss in detail the TE profile of the *Anolis* genome and what it tells us about vertebrate genome evolution. Finally, the background chapter will conclude with a discussion of the evolutionary forces that control the fates of TEs in genomes: mainly, natural selection and genetic drift.

CHAPTER 2: POPULATION DELIMITATION IN *ANOLIS CAROLINENSIS*

The green anole is a laboratory model, and its genus has provided numerous insights into the processes that govern species diversification. Despite this, patterns of genetic variation and the geographic distribution of evolutionary lineages within the species were completely unknown. Chapter 2 focuses on the delimitation of populations and an understanding of their demographic history, a requirement of any investigation into the impact of natural selection or genetic drift.

CHAPTER 3: THE IMPACT OF ISLAND REFUGIA ON GREEN ANOLE INTRASPECIFIC DIVERSITY

The green anole is native to the southeastern United States, where climatic shifts during the Pleistocene left dramatic impacts on the geographic distribution of a wide range of taxa. Chapter 3 focuses in more detail on the biogeography of green anoles, particularly in Florida, where waxing and waning sea levels produced island refugia that fueled the early diversification of the major lineages within this species.

CHAPTER 4: EVOLUTIONARY DYNAMICS OF LINE-1 TRANSPOSABLE ELEMENTS IN *ANOLIS*

LINE-1 transposable elements are autonomously replicating retrotransposons that have dominated the human genome for nearly 40 million years. In contrast, the lizard genome contains far fewer LINE-1 elements; most of which have only recently been inserted. Chapter 4 builds upon what I learned about green anole population dynamics in the previous chapters to understand how natural selection and genetic drift interact to control LINE-1 copy number in the *Anolis* genome. This is the first investigation of its kind in reptiles.

Chapter One: Background and Literature Review

Part One: Introduction to Transposable Elements

The complete or ongoing sequencing of more than 1,000 eukaryotic genomes (www.genomesonline.org) has been an extraordinary source of information for scientists thereby revolutionizing the field of genetics, development and evolutionary biology. Eukaryote genomes vary considerably in size and structure and understanding the cause(s) of these differences is fundamental for interpreting meaningfully genomic annotations. Among the genomic features that show the most variation among organisms is the abundance and diversity of transposable elements (TEs). TEs are DNA sequences that can move from one location in the genome to another location. They have considerably affected the size and structure of eukaryotic genomes. In fact, with the exception of polyploidy, the abundance of TEs is *the* major determinant of genome size differences among eukaryotes. The abundance and diversity of TEs in a genome has important evolutionary implications as TEs constitute an important source of evolutionary novelties by providing a tool-box of sequence motifs on which natural selection can act.

The number and diversity of TEs in a genome result from the interactions between the rate of transposition, the intensity of selection against new inserts and the demographic history of populations. How these different factors interact remains controversial but the complete sequencing of a multitude of eukaryotic genomes as well as recent population studies have provided new insights on the evolutionary dynamics of TEs in eukaryotes. Understanding the dynamics of TEs is important for two main reasons. First, as TEs occupy a significant fraction of genomes, knowing the mechanisms that control their copy number will help understand why eukaryotic genomes differ so much in size, structure and function. Second, the evolutionary dynamics of TEs can help decipher the interplay between selective and neutral factors in the evolution of genomic features, a highly contentious issue in the field of comparative genomics [1].

Classification and Mechanisms of Transposition

TE is a generic term that covers an extraordinary diversity of mobile elements. TEs are usually classified into two groups, often referred to as class I and class II elements, based on their mode of transposition. Class I elements, also called retrotransposons, mobilize using an RNA intermediate during transposition and encode the enzyme reverse transcriptase. Class II elements, also called DNA transposons, do not have an RNA intermediate during transposition and use a DNA intermediate for transposition. Class I elements are further divided into two categories based on the presence or absence of long-terminal repeats (LTRs).

Retrotransposons lacking LTRs – This group of TEs includes two categories, the nLTR-retrotransposons (non-LTR retrotransposon) *sensu stricto* and the *Penelope* elements. *Penelope* elements constitute the most basal group in the evolution of retrotransposons. They are structurally very diverse, they sometime retain introns and their reverse transcriptase shows some similarity to telomerases [2]. Although they are widely distributed among eukaryotes, *Penelope* elements remain one of the least studied group of retrotransposons. non-LTR retrotransposons *sensu stricto* constitute an extremely old and diverse component of eukaryotic genomes. A number of very ancient monophyletic lineages of elements, called clades, have been described, from 11 to 25 depending on the authors [3, 4], and it is likely that additional clades will be discovered when more genome sequences become available. These clades can be sorted into six groups based on structural differences (figure 1): the *R2*, *RandI*, *L1, RTE, I* and *Jockey* groups [3].

All these elements encode at least one, but more often two open-reading frames (ORF1 and ORF2). ORF2 encodes for reverse transcriptase activity and, logically, it is the ORF all clades have in common. The most basal group in the evolution of non-LTR retrotransposons are the *R2* and *RandI* groups, which have a single ORF that contains a RLE endonuclease domain near the C-terminus, in addition to the reverse transcriptase. These elements tend to insert in a sequence specific manner, as exemplified by the *R2* element which inserts specifically in 28S rRNA genes [5]. All other clades encode an APE endonuclease located near the N-terminus of ORF2 and some of them have an RNase H motif

downstream of the reverse transcriptase domain. Most clades belonging to the *L1*, *I* and *Jockey* groups have another ORF, ORF1. ORF1 is poorly conserved among clades and, depending on the clade, contains esterase, CCHC zinc knuckles or RNA recognition motifs. The mammalian ORF1 protein (ORF1p) encoded by *L1* has been the most studied, yet its function remains unclear. It contains a conserved RNA recognition motif [6] and a rapidly evolving coiled coil domain [7] which mediates the formation of trimers [8]. The ORF1p participates in the formation of ribonucleoprotein particles [9] and encode nucleic acid chaperone activity [10]. A 5' untranslated region (UTR), that has been shown to act as an internal promoter in *L1* [11], can be found upstream of the ORFs. A second UTR of unknown function flanks ORF2 in 3'.

The transposition mechanism of non-LTR retrotransposons was first deciphered for the *R2* element in *Bombyx mori* [5] and subsequently the same mechanism was demonstrated for the human *L1* [12]. Following transcription, the retrotransposon mRNA is exported to the cytoplasm where it is translated. The translated proteins remain bound to the RNA and the resulting complex is then re-imported in the nucleus where insertion takes place. A nick is made on the bottom strand of insertion site by the endonuclease encoded by the element. The 3'OH released by this cleavage is then used to prime reverse transcription of the mRNA into a cDNA. Because the reverse transcription occurs at the sites of insertion, this reaction has been named Target-Primed Reverse Transcription (TPRT). The TPRT reaction lacks processivity, particularly in the *L1* clade, and up to two thirds of the new insertions are truncated in 5' [13]. Although the mechanism of transposition of other clades has not been studied in great details, the similarity in structure of insertions belonging to the *L1*, *L2*, *RTE* and *CR1* clades suggests that all these elements are mobilized by a mechanism similar or identical to the TPRT reaction [14].

The evolution of non-LTR retrotransposons is quite complex and seem affected by the nature of the interactions with the host. This is particularly true of the *L1* clade (Figure 2). In fish and squamate reptiles the *L1* clade is represented by a multitude of lineages, which diverged before the diversification of vertebrates [15-17]. These lineages are represented by very small copy number, but, within each family elements are very similar suggesting they inserted recently. It seems that in fish and reptiles *L1* elements

do not accumulate to large number and are possibly eliminated by purifying selection. This mode of evolution contrasts drastically with the situation in mammals where *L1* has accumulated to extremely large numbers. For instance the human genome contains 800,000 *L1* copies that account for 21% of its size [18]. As *L1s* are never excised, a host's genome contains a complete repertoire of the families that have been active in the past [19]. Phylogenetic analyses in humans and other mammals revealed that L1 retrotransposons evolved as a single lineage meaning that only one family of element is active at a time, until it is replaced by a most recent family [20-22]. This mode of evolution is extremely unusual and is reminiscent of the evolution of the *influenza* virus, suggesting it might be driven by repression by the host, a hypothesis supported by the observation that a region of *L1* is evolving adaptively [7, 22]. Interestingly, this single lineage mode of evolution is also observed in *Platypus anatinus* whose genome is not dominated by *L1* but by *L2* [23]. However, up until 40MY ago, multiple lineages of *L1* were concurrently active in primates [22]. It was found that coexisting families always had non-homologous promoter sequences raising the intriguing possibility that a competition for transcription factor encoded by the host might be limiting the diversification of *L1*. This hypothesis is supported by the fact that coexisting families in mouse [20] and lizard [17], also have non-homologous 5'UTR.

Although retrotransposition acts preferentially in c*is* [24], the replicative machinery encoded by non-LTR retrotransposons can also act on other transcripts and is responsible for the amplification of a number of non-autonomous TEs, called SINEs for Short INterspersed Elements, and processed pseudogenes [25]. For instance, the human *Alu* element, which is derived from the 7SL RNA, uses the L1 replicative machinery for its own benefit and amplified to considerable numbers in primate genomes (~1,000,000 copies in human). Other SINEs are derived from tRNAs and some of them show similarity at their 3' end with the autonomous elements that mobilize them as a way to recruit the biochemical machinery necessary for transposition [26]. *L1* is not the only clade to generate SINEs as elements mobilized by *RTE* and *L2* were recently discovered [27-29].

<u>*Retrotransposons with LTRs*</u> – This group includes three subgroups, the LTR-retrotransposons (LTR-RT) *sensu stricto* and the endogenous retroviruses (ERV) which have very similar structure and mode of

transposition, and the DIRS which differ considerably in structure [30, 31]. LTR-RTs are evolutionarily more recent than non-LTR retrotransposons and it is believed that they originated by recruitment of the reverse-transcriptase domain of a non-LTR retrotransposon by a DNA transposon [32]. They are classified into two main families, the metaviridaes, which includes two subgroups *Ty3*/*gypsy* and *Bel*, and the pseudoviridaes, including the *Ty1* and *copia* elements. LTR-RTs are widely distributed in eukaryotes, in particular fungi, insects and plants, where they constitute the dominant category of TEs.

LTR-RTs have a protein coding region which is flanked by direct LTRs that regulate transcription and play a critical role during reverse transcription. The protein coding region encodes two genes: *gag* which encodes structural and nucleic acid domains required for reverse transcription and *pol* which encodes the enzymatic activities protease, RNase H, reverse transcriptase and integrase. The mechanism of transposition begins with transcription of the element and export of the resulting mRNA to the cytoplasm. The mRNA is then translated and the resulting poly-protein is cleaved by the protease. The gag proteins form a virus-like particle, which typically contains two RNA molecules as well as the integrase, reverse transcriptase and RNase. The reverse transcriptase and RNase catalyze the reverse transcription of the RNA into a linear double strand cDNA, which is then re-imported inside the nucleus and inserted back in the genome by the integrase. Many LTR-RTs have independently acquired an additional gene, *env* (envelope), which, in the case of the Drosophila *gypsy* element, confers the ability to infect oocytes [33, 34]. There are strong reasons to believe that infectious vertebrate retroviruses evolved from metaviridae after recruitment of the *env* gene.

Eventually, some infectious retroviruses infected the germ line of vertebrates and became stable residents of these genomes [35]. Although they lost their infectivity they have retained their mobility and have multiplied in their host. There are three groups of ERV, called ERV Class I, II and III which are derived from different families of retrovirus [36], yet a large number of ERVs are still unclassified because they do not show similarity with any of the currently recognized groups of exogenous retroviruses. Vertebrate genomes often contain more than one type of ERVs. For instance the human genome

contains at least 26 distinct ERV families, representing the three known class of ERVs, and the number of independent acquisition of novel ERVs is probably close to 50 [36, 37].

The third group of LTR containing retrotransposon, DIRS, is the least studied [38], although it is quite widespread in nematodes, fish, amphibians, sea urchins, slime mold and fungi [39]. DIRS elements differ from other LTR-RTs in structure, as they lack a protease or an integrase. They encode a tyrosine recombinase suggesting that insertion into the host genome occurs by a recombination reaction catalyzed by the tyrosine recombinase.

It should be noted that the majority of LTR containing retrotransposons in some genomes are not complete and are represented only by LTRs. The loss of the coding region results from homologous recombination between the two LTRs, so that a single LTR remains, usually called solo-LTR. Some LTR-RTs have successfully multiplied in the absence of protein coding capacity, such as the *Dasheng* element of rice [40]. These non-autonomous elements have LTRs but no protein coding capacity. It is believed that the LTR of these elements can still be recognized by the retrotransposition machinery encoded by complete copies and are thus mobilized by their autonomous counterparts.

DNA transposons – DNA transposons (or class II) is a general group that includes three subclasses, cut-and-paste transposons, helitrons and polintons, that don't have much in common, except that they do not go through an RNA intermediate during transposition. The cut-and-paste transposons constitute a very diverse group found in all eukaryotic phyla [41]. Cut and paste transposons have a very simple structure, containing a single ORF encoding a transposase flanked by terminal inverted repeats (TIRs). The transposase recognizes the TIRs of the element, excises the transposon and inserts it elsewhere in the host genome. At the time of insertion, duplications of the target site are generated. The length and sequence of the target site duplication, terminal motifs in the TIR and similarity in the transposase domain are used to classify cut-and-paste transposons into 15 superfamilies [3], including the widespread *Tc1/mariner*, *MuDR/Foldback*, *hAT* and *piggyback* superfamilies. Most of these superfamilies are widely distributed across eukaryotes suggesting they were already diversified in the ancestor of all eukaryotes.

Although cut-and-paste transposons move through a non-replicative mechanism, they can still amplify in the genome of their host by two means: (1) if the transposition occurs during replication and if the transposon moved from an already replicated to a non-replicated chromatid; (2) when the element is excised, the repair machinery might use homologous recombination with a chromosome still containing the insertion to repair the gap [41].

The second subclass, called helitrons, transpose by rolling circle transposition, a mechanism of transposition found in some bacterial transposons [42, 43]. They encode a DNA helicase and a nuclease/ligase, they do not have TIRs and do not generate duplication of the target site. They have now been found in most eukaryotic lineages including plants, invertebrates, vertebrates and fungi. The third subclass, polinton (also called *Mavericks*), include some of the longest TEs [44, 45]. It was recently suggested that they evolved from a Mavirus virophage [46]. They encode 5 to 9 genes including a protein-primed polymerase B, an integrase, a cysteine protease and an ATPase. The polinton transposition mechanism is called self-synthesizing: the excised copy serves as a template for synthesis of a double strand DNA copy by the DNA polymerase which is then inserted in the genome by the integrase. Polintons are also widespread in fungi, vertebrates, invertebrates and protists. The three subclasses exist as autonomous families, i.e. families with protein-coding capacities, but also as non-autonomous families [41, 43, 44]. The non-autonomous families are often derived from autonomous elements that have suffered from internal deletions. In cut-and-paste transposons, these shorter copies still possess TIRs that are recognized by the transposase encoded by complete elements and therefore retained their mobility [47]. Non-autonomous families compete with their progenitors for the transposase and often outnumber greatly their autonomous relatives [41, 48, 49].

Mode of Transmission of TEs

TEs are components of the genome and as such they are transmitted vertically from parents to offspring. However, since the invasion of the *P* element into populations of *D. melanogaster* was discovered it has been known that TEs can also be transmitted horizontally among organisms. Vertical and horizontal transmission leaves drastically different signatures. First, the phylogeny of vertically transmitted TEs is

identical to the phylogeny of their hosts whereas horizontal transfer will produce conflicting phylogenies. Second, the sequence divergence between vertically transmitted elements in different species should be similar to the background neutral divergence between the host species; horizontally transferred TEs will be less divergent (as they were inserted after the species split from a common ancestor). Third, the presence of horizontally transferred TEs will be patchy within a group whereas vertically transmitted TEs should be present in all of the descendants of a common ancestor (minus the possibility of stochastic loss of family).

Numerous evidences indicate that non-LTR retrotransposons are transmitted mostly vertically and that horizontal transfer rarely occurs. Malik and Eickbush [4] showed that the level of divergence between non-LTR retrotransposons in distantly related organism was consistent with a strict vertical model of transmission. Several studies based on a large number of L1 elements have demonstrated that the phylogeny of *L1* in mammals and other deuterostomes recapitulates perfectly the phylogeny of the host, again supporting the vertical transmission of these TEs [50, 51]. However, there are a few cases of horizontal transfer of non-LTR retrotransposons. One of the best documented cases is found in vertebrates where an element belonging to the RTE clade has been horizontally transferred from a squamate genome to bovine genomes [52]. More recently, several instances of horizontal transfer of RTE in the opossum genome were demonstrated [29]. In their recent review of the topic Schaack et al. [53] cited 14 cases of horizontal transfer involving a non-LTR retrotransposons, comprising about 6% of all known instances of horizontal transfer. Interestingly, this is exactly the proportion of non-LTR retrotransposons believed to have been laterally transferred among *Drosophila* genomes [54]. Thus, non-LTR retrotransposons are the least likely TE to be horizontally transferred. There are several reasons why this might be the case, for instance: the mechanism of transposition of non-LTR retrotransposons, or the instability of the mRNA. Another possibility results from the fact that non-LTR retrotransposons are well adapted to their host and might be unable to successfully replicate in another host. It is interesting to note that more than half of the cases cited by Schaak et al. [53] involve elements lacking ORF1, which is a region suspected to be involved in host-*L1* interactions [7]. The main mode of transmission of LTR-RTs is

vertical, although horizontal transfer has been documented in plants and *Drosophila* where it is particularly frequent and accounts for more than half of the known cases of lateral transfer [53, 54].

Most cases of horizontal transfer have been documented in DNA transposons, particularly in cut-and-paste transposons but also in helitrons [55]. Horizontal transfer has been documented in eight out of the fifteen cut-and-paste super-families and seems particularly common in the *hAT* and *mariner* superfamily [53]. Most cases of horizontal transfer have been detected in animals, including insects, but more surprisingly in reptiles and mammals [56, 57]. It was believed for a long time that the sequestration of the germ line in tetrapods presented an insurmountable barrier to horizontal transfer. By now, multiple and independent instances of horizontal transfer have been documented. It seems that some species, such as the little brown bat, the tenrec and several squamate reptiles may be more prone to horizontal transfer than other [56, 58].

Although the exact mechanism of germline colonization is not yet known, the transmission of transposons seems to be mediated by parasites [59] or by viruses [55, 60]. Horizontal transmission seems to be an important feature of DNA transposons evolution and propagation. Although DNA transposons have become stable residents that are transmitted strictly vertically in some taxa, their amplification is sporadic. Their persistence in genomes over long periods of evolutionary time is not the rule, probably because of vertical inactivation. Thus without horizontal transfer, the diversity of DNA transposons in most genomes would be considerably reduced. Another consideration is that cut-and-paste transposons require only the transposase to be mobile and they have been shown to transpose in heterologous species.

Figures Chapter One Part One

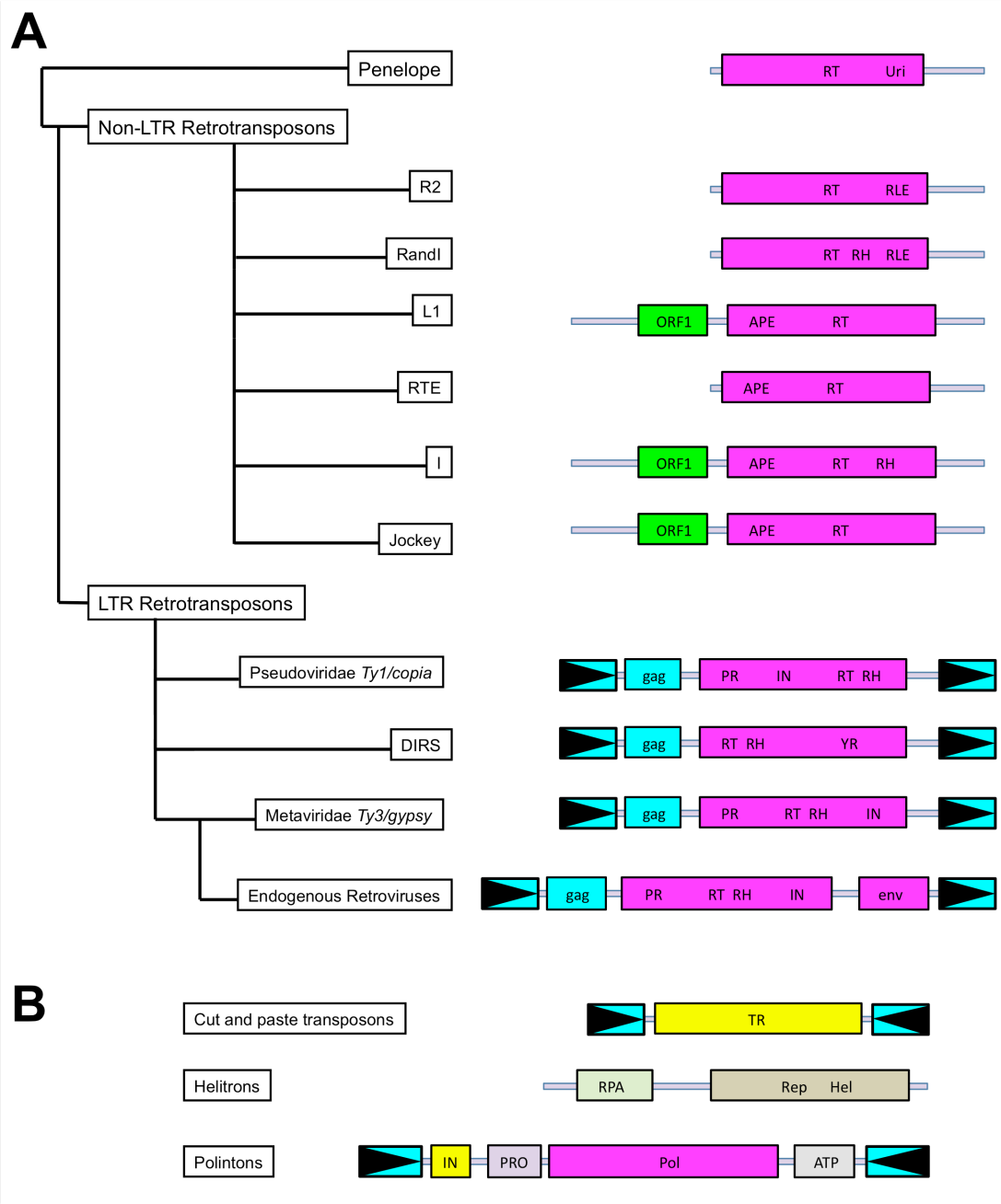

Fig. 1. (A) Schematic classification (left) and structure (right) of autonomous retrotransposons. The following abbreviations are used: APE, apurinic endonuclease; env, envelope gene; gag, gag gene; IN, integrase; ORF1, open- reading frame 1; PR, proteinase; RH, RNase H domain; RLE, restriction- like

endonuclease; RT, reverse transcriptase; Uri, endonuclease domain with similarity to group I introns; YR, tyrosine recombinase. The purple lines indicate the non- protein coding regions of the retrotransposons. The boxes represent the open- reading frames and the boxed triangles represent the LTRs. (B) Schematic structure of autonomous class II transposons. The following abbreviations are used: ATP, ATPase; Hel, helicase; IN, integrase; Pol, polymerase; PRO, cysteine protease; Rep, replication initiation domain; RPA, replication protein A; TR, transposase. The structure of the polintons can vary considerably from the type represented here.

Part Two: The Transposable Element Profile and Significance of the Anolis Genome

The original justification of the Human Genome Project was its obvious application to the genetics of medicine – to shed light on the origin and function of genes as well as the underlying variation that affects human disease [61, 62]. More recently, there has been an explosion of genomic data available for many other species [63-73], providing a robust taxonomic sampling of evolutionary groups of vertebrates. Investigators can access the NCBI and Ensembl databases and find 39 and 37 placental mammal genomes, respectively, in addition to two metatherians [74, 75]. Browsers such as the UCSC Genome Browser [76] (at http://genome.ucsc.edu) now contain the assemblies of five birds, two reptiles, an amphibian, eight teleost fish, and a cyclostome. The field of comparative genomics, concerned with the forces generating and maintaining biological variation, is thus today well equipped for addressing longstanding questions in evolution.

In 2007, the first draft of the green anole genome [77], *Anolis carolinensis*, was completed by the Broad Institute and made publicly available. *A. carolinensis* (Metazoa; Chordata; Vertebrata; Sauropsida; Lepidosauria; Squamata; Iguania; Polychrotidae) is a small lizard found in the Southeastern United States. It was an obvious candidate for a genome-sequencing project that would target a reptile. *Anolis* was "the" lizard represented in comparative vertebrate studies during the 20th century, and has been a laboratory model contributing to the fields of neuroscience, behavioral psychology and developmental biology. With over 400 species described in the New World, *Anolis* is the most speciose amniote genus and is arguably the premiere vertebrate model for the study of adaptive radiation and morphological evolution (reviewed in [78]). A remarkable variety of *Anolis* "ecomorphs" living on Caribbean islands has attracted the attention of evolutionary biologists for decades, and their research has provided insights into the mechanisms driving species diversification. Now available as a valuable genomic model, *Anolis* has much to offer our understanding of vertebrate genome evolution.

Among the genomic features that show the most variation among vertebrate taxa is the abundance and diversity of transposable elements (TEs). In fact, with the exception of autopolyploidy, the abundance of TEs is the major determinant of genome size differences among animals [79]. At one extreme,

mammalian genomes contain extremely large numbers of TEs, which can account for as much as 50% of their size [61, 64]. These genomes harbor a low diversity of mobile elements as they are dominated by the *L1* retrotransposon and its non-autonomous counterparts (SINEs). At the other end of the spectrum, teleost fish genomes are considerably smaller than mammalian genomes yet contain a comparatively more diverse TE profile, comprised of families usually represented by small numbers (<100) of very similar copies [15, 80-82].

The transition from the relatively small but diverse genomes of teleostean fish to the large, *L1*-dominated genomes of mammals is one of the most important, yet poorly understood, transitions in the evolution of vertebrates. Mammals and reptiles (or sauropsids, which include snakes, lizards, turtles, crocodilians and birds) represent the monophyletic group Amniota, which branched off from other amphibian-like tetrapods and radiated into terrestrial niches more than 300 million years ago (Figure 1). The tempo and mode of this history is well documented with transitional forms in the fossil record [83]. The genomic evolution associated with this history is less understood since we must infer ancestral states using sequence data from terminal taxa, and, from this point of view, *Anolis* helps fill an important phylogenetic gap.

A phylogenomic analysis of *L1* across deuterostomes completed before the release of the *Anolis* genome revealed that *L1* was most likely very diverse in the tetrapod and amniote ancestors, as it is found in extant amphibians. As amniotes diversified, there may have been complete stochastic loss of *L1* in certain lineages (for instance, all crocodilians and birds), with retention of *L1* diversity in some sauropsids such as lizards and snakes, along with a loss of diversity and explosion of copy number in the ancestor of mammals [50]. This analysis properly placed the evolution of a transposable element in the context of the evolution of its host across time. It is important to review more recent insights into the repetitive landscape of the *Anolis* genome, and discuss how they may impact our knowledge of vertebrate genome evolution.

The Transposable Element Profile of the Anolis Genome

The *Anolis* genome (and possibly squamate genomes in general, as will be discussed later) is characterized by an extraordinary diversity of mobile elements that far exceeds the transposon diversity found in other extant amniotes. Even more surprising is the fact that the majority of these families have recently been active, or is still active. Thus the anole genome is exposed to a tremendous level of TE activity, unparalleled among extant amniotes and possibly among vertebrates. A large number of active families of class I elements (i.e. elements that require an RNA intermediate for their mobilization or retrotransposons) were recovered from the anole genome [17, 27, 84]. *Anolis* retrotransposons are extremely diverse and representative of the two main categories of these elements – those with Long-Terminal Repeats (LTRs) and those that lack LTRs - were found to be active. Within the non-LTR retrotransposons alone, 43 distinct families, belonging to five clades (*L1*, *L2*, *RTE*, *CR1* and *R4*), show sign of recent activity [58]. For instance, in *Anolis* the *L1* clade is represented by 20 active families, the origins of which predate the split between reptiles and mammals. Each L1 family is represented by a small number (<100) of very similar copies, reminiscent of the diversity reported in the zebrafish genome [15, 85] but drastically more diverse than in modern mammals [22]. A recent study of the newly discovered *Ingi* clade of non-LTR retrotransposons found these elements in the *Anolis* genome, suggesting there may be more of these kinds of TEs to be discovered in this genome [86] . The biochemical machinery encoded by non-LTR retrotransposons can act on other transcripts and is responsible for the amplification of several SINE families. In particular, two SINEs have amplified to extremely large copy number [27]: Sauria SINE which is mobilized by an RTE retrotransposon (>200,000 copies) and Anolis SINE2 which is mobilized by a LINE-2 element (~140,000 copies). LTR-retrotransposons are also very diverse in the *Anolis* genome, as representatives of the four major groups of LTR elements (Metaviridae, Pseudoviridae, BEL/Pao and Retroviridae) have been identified [27, 84] and appear to be represented by multiple, highly conserved copies, suggestive of their recent activity [27, 84]. Interestingly, it was observed in a recent study of BEL/Pao that *Anolis* contains the highest copy number (397) of these elements across metazoans [87].

The *Anolis* genome also contains a large diversity of active class II transposons (i.e. elements that do not use an RNA-intermediate for their mobilization). At least seven autonomous families, representing 3 recognized super-families of DNA transposons (*hAT*, *Mariner*, and *Helitron*), show recent signs of activity in the *Anolis* genome [48], whereas three other types of transposons are either extinct (*Chapaev* super-family) or found in very small copy number (*PIF/Harbinger* and *Polinton/Maverick*). These families have generated a plethora of non-autonomous families that outnumber their autonomous progenitors by two orders of magnitude. The most diverse DNA transposons in *Anolis* are members of the *hAT* superfamily, which is represented by 5 autonomous and 32 non-autonomous families. This diversity results from the ability of autonomous and non-autonomous *hAT* elements to exchange genetic information. The exchange of sequence and domain, possibly by recombination, yielded a number of chimerical elements that have retained their mobility and yielded highly prolific families. For instance, multiple recombination events between non-autonomous *hobo* elements have generated at least 11 chimerical families that have amplified to several hundred copies.

Another interesting feature of DNA transposons is their ability to incorporate other transposable elements in their sequence [27, 48]. Some autonomous *hobo* elements carry up to five transposon fragments and mobilize these elements with their own sequence. This can significantly contribute to the increase in copy number of some elements. For instance, a Sauria SINE has amplified to very high copy number due to the amplification of an Helitron element it is embedded in [27].

Until recently, most sequenced amniote genomes were found to lack active DNA transposons and it was believed that this was due to the difficulty for horizontal transfer in amniote germ lines. However, the *Anolis* genome contains a number of families that show an extremely high level of similarity with elements found in distant species (e.g. bats, opossums, flatworm) that can only be explained by horizontal transmission. At least five families have been laterally transferred into *Anolis* and these families are responsible for the amplification of ~15,000 copies [56, 58]. Interestingly, the five families for which horizontal transfer was demonstrated have invaded the *Anolis* genome at different points in time, suggesting that horizontal transfer of DNA transposons is probably occurring much more frequently than

previously thought [58]. Although the exact mechanism used by transposons to invade the germ line remains unclear, it was proposed that the lateral transfer of transposons was facilitated by host-parasite interactions [59].

The Impact of Transposons on the Anolis Genome

The extreme TE diversity in *Anolis* does not translate into a much larger genome size than in other amniotes, although these genomes host a reduced number of active families. In fact, TE families in *Anolis* tend to be relatively young and ancient (i.e. divergent) families are very rare. Thus it seems that TEs are in some ways purged from the lizard genome. A first explanation comes from the analysis of the *L1* clade. *L1* copies are extremely similar to each other and divergent (i.e. ancient) *L1* copies are virtually absent from the anole genome [17]. This profile is similar to the one reported in fish and insects and is consistent with a model in which the insertion of new elements is counteracted by the removal of deleterious copies by purifying selection (which will be discussed in greater detail in Part Three of this chapter). This suggests that *L1* activity could have a high fitness cost in *Anolis*, possibly because of a higher rate of ectopic recombination [17].

However, the turn-over model provides only a partial explanation for the relatively young age of mobile elements because other TE families, in particular DNA transposons and shorter retrotransposons, do reach fixation and accumulate to large numbers. Another explanation comes from the examination of these somewhat older families, such as the *RTE-BovB* family that used to be prolific in *Anolis* but recently became extinct. When *RTE* copies are compared to each other, they exhibit a much larger number of deletions than elements of similar age in mammals, suggesting that the rate of DNA loss is much higher in *Anolis* [17], and possibly in other reptiles [88], than in mammals. Together with a low rate of fixation of insertions, a rapid decay of fixed elements in anole could explain the relative paucity of ancient elements in this genome.

Although the vast majority of TE insertions in *Anolis* are likely to be either deleterious or neutral, it is plausible that some of the abundant genetic variation introduced by transposons can be advantageous

and co-opted by the host. For instance, DNA transposons of the *PIF/Harbinger* super family show level of conservation consistent with their exaptation by the host [48]. Another observation suggesting the recruitment of TEs by the host for its own benefit comes from the observation that, in lizards, transposons tend to accumulate near developmental genes in general, and within *Hox* clusters in particular [89]. Typically, vertebrate *Hox* clusters are extremely compact and almost transposon free. In contrast, *Hox* clusters in anoles contain a large number of transposons. As transposons are restricted to the *Hox* clusters and are not as abundant in their flanking regions, it is likely that a large number of insertions were retained in *Hox* clusters because they provided an adaptive advantage. This raises the intriguing possibility that transposons have played a significant role in the diversification of a large genus such as *Anolis*, and possibly could explain the large diversity of morphologies found in squamate reptiles [90].

What Have We Learned from the Anolis Genome?

With a phylogenetic perspective, we can consider the pattern of genome evolution in vertebrates, including amniotes. Teleostean fish contain an astonishing diversity of both class I and II elements. The lone amphibian that has been fully sequenced is *Xenopus,* and it contains a similar diversity of both TE classes with high copy numbers that contribute to nearly a third of the genome. The repetitive landscape of the *Anolis* genome suggests that non-avian reptile genomes are more similar to fish and amphibians than to mammals and birds. Therefore, the most parsimonious model of genome evolution in vertebrates is one in which the ancestral amniote harbored a large diversity of class I and II elements, followed by substantial losses of transposon diversity in birds and eutherian mammals.

The completion of the *Anolis* genome emphasizes the diversity of the repetitive landscape among amniote genomes and indicates that amniote lineages have evolved drastically different strategies to cope with their intra-genomic parasites. Mammals have experienced a considerable reduction in transposon diversity yet they have much larger genomes than many other vertebrates due to the amplification of L1 retrotransposons. Population genetics and genomics studies have shown that the majority of L1 elements behave as neutral alleles and accumulate readily in the genome of their mammalian host. This does not mean that L1 activity is fully neutral. In humans, a fitness cost related to

the length of L1 elements has been demonstrated [91-93], yet it is insufficient to prevent the fixation of most elements, hence the extremely large number of copies in mammals. In contrast, the copy number of TE in *Anolis* appears to be much more regulated. Some families of elements fail to reach fixation and are probably subject to a high rate of turn-over due to a negative impact on their host genome. Like in mammals, longer elements seem to have a stronger deleterious effect than shorter ones suggesting that their negative effect could result from their ability to mediate deleterious chromosomal rearrangements through ectopic recombination [17, 48]. However, the high rate of turnover of these elements, compared to their fixation in mammals, suggest that retrotransposons impose a much heavier genetic load on *Anolis* than on humans. This raises the intriguing possibility that the rate of ectopic recombination is much lower in mammals than in *Anolis*. Such a difference in the rate of ectopic recombination was previously proposed to account for the different dynamics of amplification of TEs in mammals, insects and fish [16, 94].

A large number of transposons do reach fixation in *Anolis*, yet ancient TE families are disproportionally rare in this genome. This again is in sharp contrast with mammalian genomes that have accumulated extremely large number of TE copies since the origin of mammals so that these genomes are littered with ancient, long-extinct families. This difference could be due to differences in the regulation of DNA content among amniotes. Transposons decay faster in *Anolis* than in humans, suggesting that the rate of DNA loss could be far greater in reptiles than in mammals. If confirmed, this could be yet another drastic difference in the way some vertebrate genomes evolve.

The *Anolis* genome represents a single reptilian species, and one can predict that other reptilian genomes will bring further insights into the genomic evolution of vertebrates since the Mesozoic Era, as the divergence times of extant reptile lineages are on average much deeper than those of mammals [83]. *Anolis* is a member of the suborder Iguania and forms a monophyletic group with snakes (suborder Serpentes) called Toxicofera that arose almost 200 million years ago [95, 96]. A recent comparison of the repetitive landscapes of two snake genomes, python (*Python molurus bivittatus*) and copperhead (*Agkistrodon contortrix*), has revealed interesting patterns in squamate genome evolution [97]. While both

snakes contain a similar diversity of elements reminiscent of the situation in *Anolis*, it appears that TE copy number in the python is much more restricted than in the copperhead (comprising 21% of the genome versus 45%, respectively) even though both genomes are of similar size. The high copy number in the copperhead is probably mostly due to recent TE expansion, as evidenced by the low pairwise divergence between elements as well as the discovery of a large proportion of TE-related transcripts.

Combined with our overview of *Anolis*, the observations in these snakes are consistent with our phylogenomic hypothesis of a TE-diverse ancestral sauropsid genome. The differences in activity and copy number between toxicoferan squamates indicate that the dynamics of amplification of TEs may differ greatly within this group. Such differences could be due to variations in metabolic demands, demographic history or regulation of transposition by the host. Understanding the role of these different factors in determining the proliferation or loss of TE diversity and abundance within reptilian lineages will require further analyses, both at the molecular level and at the population level. Lessons learned through comparative analysis of more reptiles will prove to be essential elements in the still unfolding story of the vertebrate genome.

Figures for Chapter One Part Two

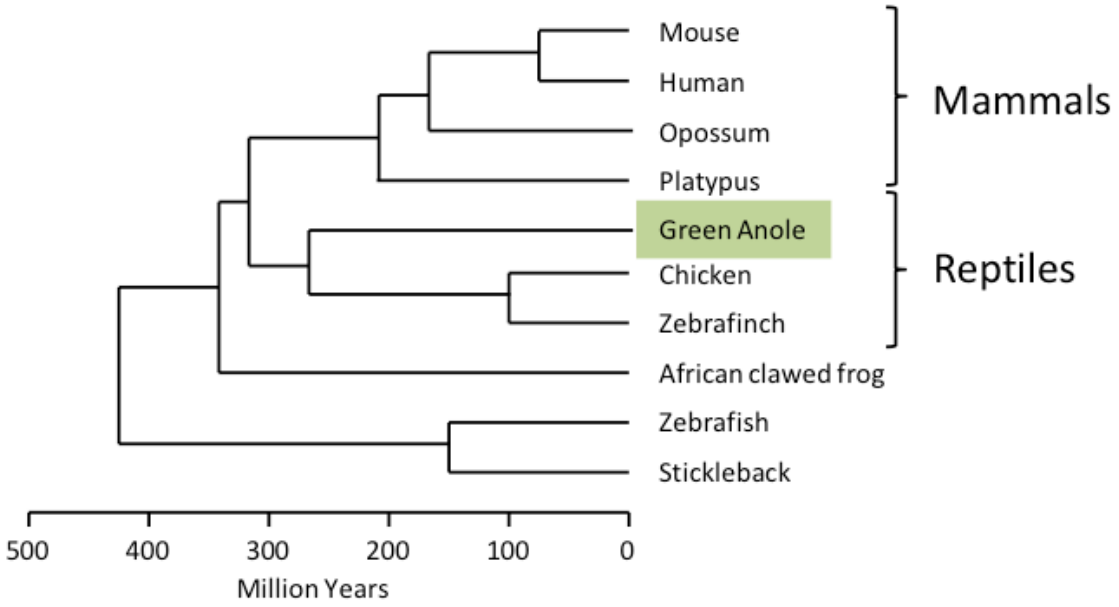

Figure 1. Phylogenetic relationships among vertebrates, highlighting the placement of the green anole.

Part Three: The Forces Controlling Transposable Element Evolutionary Dynamics

The Impact of Natural Selection on Transposable Elements

The number of TE copies in a genome is dependent on the rate at which new insertions are generated, i.e. the rate of transposition, and the rate at which insertion accumulate in the genome, i.e. the rate of fixation. The fixation (or loss) of a TE insertion is dependent on its effect on the fitness of the host. If the insertion is deleterious to the host, the insertion will have a lower chance of fixation and will most likely be eliminated from the gene pool by purifying selection. The deleterious effect TEs has been recognized since the 1970s when a phenomenon called hybrid dysgenesis was discovered in *D. melanogaster*. When females from strains of *D. melanogaster* lacking a DNA transposon called the *P* element are crossed with males that carry *P* elements, the resulting progeny is sterile and suffers from an increase in germ line transposition and an elevated mutation and recombination rate [98, 99]. Since the discovery of hybrid dysgenesis, numerous evidences suggesting a negative impact of TEs on host fitness have been described. For instance, the deleteriousness of TEs is apparent when comparing genomes that differ in TE copy number. Pasyukova et al. compared the fitness and egg hatchability among strains of *D. melanogaster* that differ by the number of TE copies. They found that, consistent with a negative impact of TEs, the strains with a larger number of insertions have a lower fitness [100].

In *Drosophila*, early surveys of insertion site polymorphisms in both natural populations and among strains revealed that the majority of TE insertions are at low frequency in populations and that fixed elements are rare [101-106]. This pattern suggests that in *Drosophila* TEs go through a rapid turnover of elements, in which the insertion of new elements is offset by the selective loss of element-containing loci [102, 107, 108]. In humans, population genetics studies have shown that the majority of *L1* elements behave as neutral alleles and accumulate readily in the genome of their host. This does not mean that *L1* activity is fully neutral. A fitness cost related to the length of *L1* elements has been demonstrated, yet it is insufficient to prevent the fixation of most elements, hence the extremely large number of copies in mammals [92, 93, 109]. The genomic distribution of TEs in *Drosophila* and human is also consistent with a deleterious effect of TE insertions. TEs tend to be more abundant in regions of low-recombination or no

recombination [92, 93, 110, 111] because low-recombining regions tend to accumulate deleterious mutations due to Hill-Robertson interactions [112] or because elements in those low recombining regions are less likely to be deleterious by mediating ectopic recombination events (see below).

Although it is widely accepted that TEs are indeed deleterious to their host, the basis for the deleterious effect of TEs has long been a matter of debate. Three non-exclusive deleterious effects of TEs have been proposed: (1) the direct effect of where elements insert (e.g., gene inactivation); (2) the effect of genetic rearrangements caused by ectopic (non-allelic) recombination between copies; (3) the effect of the transposition process *per se*. There is no doubt that TE insertions can indeed be deleterious when inserted into genes or even in introns, as suggested by the more than 60 disease-causing insertions reported in human (reviewed in [113]). It is also unquestionable that recombination between non-allelic TE insertions can create chromosomal rearrangements and large genomic deletions [114] which are very likely to be deleterious. Furthermore, the transposition process *per se* can be deleterious, for instance when the endonuclease of L1 retrotransposons makes double strand break in the host genome [115]. The question remains: which one of these mechanisms affects the dynamics of TEs in natural populations the most?

Although the issue is still debated, some recent population genetics results suggest an important role of ectopic recombination. As longer elements are more likely to be involved in ectopic recombination than shorter ones, they are more likely to be deleterious and eliminated by purifying selection. In addition, elements that belong to large families are more likely to find a partner for ectopic recombination than families with small copy numbers, thus larger families should be more deleterious than smaller ones. This is exactly what Petrov et al. [116, 117] tested using a population genetics approach. They found that, as predicted by the ectopic recombination model, selection against insertions was length-dependent and copy-number dependent. They found that long elements segregrated in *Drosophila* populations at lower frequency than short elements. They also found that many elements belonging to small copy number families reached fixation and accumulated in the *Drosophila* genome. Using the same approach, it was shown that the longest *L1* elements in the human genome were found at lower frequency in human

populations than the more truncated ones [109]. This result suggests that full-length elements are imposing a genetic load on their human host but that short, truncated elements behaved like neutral alleles, thus supporting the ectopic exchange model.

A second line of evidence in favor of the ectopic recombination model comes from the distribution of TEs across the genome. In *Drosophila* and in human, TEs accumulate in low recombining regions of the genome, possibly because insertions in low or non-recombining regions are less likely to be involved in non-allelic recombination. However, deleterious mutations are expected to accumulate in low-recombining regions of the genome whatever the nature of their deleterious effect is because of Hill-Robertson interactions. Simulation studies show that the accumulation of deleterious TEs caused by Hill-Robertson interactions would however occur only in populations of very small size and in regions of very low recombination [118]. These conditions are very restrictive and suggest that the ectopic exchange model provides a more general and likely explanation to the biased genomic distribution of TEs. In addition, a careful examination of the human genome revealed that the distribution bias of *L1* elements was also length-dependent. Short, truncated elements (<1.2Kb) are more homogenously distributed across the genome than elements longer than 1.2Kb, which tend to accumulate in low recombining regions [92, 93].

Although purifying selection is a powerful force limiting the spread of TEs, positive selection could also be acting on some insertions thus increasing their chance of fixation and their copy number. TEs can be an important source of evolutionary novelties, either by affecting the expression of host genes, by participating in regulatory networks, by incorporating into coding sequences or by creating new genes [119-122]. Some of these domestication events have had a very significant impact on the evolutionary fate of their hosts. For instance, a large number of regulatory sequences in vertebrates is derived from TEs and certainly had a profound impact on the evolution of this group [65, 123]. In *Drosophila*, some insertions have undoubtedly been positively selected as they confer resistance to insecticide [124]. However, one might wonder if positive selection in favor of TE insertions is strong enough and occurs often enough to affect the abundance and diversity of TEs in eukaryote genomes. A recent screen of

Drosophila populations outside of Africa identified at least 13 insertions that could have been under positive selection [125], although the actual number of adaptive insertions is probably between 25 and 50. This is a remarkably large number considering that *D. melanogaster* has left Africa only 10 to 16,000 years ago. Despite this high level of TE-related adaptation the *Drosophila* genome contains, as described earlier, mostly low-frequency insertions and the number of fixed insertions is relatively small. Thus adaptive insertions, as crucial as they might be as a potent source of evolutionary novelty, seem to be too rare to have a significant effect on copy number. In fact for positive selection to significantly increase copy number it would require a massive amount of favorable alleles or a general advantageous function provided by TEs. To our knowledge there are only two cases that would fit this model. Some elements in *Drosophila* have an important role in the maintenance of telomeres and in fact are the main constituents of *Drosophila* telomeres. *Drosophila* lacks telomerase and their telomeres are actually composed of three non-LTR retrotransposons named Het-A, TART and Tahre that specifically insert at the tips of chromosome and reach significant copy number because of their function at maintaining chromosome integrity [126]. Second, selection seems to be favoring the fixation of L1 elements on the eutherian X chromosome. It was proposed that *L1* can act as a booster element that would facilitate the inactivation of the X chromosome [127]. Indeed, the X chromosome of all mammals, except the opossum that lacks a Xist homolog [65], is enriched in L1 elements. It is possible that this accumulation of L1 on the X is due to a favorable effect of L1 insertion, related to a functional role of these elements in X inactivation.

The Impact of Host Demography and Life History on Transposable Elements

As TEs are obligatory parasites, their dynamics in the genome is affected by the evolution and natural history of their host. In particular, any factor that affects the effective population size (N_e) of the host will modify the equilibrium between drift and selection. When N_e is large, selection dominates over drift but any factor that decreases N_e (*e.g.* bottleneck, mating system) will strengthen drift. Thus one would expect that in populations with large N_e, selection would be more efficient at removing deleterious TE insertions than in small populations in which the rate of fixation will be higher. This can have far-reaching consequences in terms of genomic evolution because the long-term accumulation of insertions will lead to an increase in genome size, whereas one expects species with large population size to have smaller

genomes. This hypothesis was supported by a study by Lynch and Conery [1], who compared the effective population size and the genome size of a wide range of organisms. They found that species with very large population size tend to have smaller genomes than species with small population size, thus supporting an important role of non-adaptive demographic factors in shaping genome size evolution. However, a more recent analysis limited to plants [128] failed to find such a correlation and more comparative analysis will probably be required to settle this issue.

An effect of drift on the rate of fixation of TEs has also been examined using a population genetics approach. In *Drosophila subobscura* [129] and in *Arabidopsis lyrata* [130], TE insertions are found at higher frequency in bottlenecked populations than in populations that have retained a large long-term population size, an observation consistent with a reduced efficacy of purifying selection. In fact, the frequency distribution of TEs in strongly bottlenecked *A. lyrata* populations is consistent with neutrality [130]. Similarly, it was shown that the strength of selection against TE insertions was strongly reduced in populations of *D. melanogaster* that emigrated out of Africa [131]. Thus, populations subjected to strong genetic drift tend to accumulate TE insertions whereas large population will eliminate them.

Another factor that affects N_e is the mating system. For instance, inbred species face a reduced N_e relative to outbred species; consequently, the efficacy of selection should also be reduced in these taxa. In addition, inbreeding will result in an increase in homozygosity. As an element is more likely to be involved in ectopic recombination in the heterozygous state, selection against insertions will be more efficient in outcrossing species than in inbred ones. Thus, it is predicted that inbred populations will carry more fixed TEs than outcrossing ones. This was tested by comparing the selfing worm species *Caenorhabditis elegans* with its outcrossing relative *C. remanei* [132] and the self-fertilizing *A. thaliana* with the outcrossing *A. lyrata* [133, 134]. In both comparisons TE insertions were at higher frequency in the selfing species consistent with a reduced efficiency of selection. Mating system also affected the patterns of genomic distribution. Unlike what was found in *Drosophila* and human, TEs are not more abundant in low recombining regions in self-fertilizing *A. thaliana* [135] and *C. elegans* [136] suggesting a lack of genome-wide purifying selection. This is not to say that TEs can be considered purely neutral. The

density of TEs is lower in or near host genes suggesting a deleterious effect of TE insertions related to disruption of normal gene function.

Chapter Two: Multi-locus Phylogeographic and Population Genetic Analysis of *Anolis carolinensis*: Historical Demography of a Genomic Model Species

Anolis carolinensis, or the green anole lizard, is the first lepidosaurian reptile to have its entire genome sequenced [77]. Since its publication in 2011, the *Anolis* genome has already provided insights into our understanding of how vertebrates have diversified since the split between reptiles and mammals more than 300 million years ago (Mya) [137]. For instance, the sequence of events during the evolution of vertebrate axial skeleton segmentation as revealed by the *Anolis* genome [138] suggests that cyclically expressed patterns in the "segmentation clock" have not necessarily undergone a stepwise pattern from fish to mammals. The relative paucity of isochores in the *Anolis* genome [76] suggests that variation in GC composition is less integral to genomic structure than previously thought. In addition to these recent contributions to the field of comparative genomics, anoles as a group have been the focus of investigators in ecology and evolution for decades. The repeated convergent evolution of *Anolis* "ecomorphs" on the Greater Antilles has brought much attention to the mechanisms governing adaptive radiations (reviewed extensively in [78]). Thus, the *Anolis* genome will shed light on the genetic basis of not only morphological and ecological adaptation but the process of speciation itself [139].

While it belongs to a genus containing as many as 400 species across the tropical and semi-tropical New World, *A. carolinensis* is the only anole native to North America. Green anoles are widespread and abundant throughout the southeastern United States [140], occurring naturally in Florida (FL), Georgia (GA), North Carolina (NC), South Carolina (SC), Tennessee (TN), Alabama (AL), Mississippi (MS), Louisiana (LA), Arkansas (AR), Oklahoma (OK) and Texas (TX). Although this species has long served as a laboratory model in research fields such as physiology, neurobiology, and behavior [141], very little is known about the patterns of genetic diversity in natural populations. Molecular phylogenetic analyses have shown that *A. carolinensis* is nested within the paraphyletic *A. porcatus*, or Cuban green anole [142]; thus the ancestors of *A. carolinensis* are believed to have colonized North America via overwater dispersal from Cuba prior to the Pleistocene, where it occurs in the fossil record [143]. Morphological [144, 145] and genetic [146] analyses have concluded that there are significant differences between

geographically distinct populations, including major differences between the mainland and peninsular FL. In fact, Vance (1991) proposed the subspecies status of southern FL populations (*A. c. seminolus*) on the basis of dewlap coloration and number of lamellae.

The historical processes that account for the current distribution of *A. carolinensis* remain unresolved. Eastern North America is geologically and topographically complex, and a number of common phylogeographic patterns have been reported across a wide range of co-distributed taxa [147]. Genetic discontinuities observed in both animals and plants, and notably in other squamates [148-151] include "latitudinal shifts" that suggest southern refugia during glacial maxima followed by northward expansion, and sharp genetic breaks associated with mountain ranges and rivers that may have acted as common barriers to dispersal.

To test hypotheses regarding historical demography, the field of phylogeography has moved towards an emphasis on incorporating as many loci as practically possible. This is because stochastic differences between the coalescent histories of gene genealogies [152] have demonstrated that a genome-wide sampling of genetic variation will better capture the signature of demographic events such as population divergences, migration, and population size changes [153]. The availability of a complete genome sequence has allowed us to test demographic hypotheses using 10 non-coding nuclear loci designed with the *Anolis* genome database and one mitochondrial locus. We have characterized and delimited evolutionary distinct lineages within this species, estimated key demographic parameters associated with its history on the continent, and tested hypotheses postulated by other comparative phylogeographic studies of southeastern North America. The results of our analyses will have important implications for future studies of *A. carolinensis*, most notably shedding light on the effects that population structure may have on the maintenance of genetic variation in a genomic model organism.

Methods

Ethics Statement

This study was carried out in accordance with the recommendations of the American Veterinary Medical Association (AVMA) for the euthanasia of ectotherms and every step was taken to avoid needless suffering. The following protocol was approved by the Queens College Institutional Animal Care and Use Committee (IACUC) (Animal Welfare Assurance Number: A32721-01; protocol number: 135) and was administered by the authors. After capture, animals were kept in the dark in fabric bags for a maximum of four hours and were sacrificed the same day. Euthanasia was carried out in the field via intra-abdominal injection of sodium pentobarbital at a dosage of 100mg/kg of body weight. The euthanasia protocol approved by the University of Texas at Arlington IACUC (Animal Welfare Assurance Number: A09.012; protocol number A11.003) involved prolonged CO_2 exposure in a closed container followed by deep freeze. The IACUC at the American Museum of Natural History does not provide welfare assurance numbers or protocol numbers, however the approved protocol ensured that animals were euthanized humanely according to methods suggested by the AVMA as well as the American Society of Ichthyologists and Herpetologists via (1) intracoelemic injection of Tricaine Methanesulfonate (MS222) with a sodium bicarbonate buffer, and (2) once the animal was confirmed as inert, a second injection of unbuffered MS222.

Sample collection

We collected 159 anoles in NC, SC, GA, FL, AL, TN, LA and AR during 2009 through 2011 and obtained their tail or liver tissues. We also obtained the tissues of nine Texan anoles from Andre Pires da Silva of the University of Texas at Arlington, and 31 blood samples from individuals collected at additional sites in SC, GA and FL given to us by Bryan Falk of the American Museum of Natural History. A map in Figure 1 shows the geographic distribution of all samples. The GPS coordinates of each collecting locality are included in Table 1. DNA was extracted from tissues via proteinase K digestion followed by purification with the Promega Wizard Genomic DNA Purification standard protocol, except for the blood samples for which we slightly modified the protocol with a smaller final elution volume to compensate for slightly lower yield.

Marker design, amplification, sequence editing and alignment

We designed nuclear sequence loci (nDNA) *in silico* using the UCSC Genome Browser (http://genome.ucsc.edu/) [154]. We first searched the *Anolis* genome database for introns of reasonable size (<2Kb) for PCR amplification. An intron was chosen for further analysis if its presence was predicted by at least two of the tracks available for visualization on the browser (including NCBI RefSeq, Ensembl, non-lizard mRNAs from GenBank, and lizard ESTs). Although we selected introns for their length without respect to gene function, we recorded the putative function of each gene prediction (see Table 2). Primers were designed in the surrounding exons for Exon Primed Intron Crossing (EPIC) PCR [155] with the Primer3 program [156]. All primer pairs were used in a virtual PCR of the *Anolis* genome for specificity and single-copy confirmation. Anonymous nuclear loci were also designed using the UCSC Genome Browser. We first scanned chromosomes for gene-poor regions in order to avoid selective sweeps or background selection. Our chosen regions were used in a BLAST search [157] to screen for unannotated genes. Gene-free sequences were submitted to RepeatMasker [158] to search for repetitive elements, and a BLAT [159] of the *Anolis* genome was used for repeat detection and confirmation of single-copy status. Single copy regions lacking transposable elements, short tandem repeats, and/or single sequence repeats ranging from 200bp to 750bp were chosen for PCR amplification, followed by primer design using the program Primer3 and subsequent virtual PCR. All primers for PCR products used in this study are listed in Table 3. We amplified a mitochondrial region containing the nicotinamide adenine dinucleotide dehydrogenase subunit 2 (ND2) gene and downstream tryptophan and alanine tRNA genes with primers suggested by Jason Kolbe of Harvard University in 190 anoles. For each nDNA locus we amplified from a smaller geographically representative sample, resulting in 62 to 152 sequences per locus (see Table 4). PCR conditions were as follows: an initial hold for one minute at 94˚C followed by 30 cycles of 30-second denaturing at 94˚C, 30-second to 1 minute annealing at 54˚-61˚C depending on the melting temperatures of the primer pairs, and one minute extension at 72˚C, with a five minute hold at 72˚C before refrigeration at 4˚C. All PCR products were sent to the High-Throughput Genomics Unit at the University of Washington in Seattle, WA for purification and sequencing in both forward and reverse orientations.

After Sanger sequencing, we imported chromatograms into CLC Main Workbench version 5 and Geneious v5.5 [160]. Regions of poor quality at the ends of reads were trimmed and double peaks were called using the Secondary Peak Calling option (CLC) or Find Heterozygotes plugin (Geneious) using a threshold of 50% peak height. For each sample, we assembled forward and reverse reads into contigs using a reference sequence: the virtual PCR product from the *Anolis* genome database for each nDNA locus and the ND2 region of mitochondrial sequence NC_010972 from GenBank. Putative heterozygous sites in nDNA sequence reads were assessed based on quality score. Less than 5% of all reads were unusable due to heterozygous indels, and were removed from further analyses. Each contig was edited manually and the consensus sequences were extracted and aligned using ClustalW [161] as implemented in BioEdit [162], where alignments were further edited by eye. The gametic phase of each nDNA haplotype was resolved computationally using the program PHASE 2.1 [163], with a cut off of 90% probability; phase estimation was repeated twice to assure consistent and reliable haplotype reconstruction, and the haplotypes with the highest probabilities were selected for analysis. As most phylogeographic analyses include the assumption of a lack of recombination in the data set being used, we submitted the phased haplotypes for all loci to the four-gamete test [164] as implemented in DnaSP v5 [165]. Recombination-free sequence blocks were created for the data sets in which recombination was detected by the program IMgc [166], thus rendering these loci suitable for downstream analysis. A concatenated nDNA dataset was also generated for pairwise alignments and demographic inference using SequenceMatrix [167].

Phylogeographic analysis

The dataset used for phylogenetic inference included 226 ND2 sequences: 190 amplified by us; the homologous region available from the *Anolis* genome sequence (NC_010972); 30 *A. carolinensis* sequences available in GenBank (accession numbers EU106323 – EU106342); and three *A. porcatus* (AY654092 - AY654094), one *A. isolepis* (AY654022) and one *A. altitudinalis* (AY654023) to be used as outgroups. Phylogenies were reconstructed using the rapid bootstrap (RBS) Maximum Likelihood (ML) method in RAxML [168], a full ML analysis using MEGA 5.0 [169], and Bayesian Inference (BI) with BEAST version 1.6 [170]. For the RBS ML analysis, the sequence evolution model used was GTRCAT.

Bayesian Information Criterion as implemented in MEGA indicated the most likely model of sequence evolution for this sample was HKY + Gamma + Invariant sites, and we used this model for the full ML and BI analyses, with the number of gamma categories set to 4. For the full ML analysis node support was assessed with 1000 bootstrap replicates. In order to infer the timing of diversification events in the history of *A. carolinensis* with the phylogenetic information, we first conducted a preliminary BEAST analysis (25,000,000 runs, uncalibrated with an estimated mutation rate) and calculated the average Tamura-Nei and uncorrected pairwise genetic distances between recovered clades. Assuming a mutation rate of 1.3% per million years (Myr) for this region, as observed in other small lizards [171] and used previously to date diversification events in the *Anolis* genus [142, 172, 173], we estimated the divergence time of all preliminary mtDNA clades. We used this information to calibrate the final BI tree, placing a normal prior distribution encompassing the estimated time to recent common ancestor (t_{mrca}) of all *A. carolinensis* populations. We further calibrated our tree with knowledge from previous molecular phylogenetic analyses [142], which estimated the divergence of the *carolinensis* anole subgroup (for our purposes, the node shared by all *A. carolinensis* and *A. porcatus*) at 6Myr. The final BEAST analysis included all 226 sequences, two independent runs of MCMC length 100,000,000 with the evolutionary model and calibrations as stated above, a strict molecular clock, the known mutation rate, and a coalescent tree prior. Estimates of the posterior distributions of all parameters for each run were monitored with Tracer v1.4 [174] in order to assess convergence across separate runs, and once confirmed, separate runs were combined using LogCombiner included in the BEAST package.

We used several methods that allow us to delimit populations and infer their evolutionary history with the nDNA data. First, unrooted ML genealogies were constructed for each alignment in MEGA assuming the Tamura-Nei 1993 evolutionary model, and the topology of each genealogy was assessed in regard to congruence with each other and the mtDNA phylogeny. Multi-locus haplotypic data were entered into STRUCTURAMA 2.0 [175], which implements a Bayesian clustering algorithm to estimate the number of populations (a random variable K) using a Dirichlet prior and assigns individuals to each inferred population. We ran four independent chains of 10,000,000 generations. Haplotypes were also entered into the program *Structure 2.3.3* [176, 177], a similar clustering method that estimates the likelihoods of a

range of user-set values of K. *Structure* analyses were run with 100,000 steps for burn-in followed by 100,000 generations for K values ranging from 2 to 13. Each simulation was completed five times, and results files were compressed and submitted to *Harvester* [178], which selects the most likely K value based on the delta-K criterion described by Evanno [179]. *Structure* has been shown to overestimate K (see the program documentation at http://pritch.bsd.uchicago.edu/structure.html); however, it allows the user to implement a model that includes admixture, which is a common feature of population genetic data sets. *Structure* provides estimates of the proportion of each individual's genome derived from one of the K clusters. This differs from the STRUCTURAMA model we used, which estimates the posterior probability that each individual is a member of one of the K clusters using a no-admixture model. Therefore, similarities and differences between STRUCTURAMA and *Structure* results were interpreted such that STRUCTURAMA would recover genetic patterns on a larger geographic scale while *Structure* would be more sensitive to localized violations of Hardy-Weinberg equilibrium and therefore indicative of finer-scale population structure.

Recently, the utility of Bayesian clustering methods as the sole source of evidence for determining population structure has come under scrutiny [180, 181]. Therefore, as an additional assessment of population structure, we calculated pairwise F_{ST} between 22 collecting localities of sample size four or greater from the concatenated dataset in Arlequin v 3.5 [182] and the resulting distance matrix was used to construct a neighbor-joining tree in MEGA. We also calculated pairwise F_{ST} between STRUCTURAMA-inferred populations. In order to assess the degree to which the mtDNA and nDNA datasets can each explain the total genetic variation, we used Analysis of Molecular Variance (AMOVA) in Arlequin to partition groups of nDNA sequences in two ways: (1) assignment to their mtDNA clade from the phylogenetic analysis and (2) assignment to their STRUCTURAMA-inferred population. Specifically, we were most interested in the percentage of total genetic variation that is explained by differences between groups. Similar partitioning among these types of groups would suggest that both datasets recover the same phylogeographic patterns.

Analysis of genetic diversity

We calculated standard diversity statistics for each locus in DnaSP including: number of polymorphic sites (s), number of haplotypes, haplotype diversity (Hd), nucleotide diversity (π), and average number of pairwise differences per sequence (k). Summary statistics were also calculated in Arlequin for: (1) the 22 collecting localities from which we obtained four or more individuals; (2) each inferred major mtDNA clade; and (3) each STRUCTURAMA-inferred population. We also measured the mean corrected Tamura-Nei distances within and between each mtDNA clade. The uncorrected pair-wise genetic distances per locus and the concatenated nDNA data set were measured within and between each STRUCTURAMA-inferred population. All corrected and uncorrected genetic distances were calculated in MEGA.

Historical demography

In order to test for evidence of population expansion in the history of *A. carolinensis*, we calculated Tajima's D [183] and Fu's Fs [184] with 1000 permutations using Arlequin for each STRUCTURAMA-inferred population and major mtDNA clade and using DnaSP for each locus. For each population we also investigated the mismatch distribution of pairwise genetic differences in the concatenated nDNA using Arlequin, comparing the observed distributions to a unimodal expectation under a model of recent population expansion. The fit of the data to an expansion model is determined by a non-significant value of the raggedness index (R, [185]). Bayesian Skyline Plots [186] (BSP), which utilize the coalescent properties of gene trees to plot population size changes over time, were constructed for the mtDNA clades using BEAST. For each BSP, prior distributions for the root height of the population were notified by initial estimates from the 1.3%/My mutation rate. To incorporate stochastic differences between gene genealogies in the estimation of population parameters, as well as obtain posterior probabilities for the number of population size change events, we constructed multi-locus Extended Bayesian Skyline Plots (EBSP) [187] in BEAST for mtDNA clades. The EBSPs are informed by the known mutation rate used for ND2, and include an inheritance scalar to take into account the smaller effective population size of mtDNA versus nDNA. For BSPs and EBSPs, the lengths of the MCMC chains were set to achieve

effective sampling sizes (ESS) of >200 in order to avoid autocorrelation of parameter sampling and assure proper mixing.

In cases of evidence for population expansion, we tested for directionality using a series of linear regressions with nucleotide diversity as calculated with the concatenated nDNA data from collecting localities for which we collected four or more individuals. First, π was plotted against latitude, in which a negative relationship could be used as evidence for southern refugia [188]. To test for east-west expansion, we plotted π against longitude. Recently expanded populations are not expected to show isolation by distance (IBD), as not enough time will have passed for genetic drift to differentiate geographically separated subpopulations [189]. Therefore, in the widest ranging populations, we tested for IBD by implementing the Mantel Test [190], testing the correlation between the pairwise F_{ST} distance matrix and a geographic distance matrix generated in DIVA-GIS from the GPS points of each collecting locality. Significance was determined by 10,000 randomized permutations, and was used to accept or reject a null hypothesis of no correlation between geographic and genetic distances.

Results

Overview of genetic data

An overview of the genetic data is featured in Table 4. Excluding outgroups and sequences from GenBank, our mtDNA data set is comprised of 191 sequences of total length 1172bp, with 215 segregating sites. Not surprisingly, the mtDNA locus exhibited higher values of diversity statistics (k, π, number of haplotypes, and Hd) than the nDNA. The number of individuals sequenced varies among nuclear loci, ranging from 62 to 74 sequences for the four introns and from 72 to 152 for the six anonymous loci (table 1). Intronic sequences ranged in length from 688bp to 1288bp and anonymous loci ranged from 211bp to 557bp. Sequences have been deposited in GenBank (accession numbers JQ857105 - JQ858187). Almost all haplotypes were reconstructed with 100% accuracy, and the very few which were estimated at <90% had no effect when removed from downstream analyses. For the nDNA loci, π ranged from 0.00098 to 0.00572, the number of haplotypes ranged from 4 to 21, and Hd ranged from 0.255 to 0.861. Recombination was detected in three of the 10 nDNA loci (HMGCS, C1 and Gav5),

but never involved more than two events per locus, while the number of recombinant haplotypes per data set was limited to 2 or 3 individuals.

Phylogeographic analysis

All phylogenetic analyses yielded highly concordant topologies. The most likely tree from the RBS ML analysis of the mtDNA is shown in Figure 2, with posterior probabilities (pp) and bootstrap (bs) values shown above and below important nodes, respectively. The monophyly of *A. carolinensis* is strongly supported (1.0 bs and pp). As in the preliminary BI analysis, the final BI analysis recovered four major mtDNA clades with 100% pp. These clades are strongly correlated with geographic region (Figure 4A) and consist of (1) a lineage endemic to the Gulf coast region of FL in or around the Suwannee River drainage system (the "Suwannee" clade); (2) a group limited to anoles from the southern tip of the FL peninsula (the "Everglades" clade); (3) a NC clade and (4) a large clade including samples from all other localities ranging from the Atlantic coast of northern FL across the Gulf Coastal Plain to TX (the "Gulf-Atlantic" clade). Relationships within the Gulf-Atlantic clade could not be well resolved, with individuals from disparate geographic regions often clustering together with extremely low posterior support. One interesting well-supported minor mtDNA clade within the major Gulf-Atlantic clade consists of individuals collected from various localities on the western side of the Smoky Mountains in eastern TN. An unexpected result was the occurrence of two ND2 sequences from FL (one collected by us and one from GenBank) clustering just outside the NC clade to form a monphyletic group, and we address this below and in the Discussion section.

Given the mutation rate for the ND2 gene and the calibrations used, we estimated the age of the root of the BI tree to be 9.7 Myr (95% HPD 6.2-13.3), while the split between *A. porcatus* and *A. carolinensis* was estimated at 6.2 Myr (95% HPD 4.3-8.2). These estimates are close to estimates from past molecular phylogenetic studies [142], which propose very similar divergence dates for the *carolinensis* group and subgroup. The t_{mrca} for all of our samples was estimated to be 2.1 Myr (95% HPD 1.3-2.9), pointing to a Late Pliocene-early Pleistocene origin for *A. carolinensis.*

We found some discordance between the mtDNA and nDNA phylogenies, due to lower variation in the nDNA and resulting multifurcations. Even so, the unrooted genealogies were useful in visualizing the concordance that did exist across most trees, showing bifurcations between FL populations and all others (Figure 3). STRUCTURAMA estimated K=4 populations with 91% probability (2% K=3; 7% K=5). We named these populations Suwannee, Everglades, Gulf-Atlantic, and Carolinas, as they are largely congruent with the geographic distribution of the major mtDNA clades (Figure 4B), except for a few differences. First, STRUCTURAMA detected more gene flow between localities along the Atlantic seaboard: some individuals in the Gulf-Atlantic mtDNA clade were assigned to the STRUCTURAMA-inferred Carolinas population, which ranges from NC into coastal SC and GA. In addition, one SC and one GA collecting locality, both inland, contained individuals assigned to both the Carolinas and the Gulf-Atlantic STRUCTURAMA-inferred populations. The second difference consists of one individual collected in FL that clusters with NC anoles in the mtDNA phylogeny but was assigned to the Suwannee population in the STRUCTURAMA analysis. *Structure* simulations consistently encompassed the four STRUCTURAMA-inferred clusters. Using the delta-K method, *Structure* estimated a larger number of populations (K=10; Delta K = 16.41) than STRUCTURAMA, the main difference being that *Structure* detected a greater degree of clustering between GA, AL, TN, LA, TX and AR in the Gulf-Atlantic STRUCTURAMA group.

Despite these minor differences, the AMOVA partitioned the same amount of variation in the nDNA genetic data when grouped by mtDNA clade and by STRUCTURAMA-inferred group (35%, see Table 5). Therefore, both mtDNA and nDNA recover very similar geographic patterns. In both hierarchical AMOVA setups, the least amount of genetic variation existed between subpopulations within groups. This is expected if relatively little gene flow is occurring. F_{ST} values between STRUCTURAMA-inferred populations are all significant (Table 6), suggesting strong population structure with limited gene flow between adjacent clusters. The greatest differentiation exists between the Everglades and all other populations. The NJ tree derived from the pairwise F_{ST} distance matrix (Figure 5) recovered a pattern that is largely consistent with the cluster and phylogenetic analyses, including long branch lengths separating subpopulations from NC and SC, Southern FL, and those whose members are in the Gulf-Atlantic

STRUCTURAMA-inferred group and mtDNA clade. The subpopulations whose members were assigned to the Suwannee clade/population fall in slightly disparate regions of the NJ tree, close the mid-point and with short branch lengths.

Genetic diversity and historical demography

Summary statistics that were calculated per locus for each population are featured in Table 7; statistics calculated for each collecting locality with the nDNA are featured in Table 8. Averaged across all loci, haplotype diversity is lowest in the Gulf-Atlantic population and highest in the FL populations. Diversity statistics for mtDNA clades and STRUCTURAMA-inferred populations (calculated from the concatenated nDNA dataset) are shown in Table 9. Nucleotide diversity is highest in the Suwannee for both data sets, and lowest in the Gulf-Atlantic and NC for the mtDNA and in the Everglades for the nDNA. Average p-distance is greatest within the Suwannee for the mtDNA (Figure 6A) and averaged across nine nuclear loci (Figure 6B). The greatest p-distance on average from all other populations is highest in the Suwannee for mtDNA and in the Everglades for nDNA; both datasets show the closest genetic distance exists between NC and the Gulf/Atlantic (Table 10).

Tajima's D and Fu's Fs were negative for most loci (Table 7), suggesting violations of neutrality due to population size expansion. Both D and Fs were also negative in all inferred populations (Table 9). The frequencies of pair-wise differences within each population (Figure 7) are consistent with what is expected under a model of population expansion: raggedness indices (R) derived from these mismatch distributions were all non-significant (*p*-values given in Figure 6). These three indicators (D, Fs, and R) all suggest that population expansion has occurred.

The BSPs indicate that the Suwannee and Everglades clades both experienced expansions ~500-700Kya (Figure 8A and B, respectively), while the Gulf-Atlantic clade experienced a more dramatic and recent expansion ~250Kya (Figure 8C). In contrast, the NC and TN populations have remained relatively stable during the last 150-200Ky (Figure 8D and E, respectively). Multi-locus EBSPs from the Suwannee and Everglades show highest posterior probabilities for single past population size expansions. The timing of

these expansions is consistent with the single-locus BSPs (lower-bound estimates for both are between ~0.5-1Mya) (Figure 9). Effective population size estimates from all skyline plots are given in Table 11.

A strong pattern of genetic diversity loss in northern localities was not recovered, as the regression of π plotted against latitude show no significant relationship (r^2 = 0.15; p-value = 0.08). The relationship between π and longitude was positive and significant (r^2 = 0.28; p = 0.01). When we removed the localities consisting of individuals from the Suwannee and Everglades lineages the relationship between π and longitude was greater (r^2 = 0.32; p = 0.01). Therefore, a stronger signal exists for an east-to-west direction of expansion. With the Mantel Test, we could not reject a null hypothesis of no correlation between geographic and genetic distances in the wide-ranging Gulf-Atlantic population (r = 0.18; p = 0.06). Thus, evidence for a model of IBD is lacking in this wide-ranging population, suggesting that the east-west expansion was rapid and relatively recent.

Discussion

How many populations?

The mtDNA phylogeny and STRUCTURAMA analysis both indicate that there are at least four distinct green anole populations. We identify these populations as (1) the Everglades in southern FL, (2) the Suwannee on the Gulf coast of FL, (3) the Gulf-Atlantic and (4) the NC clade (Carolinas for nDNA). The differences in population assignment using mtDNA and nDNA from localities in NC, SC and GA may arise in one of two ways: (1) male-biased dispersal leading to an introgression of NC nuclear haplotypes in the south and vice-versa, or (2) the retention of ancestral polymorphisms in more slowly evolving nuclear genes. Males in the *Anolis* genus are known to disperse a few hundred meters from their place of hatching [78], which is an appropriate distance considering the adjacent geography of these populations. However, more explicit modeling of this process is needed, in which the likelihoods of competing demographic hypotheses with and without migration parameters are compared. The sole anole we collected near the central Atlantic coast of FL may represent a poorly sampled mitochondrial clade endemic to that region or, more simply, an introgressed individual with a mitochondrial haplotype derived via human-mediated dispersal. More sampling is needed to address this issue.

The topology of the NJ tree based on F_{ST} supports three of the mtDNA and STRUCTURAMA-inferred groups, with subpopulations from the putative Suwannee population occurring in disparate regions of the tree. It is important to keep in mind that this is an unrooted tree derived from the genetic distances of populations and therefore not necessarily useful for inferring evolutionary relationships. However, that these subpopulations are not separated by very long branch lengths from the midpoint of the tree may be used as further evidence for the ancestral status of the Suwannee populations (see next section). As for the *Structure* analysis, the delta-K method points to a larger K (10) than inferred by STRUCTURAMA (4). This may be because demographic histories involving population size expansions will produce an excess of low frequency polymorphisms; thus the program will use evidence from slightly divergent haplotypes to add additional clusters with minimal cost, in the process overestimating K. Still, the *Structure* results do encompass all four STRUCTURAMA groups. That STRUCTURAMA agrees so closely with the mtDNA phylogeny, and that genetic variation partitions so closely in the AMOVAs, allows us to say with confidence that four populations most accurately reflect the distribution of individuals (in our sample) across this geographic scale.

Historical demography of green anoles

It has previously been proposed that *A. carolinensis* colonized North America via overwater dispersal to FL from Cuba near the time of Plio-Pleistocene boundary [142]. Our results are consistent with this hypothesis, as the t_{mrca} of our mtDNA dataset is estimated at ~2Mya. Three lines of evidence suggest that populations on the mid- and northern Gulf coast of FL represent the most persistent remnants of this colonization event: (1) the Suwannee clade is the most basal *A. carolinensis* clade in the mtDNA phylogeny; (2) genetic diversity overall is highest in this population, suggesting long-term stability; and (3) Suwannee subpopulations tend to be closer to the midpoint of the F_{ST} NJ tree. There is a possibility that the current distribution of genetic variation represents refugial populations that were once more widespread. The genetic signatures of these populations could have been wiped out by rising sea levels and an insular history of FL during interglacial periods from the Miocene into the Pleistocene. Thus, we

also cannot rule out the extinction of ancestral populations in southern FL followed by the more recent colonization of derived populations.

Once *A. carolinensis* entered the continental mainland, there were a number of geographic and topographical constraints that may have affected its dispersal patterns and thus the current distribution of individuals. Numerous studies of terrestrial fauna have described a common phylogeographic pattern showing genetic discontinuity between populations living along rivers that drain into the Atlantic Ocean and those which drain into the Gulf of Mexico [191]. Our mtDNA and nDNA datasets show that many subpopulations along the Atlantic Seaboard – particularly those south of the Charleston Harbor watershed – cluster closely to Gulf Coastal Plain populations. The oft-observed Gulf coast/Atlantic seaboard dichotomy is not absolute in *A. carolinensis*, as there appears to be extensive gene flow along the SC and GA portions where these regions overlap.

Another important discontinuity existing on either side of the Appalachian Mountains has been observed in co-distributed taxa such as salamanders [192, 193], snakes [150, 151] and turtles [194]. This break is often extended southward below the extent of the mountain chain, on either side of the Appalachicola River, which bisects the FL panhandle as it flows into the Gulf of Mexico. This pattern holds for our data but only in FL, since the Suwannee population is divergent from other Gulf coast subpopulations west of the Appalachicola (6.5% Tamura-Nei distance between them in mtDNA versus 2.4% overall). However, both mtDNA and nDNA haplotypes easily cross the hypothesized Appalachicola barrier further north, in GA and AL. More sampling around the Appalachicola river basin is needed to test its effect as an adequate barrier to dispersal, as well as the precise geographic location of genetic breaks. Unique Carolina haplotypes in both mtDNA and nDNA datasets are not found on the other side of the Appalachians in TN, despite a relatively small geographic distance. It appears that the Appalachians have acted as a barrier to dispersal further north while there also existed some form of barrier between the mainland and the Gulf coast of FL, although these need not be related phenomena.

Whether or not Pleistocene glacial cycles have had region-wide effects on co-distributed taxa has long been a subject of debate in comparative phylogeographic studies focusing on eastern North America. Southern refugia during glacial maxima followed by subsequent northward dispersal and population size expansion has been cited as an explanation for observed genetic diversity across numerous studies [147]. Our data are not consistent with this hypothesis in three ways: (1) the inferred population expansions were estimated to have pre-dated the most recent Pleistocene glacial cycles, (2) skyline plots of northern subpopulations in NC and TN show evidence of stability, and (3) there is a lack of a significant negative correlation between nucleotide diversity and latitude. While the current geographic distribution of green anole populations may reflect ancient refugia during earlier Pleistocene glacial cycles, our data suggest that more recent advances of the Laurentide Ice Sheet (~100,000 to 10,000 years ago) have left little or no genetic signature on these populations, most notably those found at higher latitudes.

In addition to the inferred effects (or lack thereof) of late Pleistocene glaciation, the geography of genetic discontinuities in green anole populations differs from what is found in some co-distributed taxa in important ways. For instance, many of the riverine barriers that have strongly affected the cryptic fragmentation in the co-distributed common ground skink (*Scincella lateralis*) [148] have not done the same for green anoles; neither does the often observed phylogeographic break [147] at the Mississippi River hold for this species. Given the history of overwater dispersal in the *Anolis* genus [142], this is not a surprising observation, and it suggests that undetermined factors (including those inherent to the natural history of each species) have played a role in the geographic distribution of individuals in this region.

A statistically significant relationship between nucleotide diversity and longitude suggest that our data are most consistent with a hypothesis of westward expansion of *A. carolinensis* populations across the Gulf Coastal Plain during the mid- to late Pleistocene, with a minimal effect of glacial maxima during this period. Four lines of evidence suggest that anoles dispersed quite rapidly across the region: (1) mtDNA haplotypes essentially form a polytomy in the phylogenetic analysis; (2) Tajima's D, Fu's Fs and the mismatch distribution point to expansion; (3) a dramatic increase in population size inferred by the BSP, and (4) the lack of support for a model of isolation by distance.

We can attend to the assertion of Vance (1991) that certain southern FL populations constitute the subspecies *A. c. seminolus*, based on our analyses of the genetic data. While we have shown ample evidence for an independently evolving lineage at the southernmost reaches of the FL peninsula, we have detected an equally or even more divergent lineage endemic to the northern Gulf coast of FL. These lineages have been separate since the early to mid-Pleistocene, with minimal migration. When taken into account with previous morphological analyses that have concluded significant geographical differences, our genetic data point to significant polytypy in this species, although perhaps not as drastic as observed in ground skinks [148].

We have taken advantage of the resource provided by the *A. carolinensis* genome and devised a multi-locus study of the demographic history of this species in North America. Combining phylogenetics, clustering methods, and population genetics, we have demonstrated the existence of at least four distinct evolutionary lineages of *A. carolinensis*, the most recent common ancestor of which may predate the Pleistocene. These lineages have been characterized by population expansions since the mid-Pleistocene, and do not seem to have been affected by more recent glacial periods at mid-latitudes. The most striking phylogeographic breaks separate subpopulations from the mainland Atlantic Seaboard, the Gulf Coastal Plain, and the Gulf coast and southern portions of FL. We propose that these genetic discontinuities may result from a combination of barriers provided by the Appalachian Mountains and dispersal patterns along waterways that drain towards either the Gulf of Mexico or the Atlantic Ocean. We have also inferred a rapid mid- to late Pleistocene westward expansion of *A. carolinensis* populations across the Gulf Coastal Plain, and into habitats that have expanded the traditional niche of their more tropical evolutionary progenitors. Surely, the resources provided by the sequencing of the *Anolis* genome will provide ample opportunity to investigate the adaptive differences that exist at the molecular level for a species that ranges across such a wide variety of habitats.

Figures Chapter Two

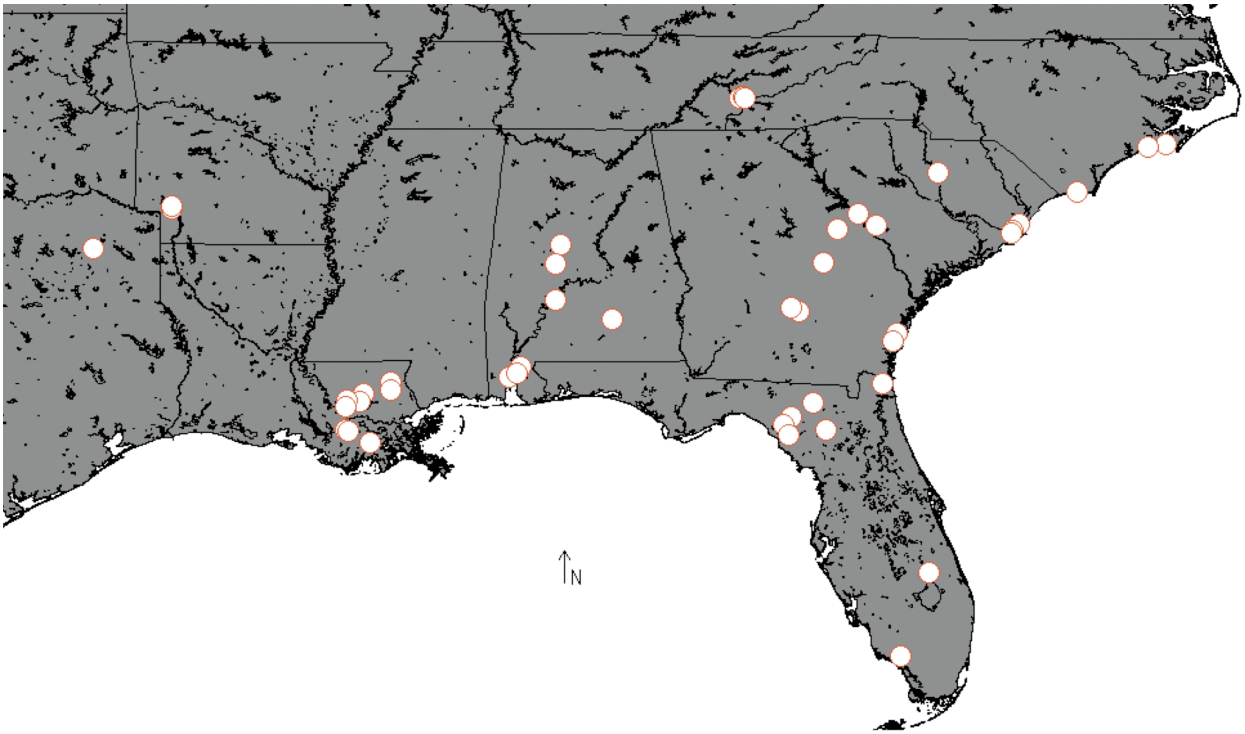

Figure 1. Sampling localities.

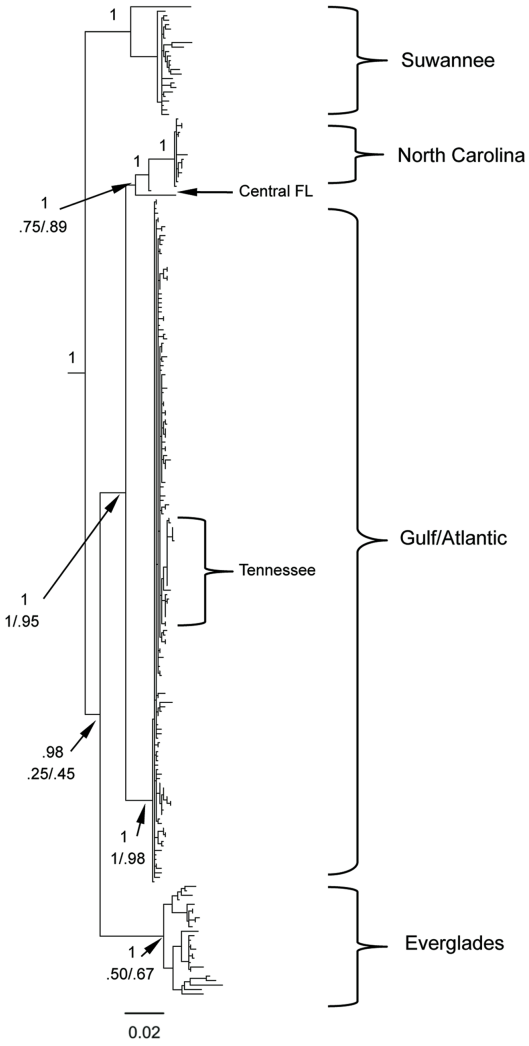

Figure 2. Phylogenetic reconstruction of the ND2 mitochondrial region. The most likely tree derived from the rapid bootstrap method (RBS) in RaxML is shown; the topologies of three methods (RBS, full likelihood, and Bayesian inference) were highly concordant. Posterior probabilities are given above important nodes; below nodes are the bootstrap values from RBS before the slash and from the full ML analysis after the slash. Nodes with 100% support in all three analyses are indicated with a 1. Outgroups (*A. isolepis*, *A. altitudinalis* and *A. porcatus*) are not shown. The four major clades (Suwannee, North Carolina, Gulf-Atlantic, and Everglades) are indicated with large brackets on the right. The Tennessee minor clade is shown, nested within the Gulf-Atlantic lineage. Two samples from central FL that clustered closer to the North Carolina clade are indicated with a black arrow.

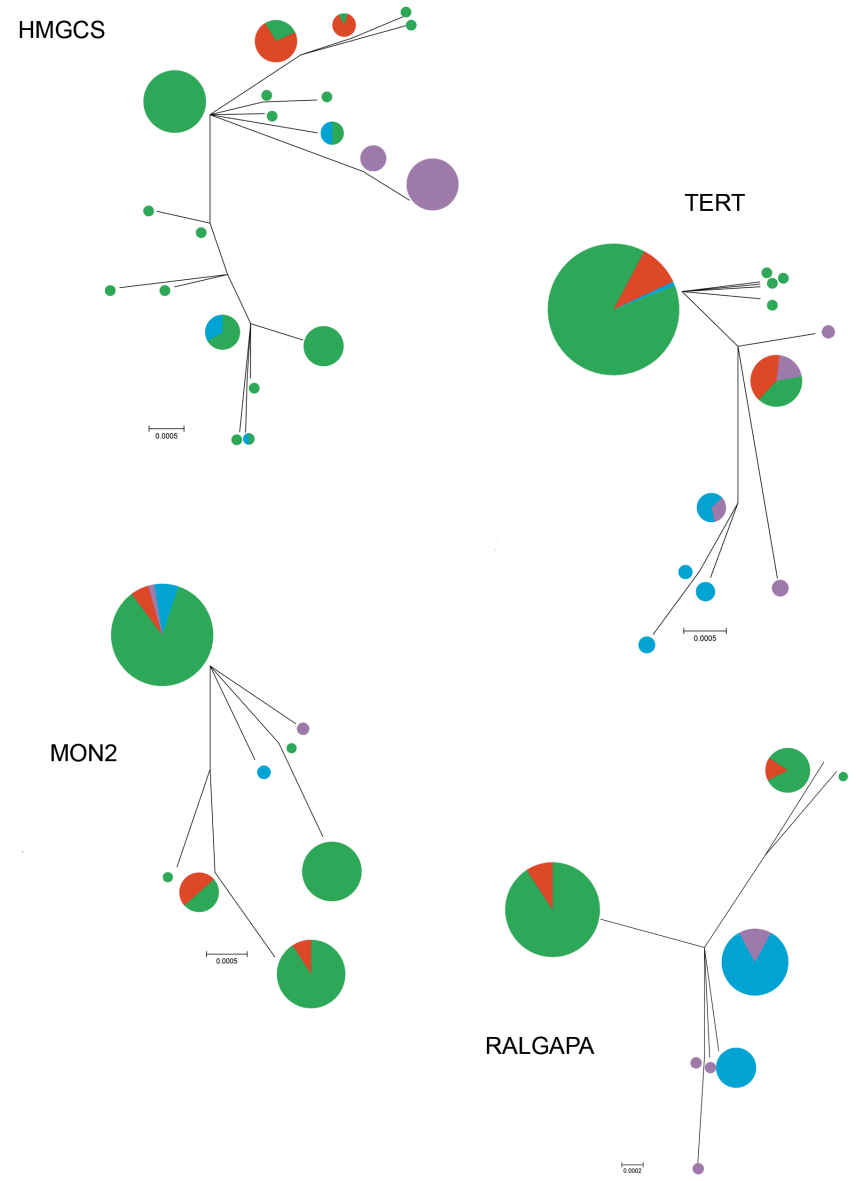

Figure 3: Unrooted ML trees for intronic sequences. We performed phylogenetic inference using Maximum Likelihood (ML) in MEGA 5.0 with 1000 bootstrap replicates (bootstrap values not shown). Trees are unrooted due to lack of an outgroup. Circles are roughly proportional to the number of individuals present at a node, and pie charts reflect proportion of individuals at each node belonging to one of four major mitochondrial clades: Gulf-Atlantic (green), North Carolina (red), Suwannee (blue), Everlgades (magenta).

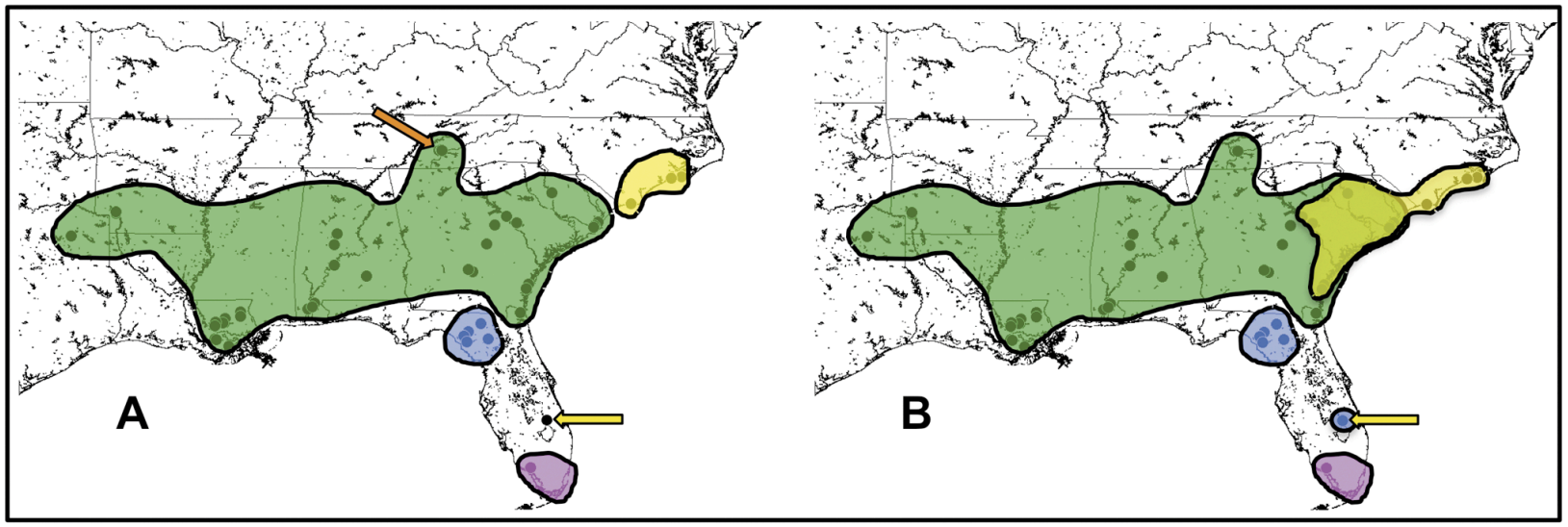

Figure 4. Geographic distribution of genetic populations. Colored shapes indicate the extent and boundaries of each inferred population. 3A shows the distribution of the four major mitochondrial clades: NC (yellow), Gulf/Atlantic (green), Suwannee (blue), and Everglades (magenta). The orange arrow indicates location of the Tennessee subpopulation. The yellow arrow indicates one individual in central FL that clusters with the NC clade. 3B shows the geographic distribution of the STRUCTURAMA-inferred genetic clusters. Color key is the same as 3A, except the yellow shape denotes the range of the Carolinas population inferred by nDNA versus the NC clade inferred by mtDNA. The yellow arrow points to the same individual in 3A, which clusters with the Suwannee population in the STRUCTURAMA analysis.

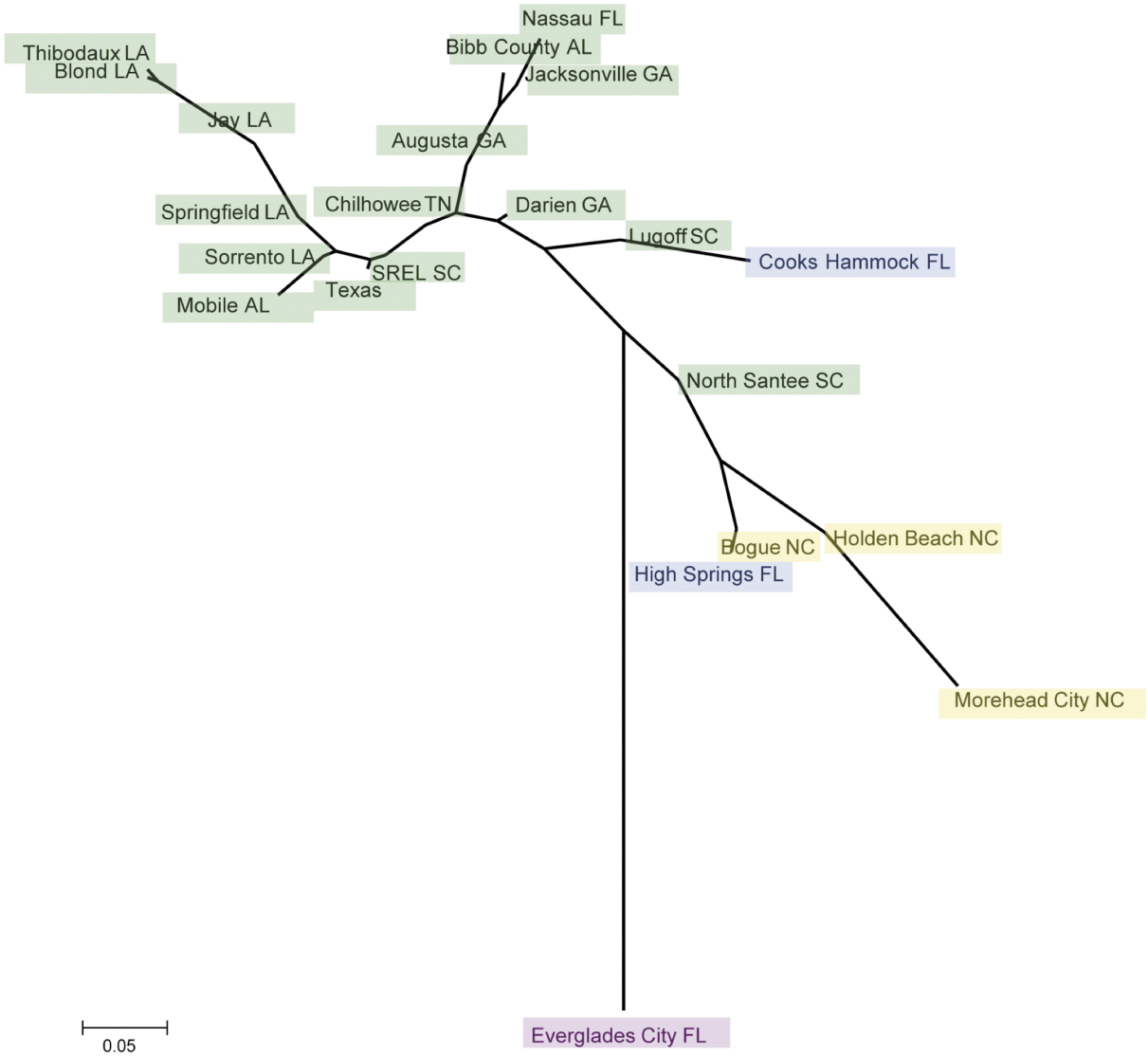

Figure 5. Neighbor-joining tree derived from pair wise F_{ST} of green anole subpopulations. Colors indicate the mitochondrial clade to which individuals in the subpopulation belong: Gulf/Atlantic (green), NC (yellow), Suwannee (blue) and Everglades (magenta).

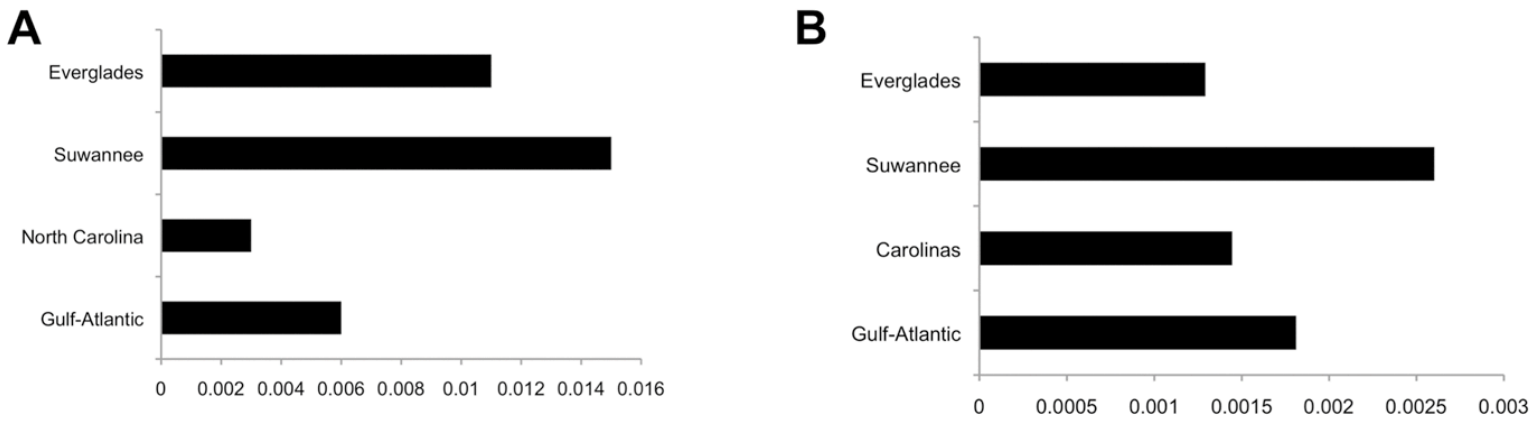

Figure 6. Pairwise distances within populations. A: Tamura-Nei corrected distance within each mitochondrial clade. B: Average p-distance across nine nuclear loci within each STRUCTURAMA-inferred population.

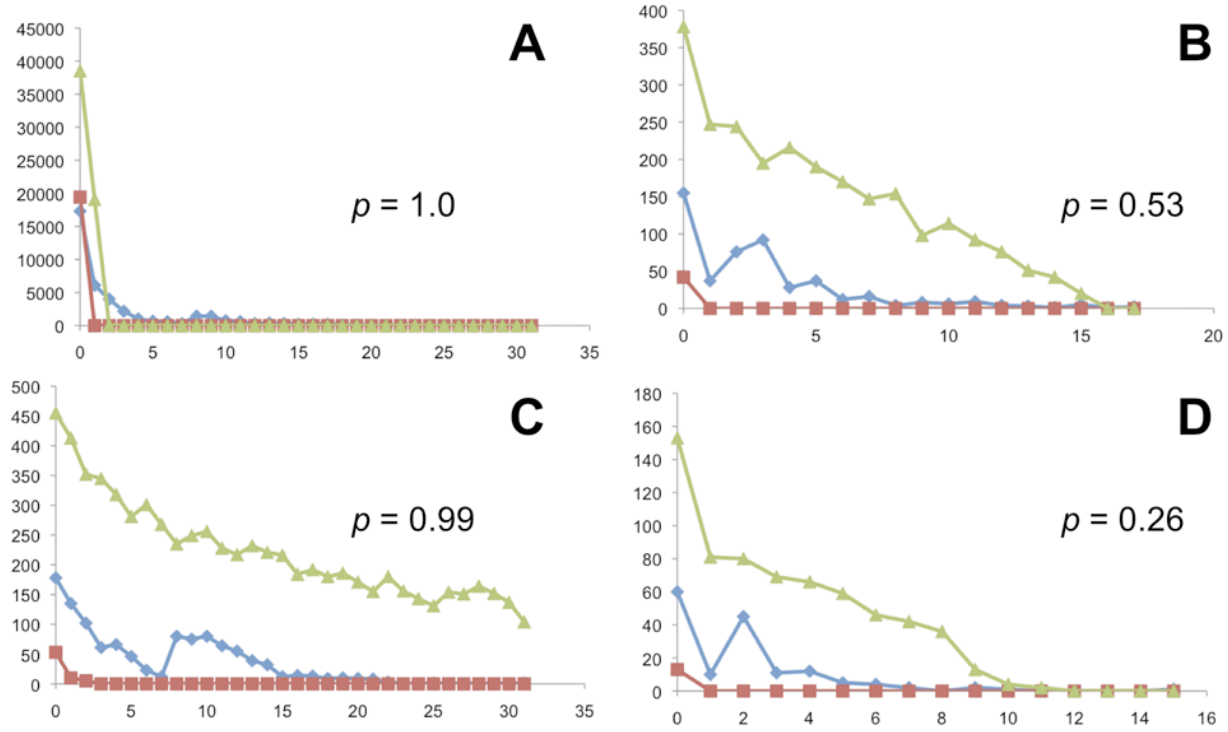

Figure 7. Mismatch distributions. The frequency distribution of nDNA polymorphisms within STRUCTURAMA-inferred populations calculated with the concatenated dataset in Arlequin. X-axes are in number of differences and Y-axes are in number of observations. Blue diamonds represent the observed data, green triangles and red squares represent upper and lower bounds expected under a model of expansion, respectively. P-values of the raggedness index for each analysis are given. A: Gulf-Atlantic. B: Suwannee. C: Carolinas. D: Everglades.

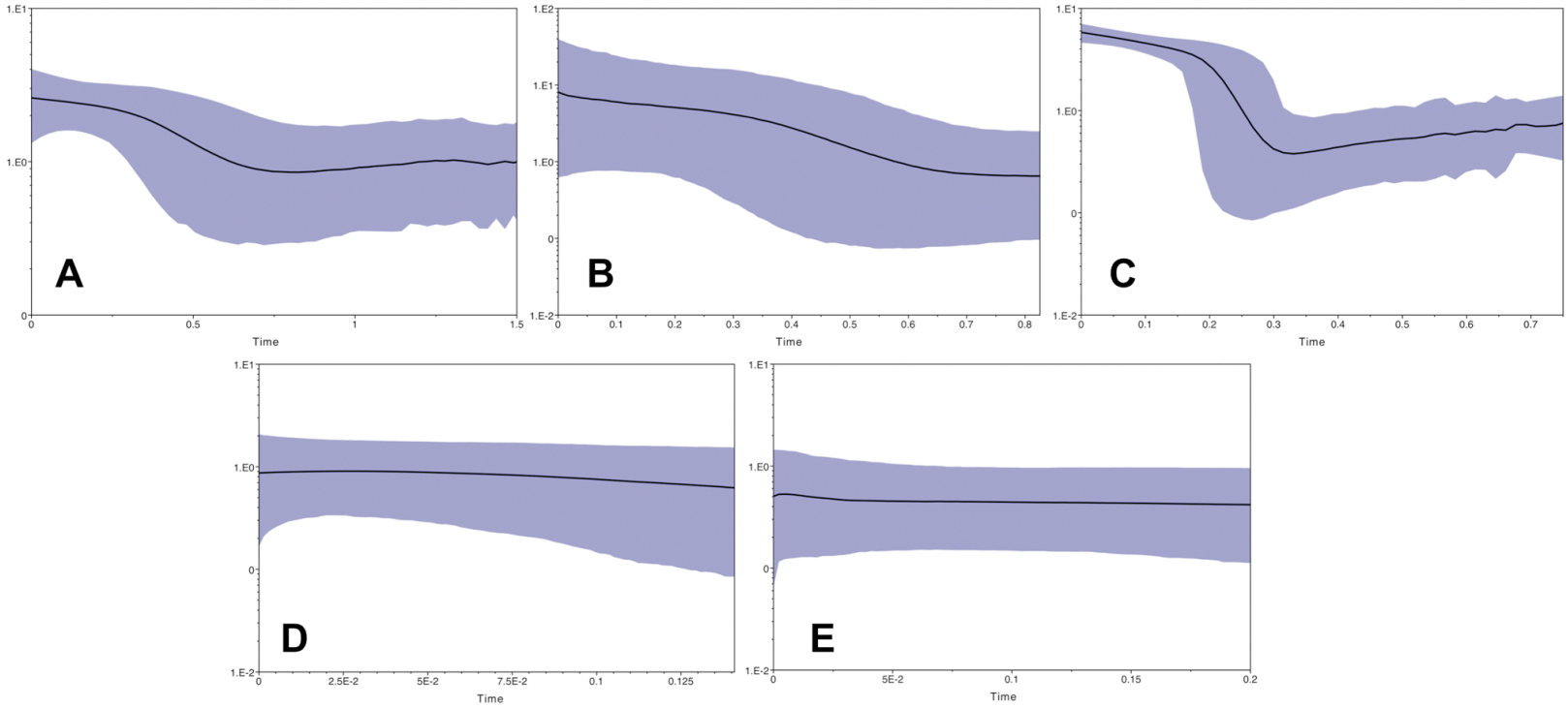

Figure 8. Bayesian Skyline Plots (BSPs). BSPs represent population size changes over time, inferred with mtDNA and an assumed mutation rate of 1.3% per million years. The X-axes are time in millions of years. Y-axes are mean effective population size in millions of individuals divided by generation time (for *Anolis* we assume a generation time of one year) on a log scale. Shaded areas encompass 95% highest posterior density (HPD). A: Suwannee. B: Everglades. C: Gulf/Atlantic. D: North Carolina. E: Tennessee.

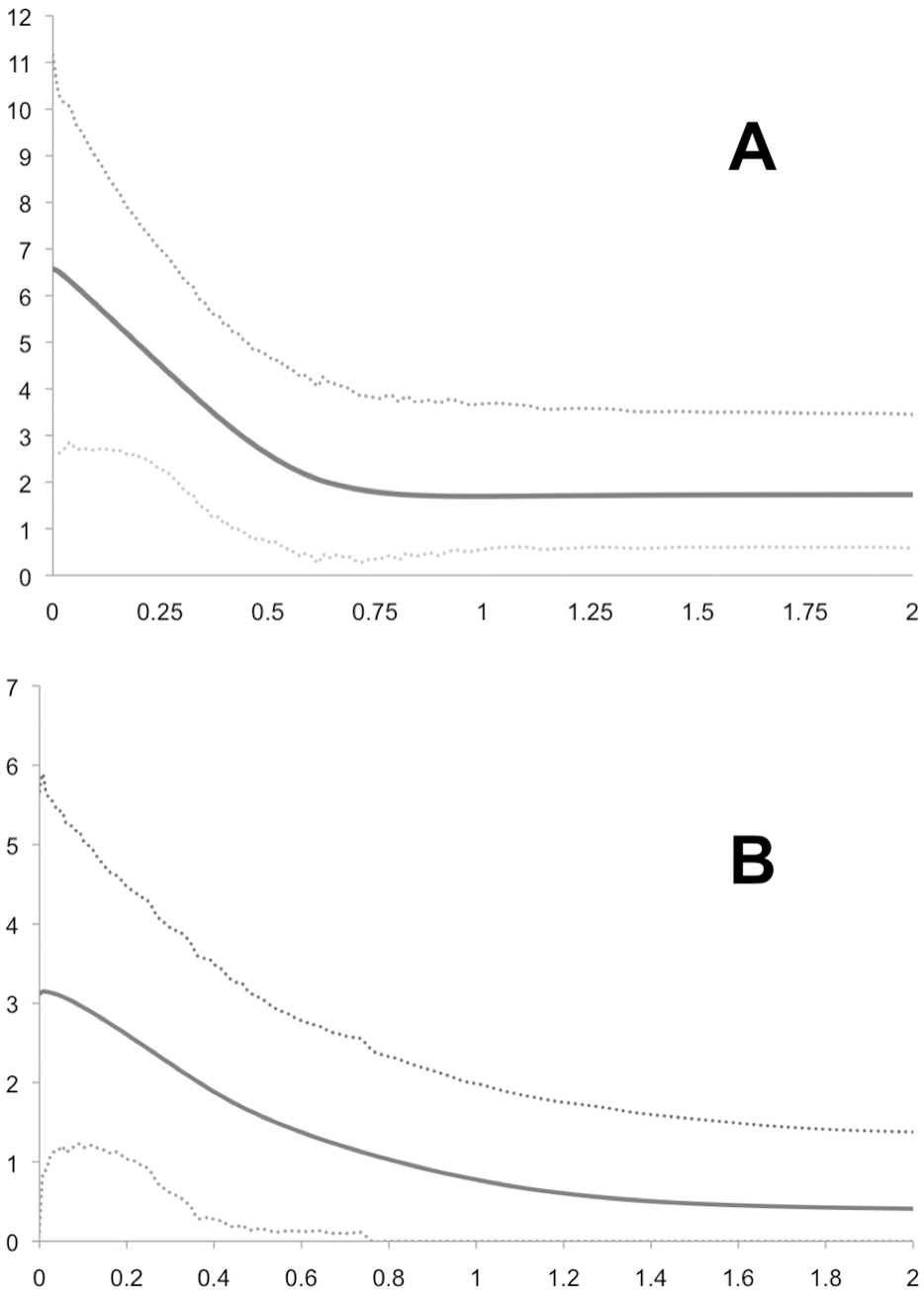

Figure 9: Extended Bayesian Skyline Plots (EBSPs). EBSPs represent population size changes over time in two of the mtDNA clades, inferred by mtDNA and multiple nuclear loci. Y-axes are effective population size divided by generation time. X-axes are in millions of years. A: Suwannee. B: Everglades.

Tables Chapter Two

County	State	Location ID	Latitude	Longitude
St. Tammany	LA	AC2	30.639267	-90.083333
St. Tammany	LA	AC3	30.50155	-90.083333
Livingston	LA	AC4	30.4316	-90.5474
Livingston	LA	AC5	30.306683	-90.610483
Livingston	LA	AC6	30.3284	-90.8272
Lafourche	LA	AC7	29.807267	-90.8563
Lafourche	LA	AC8	29.797883	-90.8129
Lafourche	LA	AC9	29.58905	-90.442933
Ascension	LA	AC10	30.223667	-90.85185
Collier	FL	AC13	25.918667	-81.310783
Okeechobee	FL	AC14	27.352883	-80.820633
Alachua	FL	AC15	29.82665	-82.597133
Suwannee	FL	AC16	30.280083	-82.819167
Lafayette	FL	AC17	30.028333	-83.1865
Lafayette	FL	AC18	29.90955	-83.314983
Dixie	FL	AC19	29.737733	-83.2363
Carteret	NC	AC20	34.727583	-76.736183
Carteret	NC	AC21	34.69675	-77.044333
Brunswick	NC	AC22	33.9194	-78.268617
Georgetown	SC	AC23	33.368067	-79.270533
Georgetown	SC	AC24	33.247583	-79.375717
Georgetown	SC	AC25	33.204483	-79.387667
Kershaw	SC	AC32	34.243983	-80.654333
McIntosh	GA	AC26	31.4682	-81.36675
McIntosh	GA	AC27	31.35295	-81.447467
Duval	GA	AC28	31.844867	-83.050667
Dodge	GA	AC29	31.9278	-83.181367
Johnson	GA	AC30	32.7019	-82.64255
Columbia	GA	AC31	33.2752	-82.388917
Blount	TN	AC33	35.54715	-84.051833
Blount	TN	AC34	35.550933	-84.020867
Blount	TN	AC35	35.550783	-84.023017
Blount	TN	AC36	35.53855	-84.07625
Blount	TN	AC37	35.55645	-84.012233
Blount	TN	AC38	35.5558	-84.00245
Blount	TN	AC39	35.531217	-83.991067
Bibb	AL	AC40	33.001667	-87.1488
Perry	AL	AC41	32.668883	-87.242433
Wilcox	AL	AC42	32.05015	-87.248367
Baldwin	AL	AC43	30.9004	-87.850967
Mobile	AL	AC44	30.733133	-88.03865
Baldwin	AL	AC45	30.79325	-87.899617
Crenshaw	AL	AC46	31.721167	-86.273167
Hempstead	AR	AR1	33.62075	-93.86015
Hempstead	AR	AR2	33.6635	-93.8435
Wood	TX	TX	32.944667	-95.193683
Aiken	SC	SREL	33.34279	-81.734333
Augusta	GA	GA	33.55019	-82.03929
Nassau	FL	FL	30.61009	-81.62317

Table 1. GPS coordinates and locality information for the samples used in this study.

Locus	Gene product
RALGAPA	ral GTPase-activating protein subunit alpha-1 isoform 1
HMGCS	hydroxymethylglutaryl-CoA synthase cytoplasmic
TERT	telomerase reverse transcriptase
MON2	MON2 homolog

Table 2. Locus names and predicted gene products.

Locus	Orientation	Oligo
ND2	Forward	AAGCTTTCGGGCCCATACC
	Reverse	GCGAATGGAAGCCCGCTGG
RALGAPA	Forward	CGTGAAGGTGTACGTCTTTTCC
	Reverse	GATCATCTTGTCCTTTATCTCCTGA
HGMCS	Forward	TTCAGGATTTGCTGCTACCC
	Reverse	TGTCTAAGCATCATGTTTTCTGC
TERT	Forward	TTTTCTGCCGCAGGTATTTC
	Reverse	TGGATGGTTCTGTCACTTCCT
MON2	Forward	ACATTCTGCCCCTGTATGGA
	Reverse	AGGTTCCTTGCCTCCTCATT
C1	Forward	GGGTGAATCTTCCCTCCTTC
	Reverse	AGGCCCGAGATTCAATGATA
R1	Forward	CATGCCTTTCCAAAAATCGT
	Reverse	CACCAGCCCCTACCAAAATA
Gav1	Forward	CCCCAGGAAATGAGATTCAA
	Reverse	AAAGTGCTGCAGGGCTTAAA
Gav3	Forward	GCCTTCTGGCTTTCTCACAC
	Reverse	TTTTGCCTTCCAAATCCAAG
Gav4	Forward	TGATTCCAACTGCGAGACAG
	Reverse	TTTATTCCCAGCCTGAAACG
Gav5	Forward	AGTCCAACCATCCTCACTGC
	Reverse	TTGGTCAAAGCTGTGCTGTC

Table 3. PCR primers used in this chapter.

	Locus	Length (bp)	#seqs	#haps	Hd	*s*	*k*	π
mtDNA	ND2	1172	191	128	0.987 ± 0.003	215	19.2062	0.02339 ± 0.00226
Intron	RALGAPA	970	130	8	0.513 ± 0.048	7	0.98366	0.00101 ± 0.00012
	HMGCS	1288	144	21	0.861 ± 0.016	22	4.07246	0.00317 ± 0.00012
	TERT	1077	124	12	0.383 ± 0.055	13	0.72528	0.00067 ± 0.00013
	MON2	688	148	8	0.479 ± 0.048	8	1.08917	0.00138 ± 0.00016
Anonymous	C1	356	144	10	0.695 ± 0.028	8	1.86616	0.00572 ± 0.00027
	R1	557	136	9	0.664 ± 0.36	12	1.60643	0.00292 ± 0.00021
	Gav1	266	212	4	0.255 ± 0.038	3	0.26585	0.001 ± 0.00016
	Gav3	455	260	13	0.402 ± 0.035	12	0.44505	0.00098 ± 0.0001
	Gav4	211	304	11	0.354 ± 035	9	0.43792	0.00209 ± 0.00024
	Gav5	350	226	11	0.547 ± 0.035	9	1.4719	0.00422 ± 0.00043

Table 4. Overiew of genetic data.

#seqs – number of haploid sequences in data set
#haps – number of haplotypes
Hd – haplotype diversity
s – number of segregating sites
k – average number of differences between sequences
π – nucleotide diversity

	nDNA x mtDNA		nDNA x STRUCTURAMA	
	Percentage of total variation	Fixation index	Percentage of total variation	Fixation index
Among groups	0.3455	Fct = 0.34549	0.3466	Fct = 0.34664
Among populations within groups	0.0459	Fsc = 0.07016	0.0336	Fsc = 0.05146
Within populations	0.6086	Fst = 0.39141	61.97	Fst = 0.38027

Table 5. Analysis of Molecular Variance (AMOVA).

Pairwise comparison	F_{ST}
Gulf-Atlantic - Everglades	0.59665
Gulf-Atlantic - Suwannee	0.20249
Gulf-Atlantic - Carolinas	0.31525
Everglades - Suwannee	0.60635
Everglades - Carolinas	0.45624
Suwannee - Carolinas	0.15882

Table 6. Pairwise F_{ST} measured between STRUCTURAMA-inferred populations with nDNA.

	Summary Statistics for Gulf-Atlantic Population					
	#seqs	#haps	Hd	pi	D	Fs
RALGAPA	72	3	0.082	0.00025	-1.48159	-0.837
HGMCS	96	13	0.748	0.00226	-0.18846	-1.079
MON2	106	5	0.405	0.00288	-0.14931	0.179
TERT	74	5	0.156	0.00015	-1.68429	-5.181
C1	94	6	0.534	0.00546	1.76473	1.528
R1	80	6	0.686	0.00273	1.11681	0.764
Gav1	130	4	0.105	0.0004	-1.3833	-4.016
Gav3	180	8	0.431	0.00136	-1.31937	-4.695
Gav4	210	7	0.243	0.00125	-1.49614	-5.875
Gav5	168	11	0.412	0.00324	-1.00792	-3.486

	Summary Statistics for Carolinas Population					
	#seqs	#haps	Hd	pi	D	Fs
RALGAPA	20	2	0.521	0.00161	2.26638	4.362
HGMCS	26	8	0.695	0.00157	-0.7596	-1.557
MON2	18	3	0.621	0.0017	1.48666	1.834
TERT	24	2	0.464	0.00043	1.23177	1.362
C1	28	3	0.204	0.00104	-1.32136	-0.674
R1	18	1	0	0		
Gav1	26	3	0.674	0.00316	1.2909	1.059
Gav3	32	4	0.181	0.00041	-1.72954	-3.49
Gav4	38	1	0	0		
Gav5	40	3	0.641	0.00549	1.61832	4.262

	Summary Statistics for the Suwannee Population					
	#seqs	#haps	Hd	pi	D	Fs
RALGAPA	22	2	0.416	0.00043	0.89527	0.62
HGMCS	10	4	0.644	0.00209	-0.21206	1.278
MON2	10	2	0.356	0.00045	0.01499	0.417
TERT	8	5	0.893	0.00176	-0.08352	-1.273
C1	12	4	0.758	0.00649	0.02646	1.655
R1	10	4	0.733	0.00355	-0.8501	0.568
Gav1	24	1	0	0	0	0
Gav3	18	2	0.471	0.00103	1.16615	1.215
Gav4	24	3	0.663	0.00234	1.46643	2.04

	Summary Statistics for the Everglades Population					
	#seqs	#haps	Hd	pi	D	Fs
RALGAPA	6	4	0.8	0.0011	1.03194	-1.685
HGMCS	12	2	0.303	0.00024	-0.19492	0.297
MON2	4	2	0.667	0.00084	1.63299	0.54
TERT	6	4	0.867	0.00239	-0.06042	-0.024
C1	4	3	0.833	0.00421	-0.75445	-0.288
R1	4	1	0	0	0	0
Gav1	20	2	0.395	0.00148	0.72261	0.976
Gav3	18	2	0.209	0.00046	-0.52899	-0.011
Gav4	20	2	0.189	0.0009	-0.59155	-0.097

Table 7. Summary statistics for each nDNA locus per population.

Locality	#seqs	k	π	D	Fs
Morehead City NC	10	0.888889	0.000178	-1.99318	-8.89663
Bogue NC	10	6.488889	0.001025	1.03078	-2.77142
Holden Beach NC	12	7.5	0.001185	0.58181	-6.14575
North Santee SC	12	10.878788	0.001719	0.77695	-4.62249
Darien GA	10	6.977778	0.001103	2.3592	-4.63494
Jacksonville GA	8	3.928571	0.000621	-0.76049	-2.50537
Lugoff SC	16	2.658333	0.00081	-1.23975	-12.47171
SREL SC	20	4.926316	0.000778	-1.98059	-10.54979
Augusta GA	26	4.206154	0.000665	-1.85808	-16.07731
Nassau FL	12	3.69697	0.000584	-2.06418	-7.01941
Bibb County AL	10	1.066667	0.00025	-2.40972	-3.66668
Mobile AL	10	2.8	0.000522	-1.777	-2.24471
Chilhowee TN	48	1.333333	0.00104	-1.17614	-23.24228
Jay LA	14	3.010989	0.000513	-1.9401	-6.32842
Thibodaux LA	30	0.496552	0.0001	-1.91803	-3.402823
Springfield LA	10	0.977778	0.000218	-2.22977	-5.94942
Blond LA	14	1.428571	0.000226	-2.84274	-8.60437
Sorrento LA	8	5.25	0.000894	-0.47329	-0.38278
Texas	18	1.196078	0.000207	-2.75193	-10.1634
High Springs FL	10	3	0.00064	-0.67067	-3.55534
Cooks Hammock FL	14	2.912088	0.000642	-2.01253	-4.72596
Everglades City FL	20	1.9421	0.000325	-2.1162	-20.17262

Table 8. Summary statistics calculated in Arlequin for the nDNA at collecting localities.

	π		D		Fs	
	mtDNA	nDNA	mtDNA	nDNA	mtDNA	nDNA
Gulf-Atlantic	0.00527	0.001422	-2.48252	-2.30077	-133.938	-26.27
Suwannee	0.02699	0.002268	-1.20207	-2.35881	-4.575	-26.04368
Everglades	0.01123	0.000853	-0.39885	-2.1162	-1.039	-20.17262
NC (Carolinas)	0.00326	0.00214	-1.58456	-0.08414	-4.576	-25.39846

Table 9. Summary statistics for each mtDNA clade and STRUCTURAMA-inferred group.

	Gulf-Atlantic	NC (Carolinas)	Suwannee	Everglades
Gulf-Atlantic	**0.001665**	0.002566	0.002569	0.003281
	0.006179			
NC (Carolinas)	0.033713	**0.002361**	0.002619	0.003423
		0.003279		
Suwannee	0.064498	0.070281	**0.002375**	0.003526
			0.014733	
Everglades	0.062848	0.061362	0.075565	**0.000851**
				0.011451

Table 10. Average pair wise genetic distances within and between green anole populations.

Main Diagonal (bold): Top entry is average p-distance within the population calculated from nDNA. Bottom entry is average Tamura-Nei corrected distance within the population calculated from mtDNA. Upper diagonal: Average p-distance between two populations calculated from nDNA. Lower diagonal: Average Tamura-Nei corrected distance between two populations calculated from mtDNA.

	Current effective population size (millions of individuals)		
Lineage (method)	**Median**	**Mean**	**95% HPD**
Suwannee (BSP)	2.6	2.6	1.4-4.0
Suwannee (EBSP)	6.4	6.6	2.6-11.1
Everglades (BSP)	4	8.1	0.13-30
Everglades (EBSP)	3	3.1	0.13-5.7
Gulf-Atlantic (BSP)	5.8	5.8	4.5-7.0
North Carolina (BSP)	0.78	0.87	0.08-1.9
Tennessee (BSP)	0.41	0.5	0.03-1.2

Table 11. Estimates of effective population size obtained from the skyline plots implemented in BEAST.

BSP – Bayesian Skyline Plot

EBSP – Extended Bayesian Skyline Plot

HPD = highest posterior density.

Chapter Three: Genetic Variation in the Green Anole (*Anolis carolinensis*) Reveals Island Refugia and a Fragmented Florida During the Quaternary

While the green anole lizard (*Anolis carolinensis*) is an emerging genomic model species [77], aspects of its evolutionary and demographic histories in North America have only recent undergone scrutiny [195, 196]. In the previous chapter, I used molecular phylogenetics of the mitochondrial NADH-2 region and statistical clustering of 10 autosomal (nDNA) loci to uncover four major lineages of green anoles that tightly correspond to geography. These lineages were: (1) the Suwannee lineage endemic to the northern Gulf coast of Florida, (2) an Everglades lineage, (3) a North Carolina lineage, and (4) a Gulf-Atlantic lineage that ranged from the Atlantic Coast of South Carolina and Georgia and across the Gulf Coastal Plain to Texas. Campbell-Staton et al. (2012) reported highly congruent patterns of mitochondrial genetic diversity. One difference between these two studies is the fact that Tollis et al. estimated the divergence time of the major green anole lineage to be close to the Pliocene-Pleistocene boundary (between ~1.5 and 3 million years ago), while Campbell-Staton et al. estimated a much deeper divergence time (between ~7 and 18 million years ago).

It has been established that *A. carolinensis* dispersed to North America from Cuba [142], due to the fact that it is phylogenetically nested within western Cuban populations of *A. porcatus*. However, the precise timing of this dispersal event is unclear. The literature, including the fossil record, agrees that *A. carolinensis* inhabited North America during the Pleistocene. Both Tollis et al. (2012) and Cambell-Staton et al. (2012) have concluded that the species has probably inhabited what is now the U.S. state of Florida for its entire history. This suggests that *A. carolinensis* may have been exposed to significant upheaval caused by Pleistocene climatic oscillations, which drove the waxing and waning of glacial epochs throughout the past ~2 million years [197, 198], repeatedly inundating the modern-day Florida peninsula with seawater and turning it into a series of archipelagoes. These physiographic changes could have had profound vicariant effects, and various patterns of genetic discontinuity and endemism in Florida [147, 199] have been explained in this light. It is likely that similar fragmentation occurred between *A. carolinensis* populations living on the Florida peninsula during the Pleistocene.

Tollis et al. (2012) recovered four green anole lineages and three of them can be found in Florida; this remarkable diversity observed across a limited geography raises the possibility of multiple ancient island refugia. There was also evidence of a possible fifth lineage endemic to central Florida, although this was based on only two mtDNA haplotypes. Campbell-Staton et al. (2012) reported more mtDNA haplotypes unique to this region, confirming the existence of a central Floridian mtDNA clade and establishing that Florida is home to at least four highly divergent evolutionary lineages. The diversity of green anole lineages in Florida versus that of the continental mainland warrants a focused examination at a higher resolution of the dispersal patterns within this relatively limited geographic region. Also, resolving the issue of the time to most recent common ancestor (T_{mrca}) for *A. carolinensis* is extremely important because without proper divergence time estimates it will remain unknown if intraspecific diversity was driven by Pleistocene climatic oscillations in Florida or by more ancient phenomena that were unique to this taxon.

Relaxed clock phylogenetic methods that utilize fossil calibrations or informative substitution rate priors have made the estimation of divergence times more accurate [200, 201]. However, research has shown that the use of a single gene or the concatenation of multiple genes can lead to error because gene-tree heterogeneity is the rule rather than the exception [153, 202]. I will use a multi-locus coalescent approach, incorporating topological discord between sampled gene genealogies and providing a more accurate estimation of divergence histories, to date important events in the history of green anoles. The results of this analysis will be compared to those in Tollis et al. (2012) and Campbell-Staton et al. (2012), and used to assess the relative merits of single-gene versus multi-locus studies. A greater geographic sampling within Florida allows for the first time a multi-locus study cataloguing genetic diversity on the peninsula. By placing conclusions from this chapter in a comparative context as well as in a holistic manner that incorporates what is known about Florida geology, I intend to elucidate the history not only a model organism but of an interesting North American biogeographic region.

Methods

Sample preparation and marker design

A map in Figure 1 shows the collecting localities across Florida, including the 49 reported previously, six new localities in Central Florida where 35 additional anoles were collected in September 2012, as well as additional localities in southern Florida that were obtained from T. Hsieh of Temple University. Table 1 shows the GPS coordinates and locality information for the new samples. Genomic DNA was extracted from tissues using the Promega Genomic Wizard DNA Extraction kit. From the new samples, we amplified by PCR an 1172bp mtDNA fragment that includes the NADH-2 gene and two adjacent tRNAs with primers published in Tollis et al. (2012). We also amplified three nDNA loci using primers that were reported in Tollis et al. (2012) (HMGCS, RALGAPA, and TERT). PCR protocols were repeated as in Tollis et al. (2012), and all products were sequenced in both forward and reverse directions at the High-Throughput Genomics Unit at the University of Washington in Seattle, WA. Chromatograms were imported into Geneious v5.5 [160], where poor-quality regions were trimmed and heterozygous sites were called using the Find Heterozygotes plugin. Forward and reverse reads were assembled into contigs and consensus sequences were extracted and aligned using ClustalW [161] in BioEdit [162], where alignments were further edited by eye. Gametic phases of each nDNA haplotype were resolved computationally using PHASE 2.1 [163] with a 90% probability cutoff.

Phylogeographic Inference

The phylogeny of the mtDNA sequences published in Tollis et al. (2012) including the outgroups used in that study (*A. isolepis*, *A. altitudinalis*, and *A. porcatus*) in addition to 25 additional sequences from GenBank (see Table 2) and 45 sequences new to this study for a total sample of 299, was reconstructed with MrBayes3.2 [203]. The analysis was run for 20,000,000 generations, and we sampled 10,000 trees, discarding the first 1000 as burn-in. We applied the HKY + Gamma + Invariant sites model of sequence evolution as it was the most likely according to Bayes Factors using MEGA 5.0 [169]. We also reconstructed phylogenies for the HMGCS, RALGAPA and TERT genes, as well as a concatenated nDNA data set using Maximum Likelihood (ML) estimation as implemented with PhyML [204] with 500

bootstrap replicates to assess node support and the TN93 model of sequence evolution, which was also determined in MEGA. For the nDNA trees, we used *A. porcatus* as an outgroup.

In order to assess the number of populations in our sample without using *a priori* information about population structure (i.e. from the gene tree phylogenies), the entire nDNA dataset for 158 anoles was loaded into the Bayesian clustering program Structure 2.3.3 [176], which can estimate the likelihood of a user-set number of *K* clusters. We ran the admixture model for 60,000 generations, with 10,000 discarded as burn-in. Analyses for a range of *K* values from 1 to 10 were repeated three times each, and the results files were submitted to Structure Harvester [178], which chooses the most likely number of clusters using the ΔK method [179].

Genetic Variation and Historical Demography

As a comparison of genetic variation between Florida and the mainland, we computed the average pair wise genetic distance both within Florida and within the continental mainland using the Kimura 2-Parameter correction in MEGA 5.0. Taking into account where there was agreement between the phylogenetic and cluster assignments, we computed standard measurements of sequence diversity within all populations for each gene using DnaSP version 5 [165] including the number of haplotypes, hapoltype diversity (Hd), nucleotide diversity (π), Theta per site ($\Theta = 4N_e u$), Tajima's D and Fu's Fs. We also calculated F_{ST} and *Nm* [205] for each gene to estimate the extent of population subdivision and the extent of migration, respectively. We also calculated pair wise F_{ST} between all populations for each gene in DnaSP and for the entire nDNA dataset in Arlequin v3.5 [182].

To understand the historical demography of the newly sampled Central Florida clade, we first examined the mismatch distribution of nDNA polymorphisms, comparing the observed data to what is expected under a model of population size expansion. We also constructed an Extended Bayesian Skyline plot [187] for this population, which uses the coalescent histories of separate loci to estimate the number and extent of population size changes in the past.

Species Tree Estimation

To account for mito-nuclear discordance as well as stochastic differences in the coalescent histories of the sampled gene genealogies while making inferences about the evolutionary history of *A. carolinensis*, we used *BEAST [206]. *BEAST can implement a probabilistic framework that uses sequence information from different loci and multiple individuals per taxon. By incorporating prior probabilities on substitution rates, *BEAST will jointly estimate the species tree as well as the individual gene trees that evolve within it, taking deeper gene tree coalescence into account and providing better estimates of divergence times [207]. The goals of this analysis were twofold: (1) to obtain a population phylogeny representing the intraspecific evolutionary history, and (2) to estimate the divergence times of the major green anole lineages. Partitions included the NADH-2, HMGCS, RALGAPA and TERT alignments. Site models implemented were HKY + Gamma + Invariant Sites for NADH-2, and TN93 for the nDNA partitions, as determined previously in MEGA. We used sequences obtained from *A. porcatus* as an outgroup to root the species tree, and assigned each *A. carolinensis* individual a discrete trait representing one of the major green anole lineages as determined by two criteria: (1) its mtDNA clade and (2) >95% of its genome belonging to a specific nDNA cluster. Each partition contained eight individuals per lineage. For divergence time estimation, we used the NADH-2 mutation rate of 1.3% per million years and placed normally distributed prior probabilities around the mutation rate parameters for each nuclear gene with a starting mean of 1. This analysis used a Yule prior, which assumes an unknown yet constant birth rate of lineages.

The *BEAST results were obtained by combining the parameter files from two independent MCMC chains of length 500,000,000 each. To assess convergence between these runs we monitored the effective sample size (ESS) values and consistency of parameter estimates using Tracer v1.5, and once confirmed, the separate runs were combined using LogCombiner. The initial 10% of each run was ignored as burn-in. From the combined results, we sampled parameters and trees every 100,000 for a total of 10,000. Out of the 10,000 sampled genealogies, we obtained the maximum-clade credibility tree with divergence time estimates for the species tree (as well as for the embedded mtDNA gene tree) using TreeAnnotator, discarding the first 1,000 samples as burn-in.

Results

Phylogeographic analysis

The mtDNA phylogeny recovered four major clades, and the topology and geographic distribution of clades were highly congruent with the results published in Tollis et al. (2012) and Campbell-Staton et al. (2012) (Figure 2). The four clades include: (1) the Suwannee clade endemic to the central and northern Gulf Coast of Florida up to the eastern side of the Apalachicola River, (2) the Everglades clade centered around the Gulf and Atlantic coasts at the southern tip of the Florida peninsula and northwards up to Lake Okeechobee, (3) a clade featuring a paraphyletic Central Florida lineage from the central Atlantic Florida coast, with a North Carolina lineage nested within it, and (4) the Gulf-Atlantic clade which includes individuals from localities along the Atlantic coasts of South Carolina and Georgia, across the Gulf Coastal Plain and the Mississippi River into Texas. This wide-ranging clade includes some populations from northern Florida (see below). All of these nodes featured >95% posterior probability support.

Within the major mitochondrial clades of *A. carolinensis* there is a considerable amount of diversity, especially in Florida. For instance, the Suwannee clade consists of two well-supported minor clades (both with 100% posterior probability), the first of which ranges primarily along the Gulf of Mexico coastline from Tampa Bay to just east of the Apalachicola River, while the second one is centered along the Central Highlands that run inland down the northwestern part of the peninsula. There is also considerable structure within the Everglades clade: a lineage from the Gulf Coast of very southern Florida, near Everglades City (96% posterior support), and a more geographically diverse lineage (98% posterior support) that consists of individuals from the Atlantic Coast in and around Miami as well as a nested lineage from South Bay just south of Lake Okeechobee (100% posterior support). The third major Floridian lineage, Central Florida, is perhaps the most complex because it consists of two deep mitochondrial lineages (both with 100% posterior support) across its relatively limited geographic range in Florida; one of which contains a nested clade (100% posterior support) that is endemic to North Carolina. Interestingly, anoles sampled near Jacksonville, Florida in Nassau County, fall within the Gulf-Atlantic clade. These anoles are more closely related to conspecifics greater than 1000km away in Texas than

they are to conspecifics less than 50km to the south in neighboring Florida counties. There are also individual mtDNA haplotypes sampled from near Panama City, west of the Apalachicola River, which were assigned to the Gulf-Atlantic clade.

There was significant discordance across the nDNA gene tree topologies and branch lengths and very little statistical support overall (Figure 3); however, qualitative similarities across these gene trees included monophyletic groups from NC and the Atlantic seaboard, and the Everglades, Suwannee, and the Gulf Coastal Plain respectively. The most likely tree from a ML analysis of a concatenated nDNA alignment is shown in Figure 3D. Much of the gene tree discordance was due to differential placement of sequences from individuals across Florida, most notably from South Bay, which phylogenetically groups with the Everlgades mtDNA clade but shows very little evidence of affinity to the Everglades based on the nDNA.

The ΔK method from our Structure analysis indicated that the most likely number of clusters in our nDNA dataset was five (Figure 4). The geographic distribution of these clusters closely mirrored that of the five lineages of the mtDNA phylogenetic analysis (Figure 5A and B). The population near Nassau, Florida clusters with the Gulf-Atlantic population rather than the other Florida populations, as it does in the mtDNA analysis. Within each cluster there was no discernable structure such as within the major mtDNA clades. One disagreement with the mtDNA phylogeny that was shared between nDNA gene trees and the clustering method was that individuals collected in South Bay were assigned to both the Central Florida and Suwannee nDNA cluster, with very little evidence of persistent Everglades haplotypes in this area. This differs from the mtDNA phylogeny in that all South Bay mtDNA haplotypes group with the Everglades. Another point of mito-nuclear discordance was the detection of significant nDNA admixture between the Gulf-Atlantic and North Carolina populations that extends into southeastern Georgia, which is notably absent in the mtDNA analysis. Overall, the Structure analysis agrees with the mtDNA phylogeny in that green anoles in Florida belong to four of the major evolutionary lineages: Central Florida, Everglades, Suwannee, and Gulf-Atlantic.

Genetic Differentiation and Historical Demography in Florida

The average genetic distances between individuals within Florida are much greater than those of the continental mainland (0.0538 versus 0.0113 mitochondrial and 0.0029 versus 0.0014 autosomal). Table 3 shows the genetic diversity overview for each gene. In general, genetic diversity (measured by Hd, π or Θ) is greatest in the Central Florida population, with some exceptions where diversity estimates in the Everglades and Suwannee are greater. Non-significant negative values of Tajima's D and very few low negative values of Fu's Fs across all genes suggest a lack of departures from neutrality due to population bottlenecks or expansions in Florida. The fact that Central Florida harbors a relatively large amount of genetic diversity is consistent with the mismatch distributions, which reject a model of expansion, and the EBSP (Figure 6), which places most of posterior probability under a model of zero past population size changes. Across all five green anole populations, F_{ST} calculations are significant and *Nm* is consistently small (<<1) (Table 4), suggesting strong population subdivisions with relatively low amounts of gene flow and few migrants. The pair wise calculations F_{ST} for each gene show highly significant values for every population comparison (Table 5).

Species Tree Estimation

Parameter estimates between the separate *BEAST runs were highly concordant, suggesting convergence, and the ESS values were very high for all parameters (mostly over 4000), suggesting proper mixing of the MCMC chain. The species tree topology (Figure 7) differed significantly from that of the mtDNA phylogeny, suggesting the early divergence of the Everglades population, followed by a split between Central Florida and North Carolina, and then another more recent split between the Suwannee and Gulf-Atlantic lineages. The divergence time estimates were more similar to the calibrated mtDNA-based results published in Tollis et al (2012) than to those of the relaxed clock analysis Campbell-Staton et al (2012), providing more support for the hypothesis that the major evolutionary events in this species' history occurred during the Pleistocene. Our *BEAST analysis placed the T_{mrca} for *A. carolinensis* and the branching off of the Everglades population at ~2.1 million years (1.3 – 3.1 95% HPD) ago. The next divergence was between populations that live along either the Atlantic Coast, or the Gulf Coast and more inland across the Gulf Coastal Plain, and occurred ~1.4 million years ago (0.83 – 2.3 95% HPD). Along

the Atlantic Coast, the split between Central Florida and North Carolina occurred ~1.3 million years ago (0.74 – 1.8 95% HPD). Meanwhile, the Gulf-Atlantic lineage split off from the Suwannee ~750,000 years ago (0.45 – 1.1 95% HPD).

Discussion

We used phylogenetic and clustering methods to delimit five distinct evolutionary lineages of green anoles, using a greater geographic sampling across the Florida peninsula to further extend the findings from Chapter 2. The lineages we recovered that were identical to those described in Chapter 2 are: the Suwannee from the Gulf Coast of Florida, the Everglades, the Gulf-Atlantic, and North Carolina. The new lineage we are reporting here is Central Florida, a genetically diverse and paraphyletc mtDNA lineage endemic to the central Atlantic Coast of the Florida peninsula, and actually contains the North Carolina clade. Statistical clustering of the nDNA also supports this fifth population. Inferences of the historical demography for Central Florida include long-term stability with no evidence of population size changes. This is in contrast to all other lineages, which have experienced recent population size changes – as in the Gulf-Atlantic lineage ~300,000 years ago – or more ancient population size changes – as in the Suwannee or Everglades ~500,000 – ~1 million years ago (see Chapter 2 of this dissertation). Overall, we have found that the greatest genetic diversity within this species exists in Florida.

It is worth noting that the earliest instances of *A. carolinensis* in the North American fossil record are from the Late Pleistocene [143]. Our results consistently suggest that the reason for this is likely due to the fact that the history of green anoles on the continent may not extend much further beyond the Pleistocene. In Chapters 2 and 3 of this dissertation, I used multi-locus and coalescent-based methods that take into account the differing histories between genes across the mitochondrial and nuclear genomes to estimate both divergence times and demographic events. All analyses have come to the same conclusion: the most recent common ancestor of North American green anoles probably lived just over two million years ago and all of the major demographic events in the history of *A. carolinensis* occurred during the Pleistocene. Tollis et al. (2012) used a time-calibrated phylogenetic approach that centered on previous divergence time estimates within the *carolinensis* subgroup in conjunction with the pair wise genetic

distances of the major lineages within *A. carolinensis*, and a strict molecular clock to estimate a T_{mrca} that straddles the Plio-Pleistocene boundary at ~1.3 – 2.9 million years. We also tested demographic hypotheses within each population using both the single mtDNA gene and multiple nuclear loci; the timing of these inferred demographic events closely mirrored the divergence time estimates from the calibrated phylogenetic analysis. In contrast, Campbell-Staton et al. (2012) used a relaxed molecular clock model, with the same mtDNA region and fixed the mutation rate at 1.3% per million years, to estimate a T_{mrca} of ~6.8 – 17.8 million years, with the radiation of the lineages within *A. carolinensis* occurring well before the Pliocene and sometime during the Miocene.

We suggest that the dramatic contrast of our inferences to those made in Campbell-Staton et al. (2012) stem from the use of a single mtDNA gene to reconstruct the timing and order of population divergences within a species. The risk of over-interpretation when using one mtDNA gene lies in many factors. First, it is well established in evolutionary biology that during lineage diversification, gene histories will diverge before populations. Since gene trees are embedded within species trees, it is expected that dates estimated from gene trees will overestimate the true divergence times. An example is the mtDNA analysis from the species-tree approach used in this chapter. While the divergence time estimates for all green anole populations when incorporating all loci lie within ~3 million years, the embedded mtDNA analysis produced much more ancient divergence time estimates that are more similar to those in Campbell-Staton et al. (2012), including a T_{mrca} of 5.7 – 8.4 million years (95% HPD). Another confounding factor in single-gene phylogeographic inference is gene tree heterogeneity due to random effects of coalescence in finite populations [153]. This is obviously the case for *A. carolinensis* and is apparent when one compares the topologies of the mtDNA and nDNA phylogenies (Figures 2 and 3, respectively). Other problems, related to using a relaxed clock model explicitly, include the relatively rapid substitution rates in mitochondrial genomes, which may cause over-saturation in more ancient lineages, thus severely biasing divergence time estimates at larger time scales [208]. At the other temporal end, applying a relaxed clock within species may be confounded by ancestral sequence polymorphisms that have not yet sorted between lineages, leading to an apparent yet false inflation of molecular clocks at shorter time scales [209, 210]. These were the reasons why Tollis et al. (2012) applied the admittedly somewhat complicated

calibration approach to estimating the *A. carolinensis* T_{mrca}; we wanted to effectively rein in the tendency of mtDNA to overestimate. In the current chapter, we used *BEAST to implement a coalescent framework using multiple loci with differing effective population sizes (i.e. from both the mitochondrion and the nucleus) and topologies to obtain what we believe to be a more biologically realistic estimate of the timing of diversification events in the history of *A. carolinensis*. We have shown empirically that this method produces more recent divergence time estimates than relaxed clock models using a single gene (such as implemented in BEAST). Our findings are consistent with a previous study in birds (jays of the genus *Aphelocoma*) [207], which compared *BEAST to BEAST and concluded that the multi-locus coalescent analysis provided more robust divergence time estimates that were more shifted towards the present.

Due to the discordance across the mtDNA and nDNA gene tree topologies, elucidating the branching order of green anole populations in time has proven to be complicated and requires more than a simple interpretation of the mtDNA phylogeny. However, there are aspects of *A. carolinensis* evolutionary history that can be inferred from the gene trees. For instance, in addition to the well-supported geographic patterns in the mtDNA phylogeny (and to some degree in the nDNA phylogenies), there is strong structure present across all genes if one examines their global and pair wise F_{ST}, suggesting that all of these data sets contain valuable signatures of population history (Tables 4 and 5). Also, there are reciprocally monophyletic lineages across both mtDNA and nDNA gene trees, including the Everglades and North Carolina, which suggests the long-term geographic isolation of these populations. The non-phylogenetic approaches (i.e. the clustering and distance methods) support very well the five-lineage scenario and allow for robust population delimitation. From the species tree it is obvious that the coalescent histories of the nuclear genes are driving the topology differences when compared to the mtDNA phylogeny. This may be due to possible artifacts of mtDNA evolution such as selective sweeps or incomplete lineage sorting resulting from relatively rapid population fragmentation during the diversification process.

It is clear that Florida, which comprises less than 15% of *A. carolinensis* total natural range, is the cradle of green anole diversity. The populations found along the Florida peninsula are the most ancient

according to the phylogenetics, and by every measure genetic variation is greatest in all of these populations when compared to more wide-ranging conspecifics across the mainland. As all of our inferred demographic events in Florida occurred hundreds of thousands of years before our inferred demographic events on the mainland, it is probable that Florida green anole populations have been afforded the longest stretches of time with general stability. This is the most apparent in the Central Florida lineage, which is comprised of two deeply divergent mitochondrial lineages, contains individuals with haplotypes from three of the nDNA clusters, and harbors elevated sequence diversity when compared to all other populations. Also, while most populations show at least some evidence of effective size changes at some point in the past, Central Florida seems to have been demographically stable for most if not all of its history. The population size changes inferred for the Suwannee and Everglades populations were much more ancient than the ones inferred for the mainland populations. In contrast to this relative stability, continental mainland populations are either isolated, as in the North Carolina population, or are phylogenetically young and have much more recently undergone dramatic range and size expansions, as in the Gulf-Atlantic population (see Chapter 2). As divergent mtDNA haplotypes from two clades are found on either side of the Apalachicola River along the panhandle of Florida, it seems that this river system had provided some kind of dispersal barrier, which has been observed in other taxa [147].

If the “Out of Cuba” hypothesis is correct, then Florida provided the first stepping-stone for the invasion of *Anolis* lizards into the mainland southeastern United States [142]. Florida has a dynamic geologic history [197] and according to our estimates *A. carolinensis* may have arrived just in time for dramatic physiographic upheaval with possible vicariant effects. While Florida maintained its form as a peninsula throughout the Pliocene, the Pleistocene was the era of glacial epochs. The first of these began ~2 million years ago and caused the receding of the surrounding ocean and the augmentation of the Florida peninsula, particularly along its western edge. During interglacials throughout the Quaternary, the sea would rise hundreds of feet above its present level, effectively reducing Florida to a series of islands. The central Florida region corresponds to these archipelagoes, primarily along the backbone provided by the Central Highlands, which run north-to-south through the peninsula. West of the highlands in the northwestern part of the peninsula are other areas of isolated elevation. This dynamic history of Florida

has been invoked to explain numerous taxonomic discontinuities observed between Florida and the mainland as well as within the Florida peninsula itself, both in the historical literature [211] and more recently in the field of molecular phylogeography [147]. The geographic distribution of green anole lineages in Florida coincides nicely with the physiographic history: one needs only to look at the similar distribution of the divergent lineages (Figure 5) and positive and negative geologic elements on the peninsula (Figure 8). Our estimates of the timing of demographic events for *A. carolinensis* land squarely in the one of the most complicated parts of Florida's past; therefore, a plausible hypothesis is one where island refugia during the Pleistocene was the main driver of green anole diversity.

Given our divergence time estimates and the order of population splits within *A. carolinensis*, as well as the underlying geology of the region in which it lives, we may attempt to reconstruct the most likely historical biogeographic scenario (Figure 9). Green anoles most likely dispersed to the southern tip of Florida from Cuba during the Pliocene, when the peninsula was fully intact. Physiographic upheaval due to seawater inundation throughout the Early and Middle Pleistocene would have driven population fragmentation into island refugia, followed by gradual genetic divergence via drift. Once the peninsula reconnected to the mainland, populations living along the Atlantic Coast of Florida would have been able to disperse northward towards North Carolina, prevented from dispersing westward by the barrier posed by the Appalachian Mountains. Meanwhile, populations on the Gulf coast of Florida near the modern-day panhandle would have dispersed along the river drainage systems of the Gulf Coastal Plain. The lack of mountainous barriers along this landscape resulted in the dramatic westward expansion of green anoles across the Mississippi River and into Texas, which began approximately 300,000 years ago (see Chapter 2). More recently, secondary contact between Gulf-Coastal Plain and Atlantic Coast lineages produced admixed populations in the region just south of the Appalachian Mountains.

Conclusion

A. carolinensis is widespread and abundant throughout the southeastern United States but is most likely a relatively recent immigrant to the region, as our analyses suggest that the most recent common ancestor of modern populations lived during the Late Pliocene or Early Pleistocene. Upon dispersal to

Florida from Cuba, dynamic geological and topographic histories profoundly affected the geographic distribution of individuals in ways that are both consistent with and unique when compared to co-distributed taxa. We have found evidence suggesting that island refugia along the modern-day Florida peninsula drove the early diversification of green anole lineages. These Florida lineages show evidence of being the most ancient and the most stable in terms of population size over their demographic histories. After dispersal onto the continental mainland, two different founding green anole populations most likely undertook separate migrations along the river drainage systems of the Atlantic Coast and the Gulf Coastal Plain, respectively. Using a multi-locus coalescent model, we conclude that previous analyses based on a single mtDNA gene trees and relaxed clock phylogenetic models are biased towards older divergence time estimates. This study adds to the growing body of evidence that methods accounting for deeper gene tree coalescence are more desirable as they incorporate uncertainty and differences across all sampled loci.

Post-script: Human-mediated dispersal and hybridization between species

The dynamic history of green anoles in Florida appears to experiencing a new chapter, as a scenario of human-induced introgression is playing out. *A. carolinensis* is native to the southeastern U.S. and is nested phylogenetically within the paraphyletic *A. porcatus*, or Cuban green anole [142, 195]. Recently introduced to Florida, *A. porcatus* has become well established in the urban setting of Miami. Hybrids of these two species have been catalogued based on morphology and mitochondrial haplotypes [173], however any genetic impact on native populations remains unknown. The fact that these two species are sister taxa provides an opportunity to discover which genomic regions are implicated in their divergence. A more practical matter is that this introgression needs to be accounted for in future genetic studies of *A. carolinensis*, which is an experimental model organism with a complete genome sequence [77].

A nucleotide BLAST of a NADH-2 haplotype collected in a Miami public park returns a top hit not to *A. carolinensis* but rather *A. porcatus* (GenBank accession number AY654046, e-value = 0, 99% identity) (see Figure 2). In addition, the nuclear genes we sequenced for this individual were unique at a number of segregating sites in all three genes when compared to our *A. carolinensis* panel (P5, Figure 10, and

see Figure 3). This led us to conclude that this individual is in fact *A. porcatus*. Another individual from the same collecting locality has a NADH-2 haplotype from the Everglades (GenBank accession number EU106332, e-value = 0, 98.3% identity) but an nDNA SNP profile identical to the putative *A. porcatus* specimen (P4, Figure 10 and see Figure 3). Therefore, we have evidence that hybridization is occurring between native and Cuban green anoles. While this has been documented in the past based on discordance between morphological species assignment and mtDNA haplotypes [173], we report here for the first time a description of *A. carolinensis* – *porcatus* hybrids based on mito-nuclear genetic patterns.

Invasive species affect native communities, especially when they can hybridize with native taxa. Where they occur, these new hybrid zones become living laboratories of evolution. When contact between allopatrically diverged species is reestablished, a new mixture of parental genotypes is exposed to natural selection. Introgression of foreign alleles can be a potent source of genetic variation and a resource for adaptation [212, 213]. At the same time, gene flow can homogenize populations and slow adaptive divergence, creating a conservation concern with the potential of harm to the genetic integrity and long-term survival of the native taxon [214, 215]. Perhaps of most interest to evolutionary biologists, the amount of introgression between recently diverged species is a direct measurement of reproductive isolation, which may be the most important factor in the speciation process [216].

Figures Chapter Three

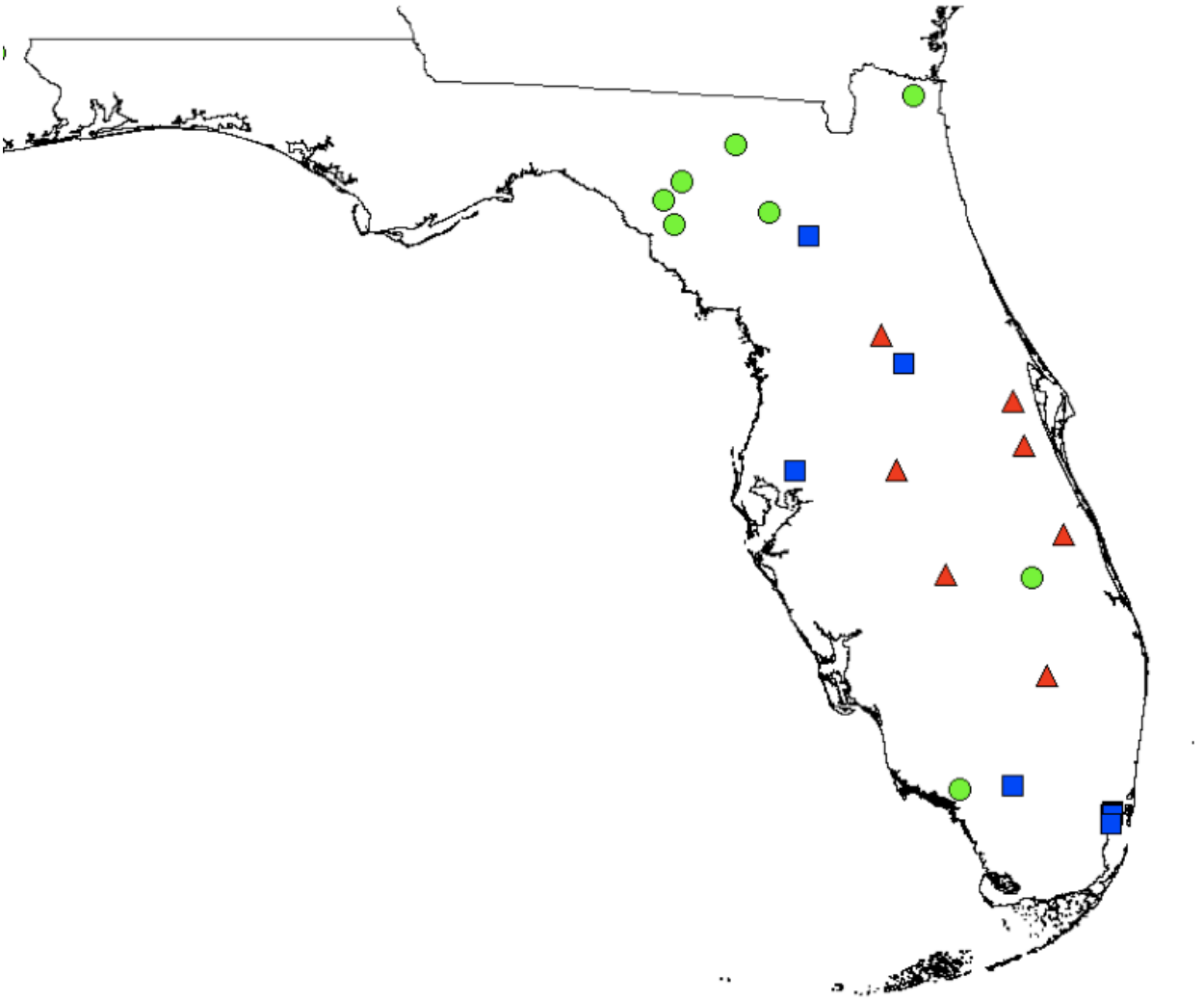

Figure 1. Collecting localities in Florida indicated by colored shapes. Green circles: collecting localities whose samples were reported in the previous chapter and were collected between 2009 and 2010. Red triangles: localities where green anoles were collected during September 2012. Blue squares: localities where samples given to us by T.Hsieh of Temple University in October 2012 were collected.

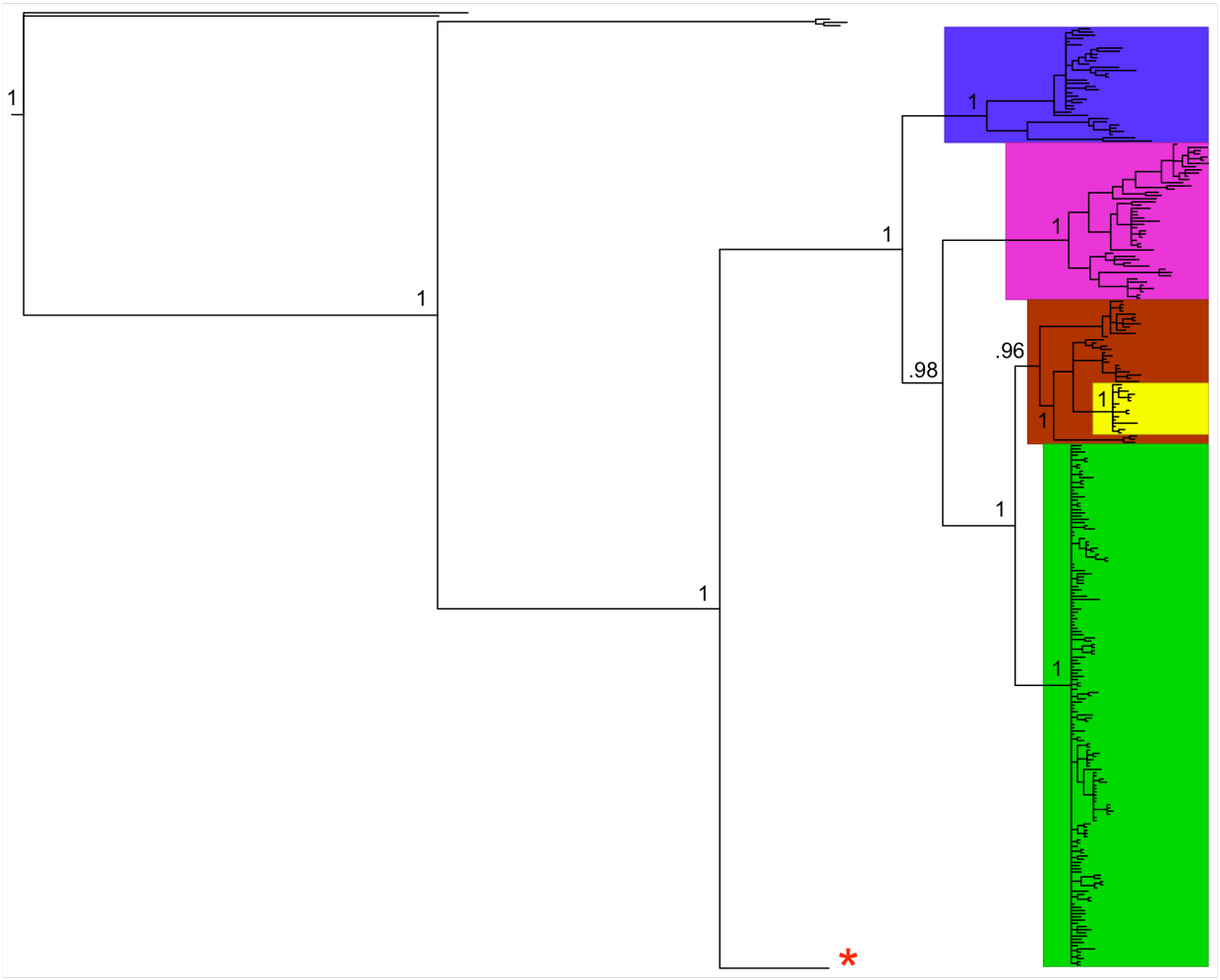

Figure 2. Majority consensus tree from a phylogenetic analysis conducted in MrBayes3.2 on the mitochondrial NADH-2 region (N=299). Outgroups (in phylogenetic order from the root) are *A. isolepis* and *A. altitudinalis*, and *A. porcatus*. The five major mitochondrial lineages of *A. carolinensis* are highlighted by colored boxes: blue – Suwannee; magenta – Southern Florida including the Everglades and South Bay; brown – Central Florida; yellow – North Carolina; green – Gulf-Atlantic. Posterior support for important nodes is shown. Red asterisk indicates one individual collected in Miami by T. Hsieh whose NADH-2 haplotype falls out of *A. carolinensis* and closest BLAST match is *A. porcatus*.

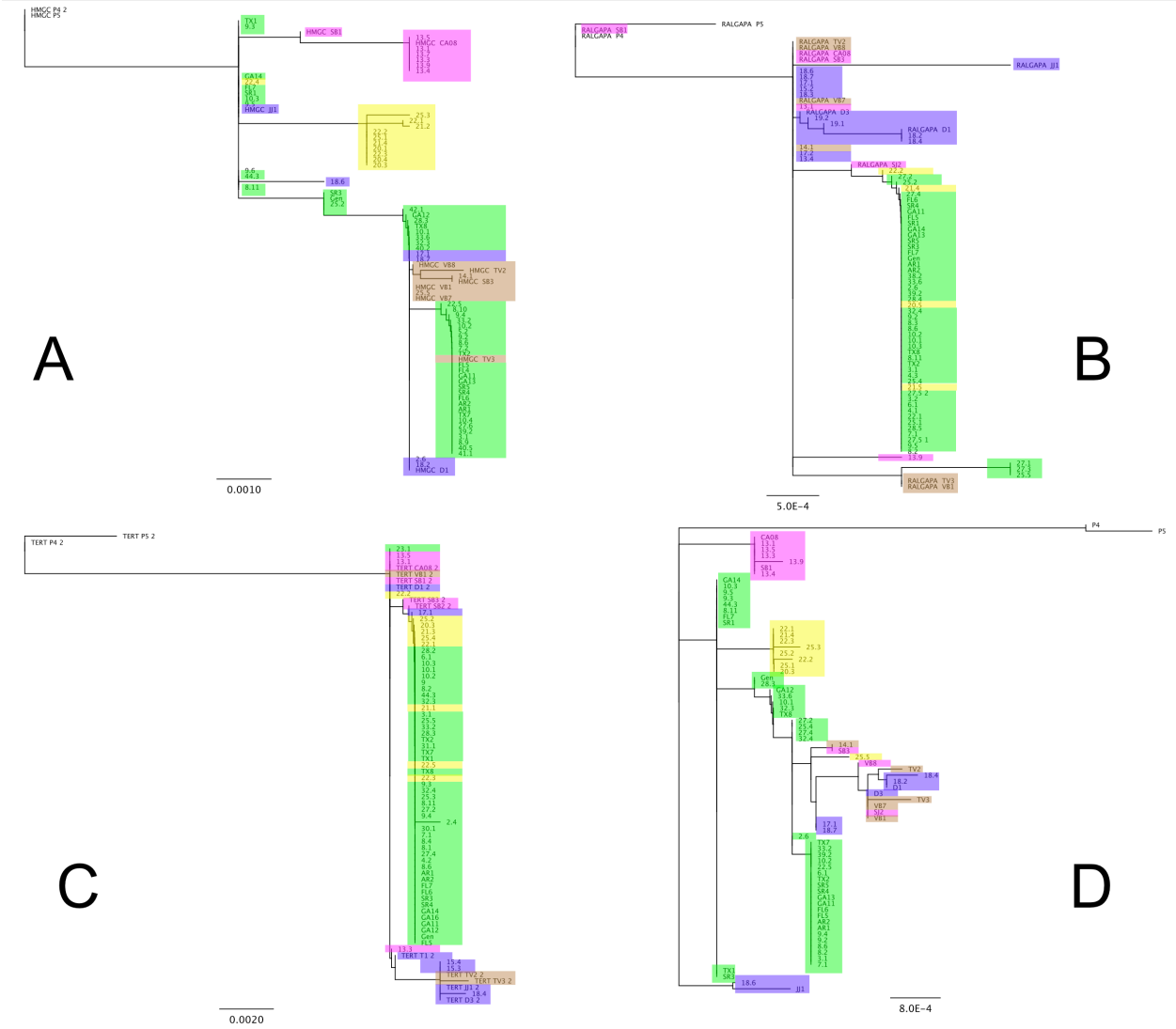

Figure 3. Most likely trees from PHYML Maximum-Likelihood phylogenetic analyses using (A) HMGCS, (B) RALGAPA, (C) TERT, and (D) a concatenated alignment of all three nDNA genes. Colored boxes indicate mitochondrial lineage assignment of each individual. Blue: Suwannee. Magenta: Everglades. Brown: Central Florida. Yellow: North Carolina. Green: Gulf-Atlantic. Across analyses, two individuals (P4 and P5) consistently fall as outgroups. The P5 mtDNA haplotype BLASTs to *A. porcatus* NADH-2 in GenBank.

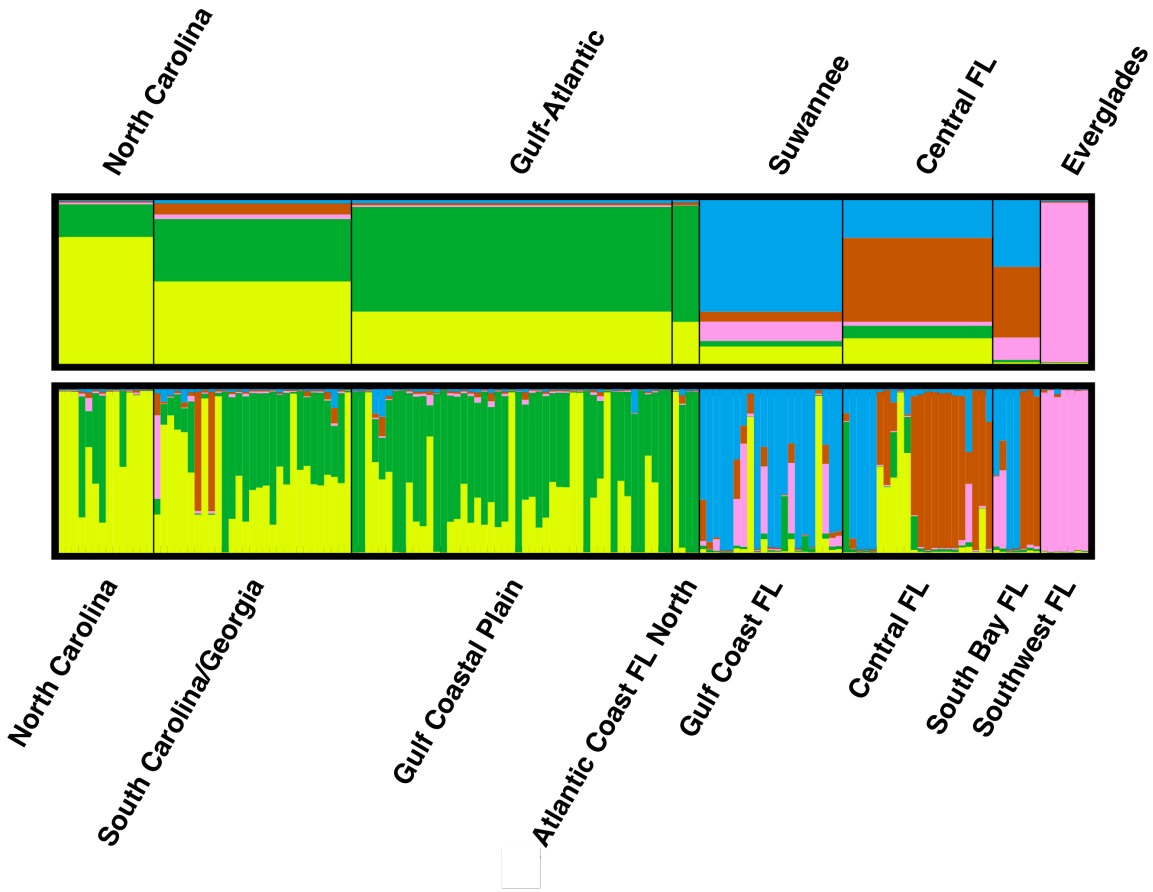

Figure 4. Visualization of the Bayesian clustering analysis from the program *Structure* with the most likely number of clusters K=5 as determined by the ΔK method. Bottom bar: x-axis represents each individual (arranged by geographic region, labeled below) and colors along the y-axis represents the proportion each individual's genome derived from one of the K clusters. Top bar: colored boxes represent the proportion of each of the five nDNA clusters found in each geographic region (corresponding mtDNA clade labeled above). Barplots were produced using the program *DISTRUCT* [217].

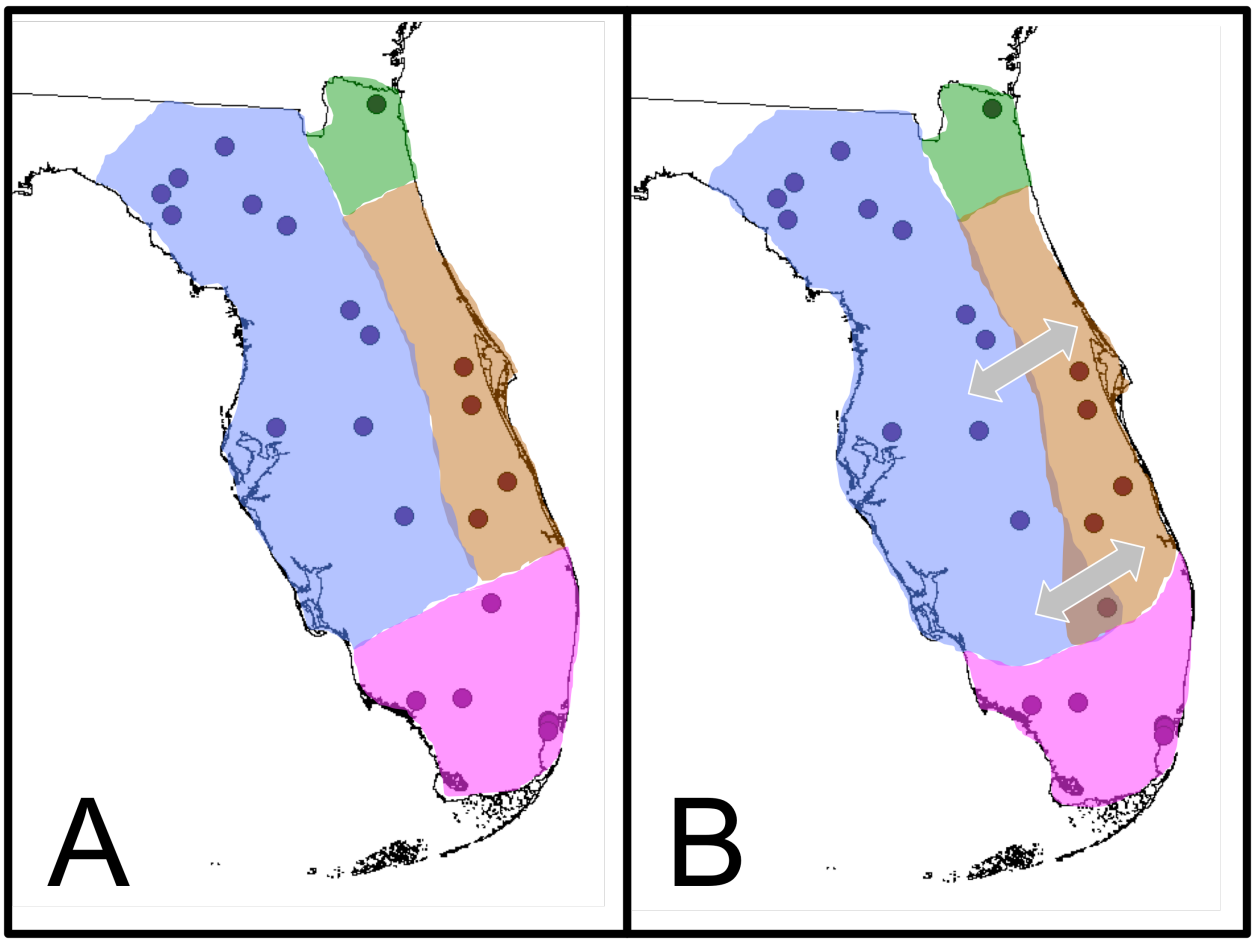

Figure 5. Colored polygons depicting the geographic distribution of mitochondrial lineages (A) and nDNA clusters (B) of green anoles in Florida. Green – Gulf-Atlantic lineage dominates Nassau County newar Jacksonville; Blue – Suwannee; Brown – Central Florida; Magenta – Everglades. Gray arrows indicate admixture between adjacent populations inferred from genetic cluster analysis.

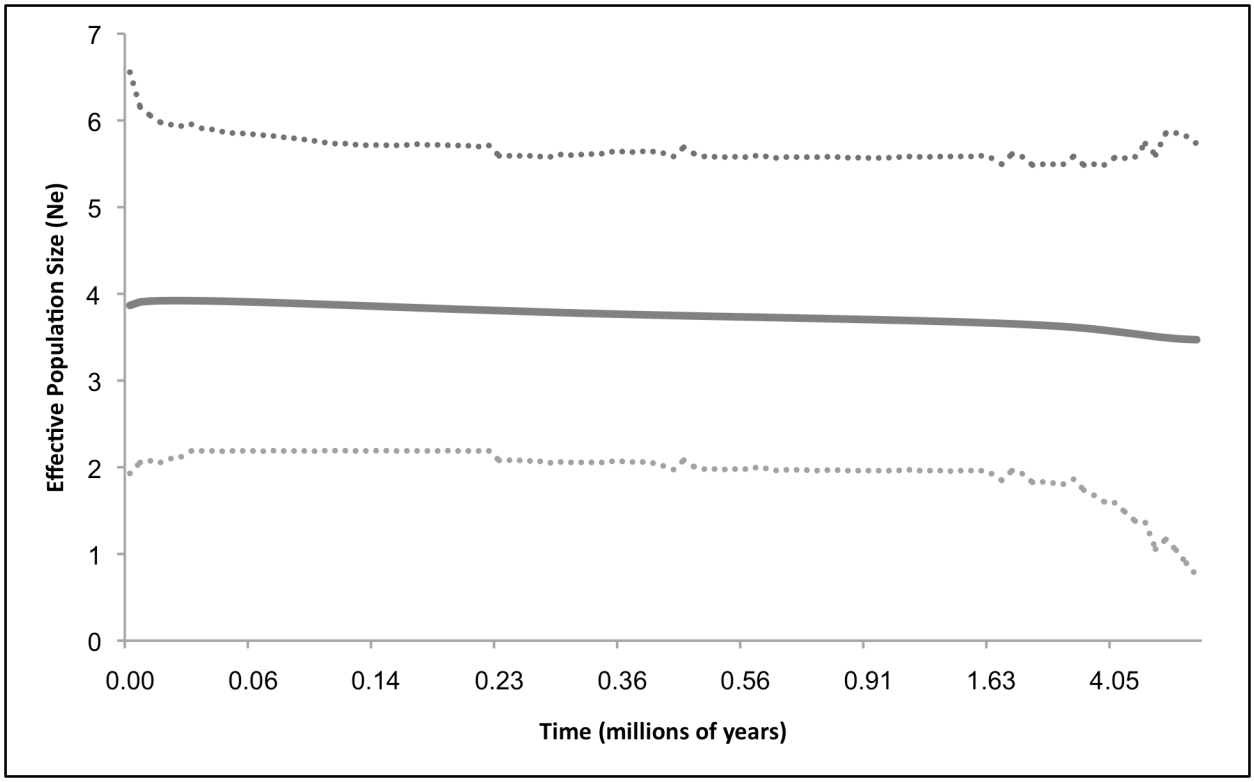

Figure 6. Extended Bayesian Skyline Plot depicting population size changes over time within the Central Florida population. The solid line is the mean estimate from the posterior distribution of population size at time intervals. Dotted lines indicates the upper and lower bounds of the 95% HPD (highest posterior density). The population size function in this analysis is Ne/generation time; for *A. carolinensis* we assume a generation time of 1 year.

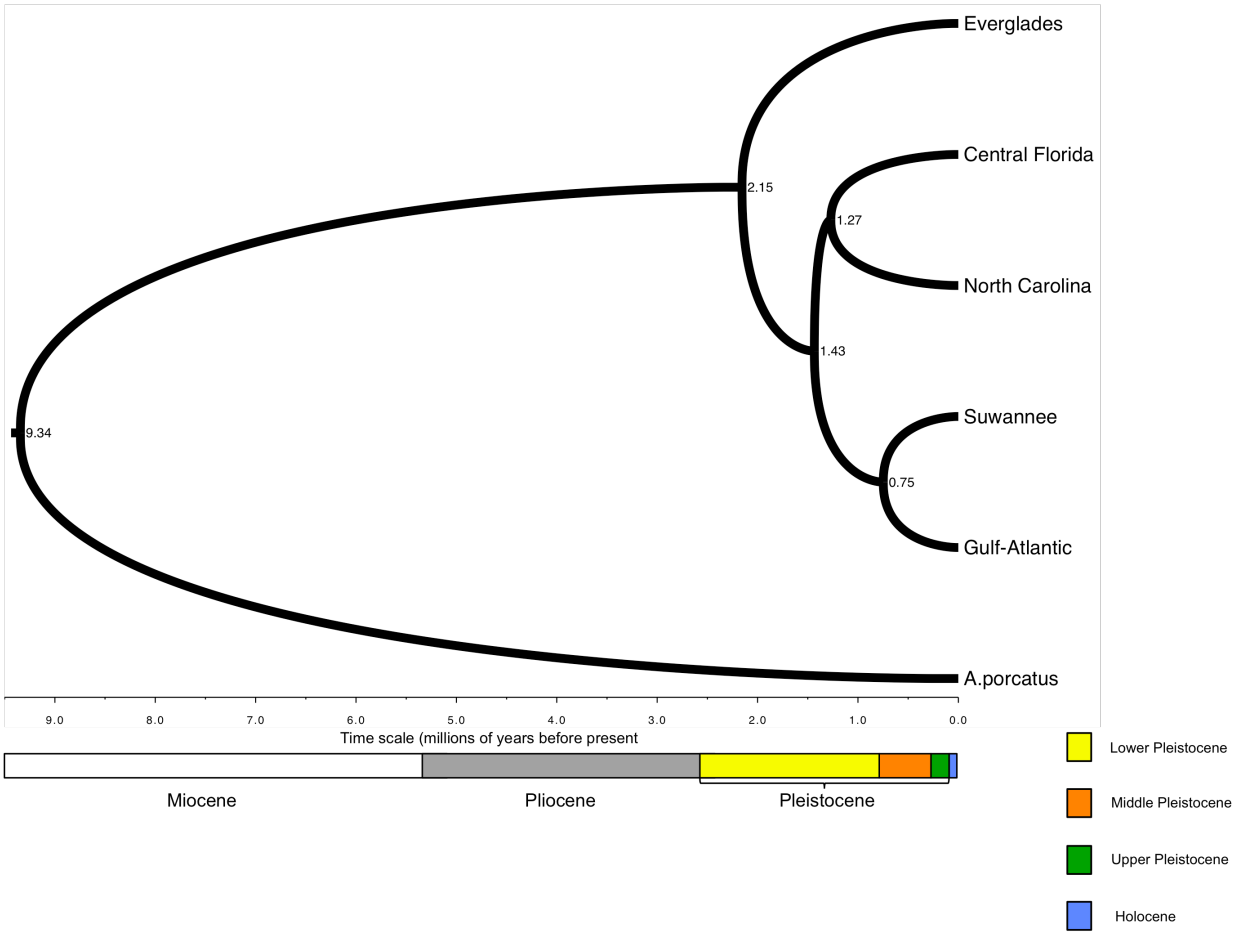

Figure 7. Maximum clade credibility tree from a *BEAST analysis using 1 mtDNA and 3 nDNA loci to show the branching order and divergence times (node values) for the five major lineages of green anoles.

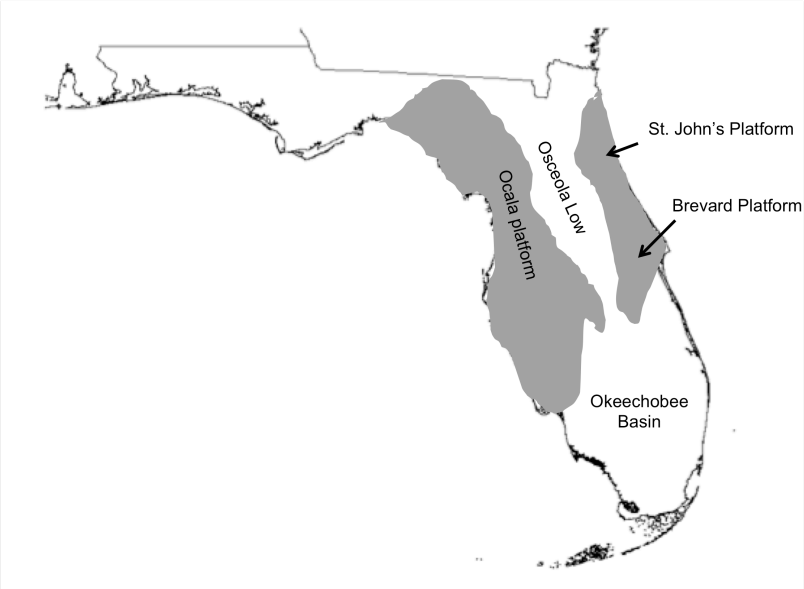

Figure 8. Major geologic structural elements of the Florida peninsula. Positive structures such as the Ocala, St. John's, and Brevard platforms are indicated in grey. Lower elements such as the Osceola Low and the Okeechobee basin are indicated as well. Adapted from Lane (1994) [197].

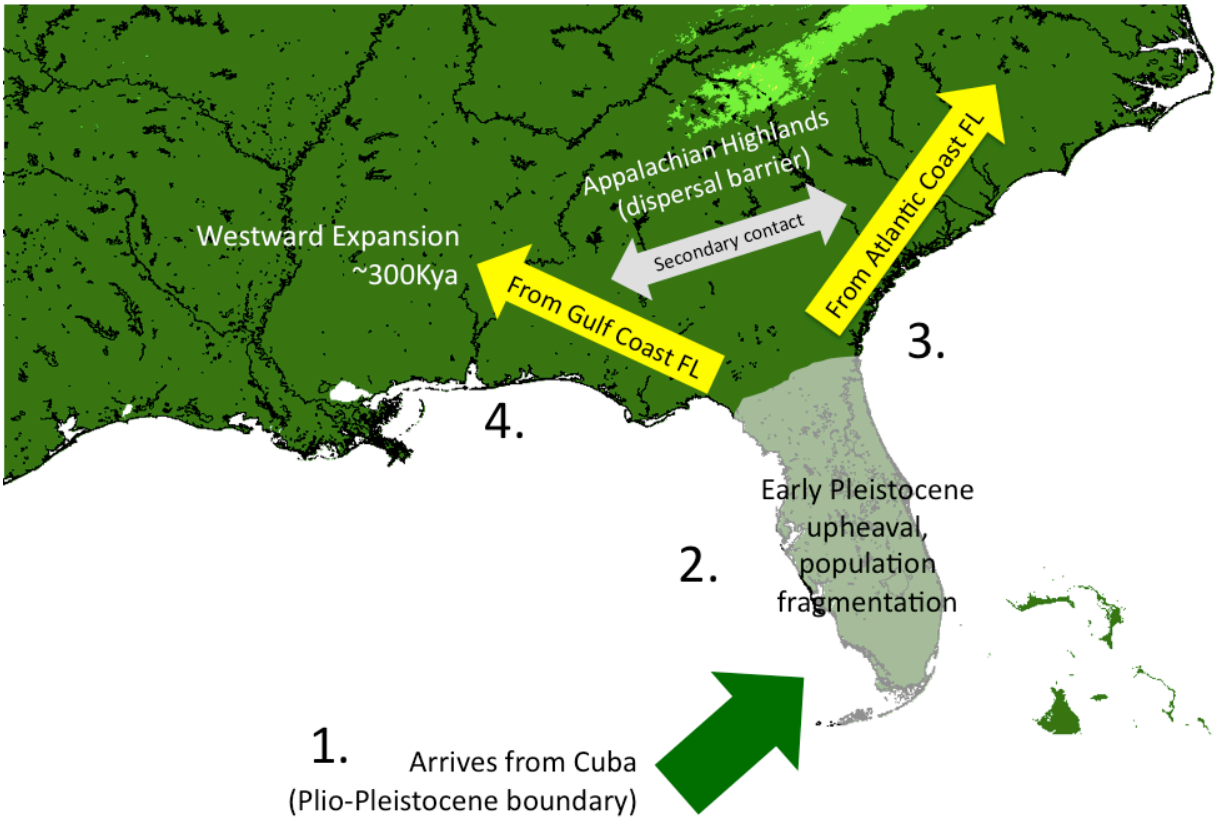

Figure 9. Hypothesized biogeographic scenario depicting the major events in the history of *A. carolinensis* inferred in this chapter. (1) Green anoles disperse from Cuba to the southern tip of Florida between 1.3 and 3.1 million years ago, most likely at the Plio-Pleistocene boundary. (2) During the Lower Pleistocene, physiographic upheaval driven by rising sea levels during glacial cycles cause population fragmentation and divergence. (3) During the late part of the Lower Pleistocene, the peninsula connects to the mainland again, and green anole populations living along the Central Atlantic Coast of Florida are able to disperse northwards towards the Carolinas. High elevations along the Appalachian Highlands (lighter green indicates >750 meters elevation) most likely provided a dispersal barrier westward as unsuitable habitat. (4) During the Middle Pleistocene, populations living along the Gulf Coast of Florida began to colonize watersheds along the Gulf Coastal Plain, resulting in a westward expansion of green anoles across the Mississippi River and into Texas.

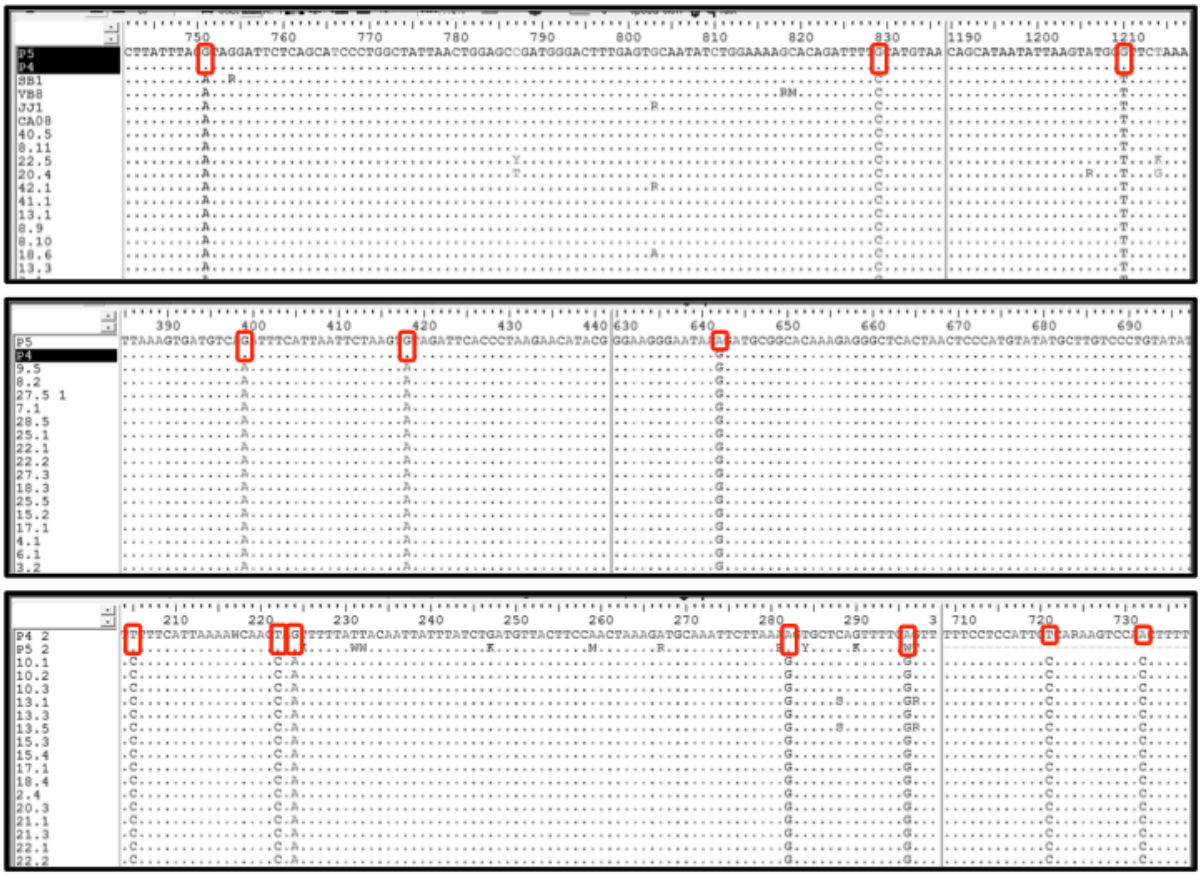

Figure 10. *A. porcatus*-specific SNPs present in two individuals (P4 and P5), indicated by red boxes, within the context of three nDNA alignments (HMGCS, RALGAPA and TERT: top, middle and bottom respectively).

Tables Chapter Three

Location	Description (if applicable)	Latitude	Longitude	Source
East Titusville		28.5437777	-80.942167	BL
East Cocoa		28.243611	-80.870556	BL
West Vero Beach		27.640278	-80.59475	BL
South Bay		26.683333	-80.716884	BL
Lake Josephine		27.366783	-81.400155	BL
Lake Alfred		28.078317	-81.729533	BL
Miami	Coral Gables	25.736883	-80.267133	HL
Miami	Coral Gables Library	25.739417	-80.2654	HL
Miami	Park on SW 40th St	25.734517	-80.274183	HL
Miami	Big Cypress	25.931133	-80.943883	HL
Miami	San Jacinto Park	25.757433	-80.267967	HL
Miami	Alhambra Circle	25.75175	-80.264317	HL
Miami	Across from library	25.738915	-80.265445	HL
Miami	Matheson County Park	25.679317	-80.2734	HL
Tampa	Tampa	28.06165	-82.4164	HL
Deer Island	Deer Island	28.796944	-81.678333	HL
Gainesville	Gainesville	29.651999	-82.324993	HL

Table 1. Information about the Florida collecting localities unique to this chapter. BL: Boissinot Lab; HL: Hsieh Lab.

Accession #	City
JX524368	Jacksonville
JX524369	Jacksonville
JX524370	Jacksonville
JX524371	Jacksonville
JX524372	Jacksonville
JX524373	Chuluota
JX524374	Chuluota
JX524375	Chuluota
JX524376	Chuluota
JX524377	Chuluota
JX524408	Palm Beach
JX524409	Palm Beach
JX524410	Palm Beach
JX524411	Palm Beach
JX524412	Palm Beach
JX524413	Naples
JX524414	Naples
JX524415	Naples
JX524416	Naples
JX524417	Panama City
JX524418	Panama City
JX524419	Panama City
JX524420	Panama City
JX524421	Panama City
JX524422	Panama City

Table 2. GenBank accession numbers and Florida locations for 25 NADH-2 sequences used in this chapter.

Gene	Population	#haps	Hd	π	Θ	D	Fs
HMGCS	Central Florida	16	0.893	0.00499	0.00506	-0.04405	-1.275
	Everglades	2	0.264	0.00021	0.00024	-0.34144	0.186
	Suwannee	6	0.675	0.00262	0.00329	-0.79339	0.728
	North Carolina	4	0.6	0.00114	0.00197	-1.46261	0.907
	Gulf-Atlantic	13	0.753	0.00225	0.00244	-0.22659	-1.187
RALGAPA	Central Florida	8	0.656	0.00086	0.00168	-1.32791	-4.071
	Everglades	6	0.893	0.0014	0.00159	-0.52474	-3.746
	Suwannee	7	0.66	0.00107	0.00227	-1.59816	-2.404
	North Carolina	2	0.356	0.0011	0.00109	0.02107	2.338
	Gulf-Atlantic	3	0.076	0.00023	0.00084	-1.49607	-0.947
TERT	Central Florida	16	0.905	0.00292	0.0048	-1.35152	-7.19
	Everglades	4	0.867	0.00239	0.00242	-0.06042	-0.024
	Suwannee	10	0.863	0.00223	0.00363	-1.41478	-3.683
	North Carolina	2	0.44	0.0004	0.00029	0.84228	0.944
	Gulf-Atlantic	5	0.141	0.00013	0.00074	-1.68341	-5.44
NADH-2	Central Florida	21	0.992	0.03576	0.02862	1.00251	-1.893
	Everglades	12	0.987	0.01684	0.01896	-0.50537	-1.546
	Suwannee	25	0.997	0.02706	0.03702	-1.06254	-5.925
	North Carolina	12	0.958	0.00328	0.00547	-1.61885	-5.61
	Gulf-Atlantic	80	0.978	0.00531	0.02155	**-2.41167**	-112.233

Table 3. Polymorphism overview for each gene within five green anole populations.

#haps – number of haplotypes

Hd – haplotype diversity

π - nucleotide diversity

Θ - Theta per site ($4N_e u$)

D – Tajima's D. Statistically significant values in bold.

Fs – Fu's Fs

Gene	F_{ST}	*Nm*
ND2	0.65285	0.13
HMGCS	0.52482	0.23
RALGAPA	0.39908	0.38
TERT	0.36801	0.43

Table 4. Estimates of genetic differentiation and gene flow across Florida for four genes.

Population 1	Population 2	HMGCS	RALGAPA	TERT	NADH-2
Gulf-Atlantic	Everglades	0.7686	0.55498	0.46658	0.80269
Gulf-Atlantic	Suwannee	0.0825	0.61162	0.62357	0.68086
Gulf-Atlantic	North Carolina	0.57286	0.04871	0.1854	0.86165
Gulf-Atlantic	Central Florida	0.09638	0.68025	0.4702	0.50528
Everglades	Suwannee	0.76766	0.10532	0.28301	0.64174
Everglades	North Carolina	0.86352	0.36216	0.31365	0.81747
Everglades	Central Florida	0.5628	0.17292	0.13067	0.47306
Suwannee	North Carolina	0.60442	0.41168	0.54353	0.73003
Suwannee	Central Florida	0.10604	0.27986	0.1323	0.48819
North Carolina	Central Florida	0.40845	0.42515	0.35287	0.46326

Table 5. Pair wise F_{ST} between each green anole population per gene calculated in DnaSP.

Chapter Four: Population Dynamics of LINE-1 Retrotransposons in the Green Anole Lizard (*Anolis carolinensis)*

Autonomous non-LTR (long terminal repeat) retrotransposons are transposable elements (TEs) that can "copy and paste" themselves via an RNA intermediate, in a process that is enabled by their own reverse transcriptase domain. Non-LTR retrotransposons have proliferated in eukaryote genomes, and they are the main driver of genome size and structural variation between vertebrate lineages [61, 81, 218]. A single type of non-LTR retrotransposon known as LINE-1 (Long Interspersed Nuclear Element, L1 hereafter) dominates the human genome [61]. Inactive and ancient L1 "fossils" and their non-autonomous counterparts, including the *Alu* interspersed repeats, may account for over two-thirds of the human genome [219]. Most human L1 DNA is the result of past amplifications from many millions of years of placental mammalian evolution [21]. The abundance of L1 in the human genome is typical of eutherians [64, 67, 70-72] and accounts for their relatively large genome sizes. In contrast, compact teleost fish genomes contain several active and highly diverse types of non-LTR retrotransposons (sometimes including L1), many of which have produced recent copies; however, they are far less prolific than in mammals [15, 80, 81]. Recent analyses of some squamate reptile genomes [17, 77, 97] have revealed highly divergent repetitive landscapes that are more fish-like than mammal-like. For instance, in the green anole lizard *Anolis carolinensis*, L1 is comprised of 20 families whose divergence predates the amniote ancestor [17] (Figure 1). Within each *Anolis* L1 family, sequence dissimilarity is very low (on average <1%), suggesting that most insertions are recent. The fact that the profile and abundance of non-LTR retrotransposons is similar among non-mammalian vertebrates may mean that mammals have significantly diverged from the amniote ancestor in terms of how they deal with their intragenomic parasites.

Understanding the forces controlling TEs can reveal how genomic profiles diverge between lineages, and to this end models of TE evolutionary dynamics have considered how the selective loss of element-containing loci can be offset by genetic drift in finite populations [102, 220, 221]. The fact that TE copy number is very low in *Drosophila* and that TE insertions segregate as rare alleles in *Drosophila*

populations led researchers to suggest a model in which purifying selection against TEs causes a high turnover of elements and thus prevent the accumulation of copies in the genome [101, 106]. This is in contrast to the human genome, in which the majority of L1 copies seem to be selectively neutral and accumulate readily. However, in both *Drosophila* [222] and human [91], size seems to matter most, as non-LTR retrotransposon insertions that are longer, including those that are full-length, are found at lower population frequencies than those that are truncated. Longer non-LTR retrotransposons also have a tendency to accumulate in non-recombining genomic regions such as Y chromosomes [92, 93, 117]. These observations suggest that longer elements may behave as deleterious alleles due to their ability to mediate ectopic recombination [223]. Another potential factor that may contribute to the deleteriousness of longer elements is the fact that full-length non-LTR retrotransposon copies contain all the intact enzymatic machinery necessary for replication and proliferation, thus leaving the possibility that selection acts against the retrotransposition process itself [224, 225]. Other studies of TE dynamics that have taken into account the demographic histories of host populations have shown that TEs are able to reach higher population frequencies regardless of their selective effect due to random processes such as drift and founder effect. The differential fixation of TE copies in colonized versus native populations of both *Drosophila subobscura* [129] and *Arabidopsis lyrata* [130] strongly suggests that the efficiency of purifying selection against deleterious TE alleles is reduced in populations of small effective size.

While full-length L1 elements in the human genome have been shown to be deleterious [91, 92], the sheer fact of the vast number of elements in the genome suggests that mammals can somehow tolerate the widespread occurrence and fixation of L1. The diverse and ancient L1 families in *Anolis* are comprised of mostly recent insertions, suggesting either a high turnover due to strong purifying selection or a high rate of DNA loss. Both of these factors were invoked to explain the accumulation and decay of non-LTR retrotransposons in the stickleback fish *Gasterosteus aculeatus* [82], in which truncated elements can reach fixation yet full-length elements are rare. It is possible that the similar genomic profiles of fish and reptiles are the result of similar processes. However, there has never been a study of TE frequencies in natural populations of reptiles to address this matter. My purpose in this chapter is to

understand the evolutionary forces controlling L1 in the *Anolis* genome by examining their allelic frequencies in natural populations.

Green anoles are widespread and abundant in the southeastern U.S., and their demographic history across this landscape has recently received attention [195, 196]. In the present study we consider five distinct evolutionary lineages, which shared a common ancestor ~2 million years ago and show evidence of limited gene flow: (1) the Everglades population which is geographically limited to the southern part of the Florida peninsula; (2) the Suwannee population which inhabits the Gulf Coast of the Florida peninsula; (3) the Central Florida population which primarily is restricted to the Atlantic Coast of peninsular Florida; (4) the North Carolina population which exists in that state at the northern limits of the species range along the Atlantic Coast; and (5) the Gulf-Atlantic population, which extends from South Carolina and Georgia but also along the Gulf Coastal Plain and across the Mississippi River into Texas. The oldest populations are found today in Florida, pre-date the Early Pleistocene, and are putatively more stable and at or near equilibrium. A Mid- to Late-Pleistocene dispersal of green anoles into the continental mainland resulted in a dramatic expansion across the Gulf Coastal Plain, and a relatively young refugial population in North Carolina. These insights provide a useful backdrop for understanding the effects of selection and demography on L1 evolution in reptiles. Inferences that have been made about the different demographic histories of these lizards can provide a good test for hypotheses about the alteration of the selection-drift balance in finite populations.

Materials and Methods

Sample Collection and DNA Extraction

This study focused on a sample of 158 green anoles collected across the U.S. states of North Carolina (NC), South Carolina (SC), Georgia (GA), Alabama (AL), Florida (FL), Tennessee (TN), Arkansas (AR), Louisiana (LA) and Texas (TX) between 2009 and 2012. Specimens were caught by hand or noose and tissue samples were taken in the form of tail clippings or, if dissected, muscle or liver, which were preserved in ethanol. Protocols were established in accordance with and approved by the Queens College Institutional Animal Care and Use Committee (Animal Welfare Assurance Number: A32721-01;

protocol number: 135). DNA was extracted from all tissues with the Promega Wizard Genomic DNA Purification kit.

Collection of L1-Containing Loci

The 3' ends and genomic flanking regions of *Anolis* L1 inserts were cloned from each of the five green anole populations: Everglades, Suwannee, Central Florida, Gulf-Atlantic and North Carolina. To obtain L1 inserts that were not present in the database and unique to individuals or populations, and to minimize bias in collecting L1-containg loci, we used the following cloning strategy. For each population, the genomic DNA from five individuals was pooled in equal proportion to obtain ~2 nanograms, the concentration and purity of which was verified using a NanoDrop 2000 spectrophotometer. We digested the pooled DNA samples with NEBNext® dsDNA Fragmentase® in order to obtain randomized genomic fragments of 1kb to 2.5kb, which was verified by electrophoresis on a 1% agarose gel with a Promega Bench Top 1kb DNA ladder. The endonucleases comprising the Fragmentase produce overhangs, and these were polished by incubation at 12ºC for 30 minutes with T4 DNA polymerase followed by heat-inactivation (20 minutes at 75ºC) of the polymerase to produce blunt ends. The 5' hydroxyl groups were phosphorylated by incubation at 37ºC for 30 minutes with T4 polynucleotide kinase (with 5% polyethylene glycol) followed by heat inactivation of the kinase at 75ºC for 20 minutes. The DNA fragments were then ligated to 10uM of double-stranded anchor (5' – TAGCTACAGCTGTAGCTGACAT – 3') with T4 DNA ligase at room temperature for three hours. To ensure that the anchors ligated sufficiently, we performed a PCR using the putatively ligated DNA with the ds anchor as a primer, and checked for DNA smears of appropriate size (1kb to 2.5kb) on a 1% agarose gel.

We then took a series of enrichment steps to capture L1-containing loci from different *Anolis* L1 families using primers designed from L1 family-specific DNA alignments of elements collected from the database (shown in Table 1). The L1 families we focused on were L1AC18 and L1AC20 as described in Novick et al. (2009); the reason we chose these families is because they represented the range of copy numbers found within *Anolis* L1 families; L1AC18 contains 144 copies including 24 that are full-length and 120 that are truncated, and L1AC20 contains 75 copies including 22 that are full-length and 53 that are truncated.

We performed an asymmetrical PCR on the anchor-ligated DNA with a 5 to 1 volumetric ratio of a10uM family-specific L1 biotinylated primer (shown in Table 1) and the 10uM single strand anchor. These PCR products were then captured using streptadivin-coated magnetic beads (M-280 Dynabeads) following the procedure recommended by the manufacturer, after which a second enrichment PCR was performed using bead-captured DNA, the single strand anchor as a primer and a second nested L1 family-specific primer (shown in Table 1). Purified PCR products were ligated into plasmids using a pGEM-T® Easy Vector kit (Promega), and the ligated vectors were transformed into JM109 *E. coli* competent cells. Bacterial colonies were grown overnight on plates with LB agar + ampicillin + IPTG + X-gal and were blue-white screened. Positive clones were picked and incubated overnight in 300uL of SOC media with ampicillin in 96-well plates. We amplified the cloned products by PCR using primers located in the plasmids (Sp6 and T7). Since the goal was to determine whether or not our captured L1 insertions were unique to populations and individuals, we needed enough flanking region that could be mapped to the database. Therefore, our biotinylated and nested primers were designed to be <150 bases from element 3' ends, and we selected inserts that were at least ~500bp in length. The vector primers were used for Sanger sequencing by the HT-Seq facility at the University of Washington, Seattle, WA. Forward and reverse reads for each sequenced clone were assembled into contigs using Geneious v5.5 [160], and their consensus sequences were extracted and used for further analysis. After removing vector sequence, each L1-containing clone with enough flanking region (~50bp) was used in a BLAT search [159] of the May 2010 release (Broad Institute version AnoCar2.0) of the *Anolis* genome on the UCSC Genome Browser [154] (www.genome.ucsc.edu). If the entire query, consisting of an L1 3' end and flanking sequence, could be matched unambiguously to a specific location in the genome, then the insertion site was deemed occupied. A novel insertion was recorded when the BLAT returned a match of only the flank with no upstream L1 3' end, indicating an empty insertion site in the database.

We added to this dataset a collection of L1-containing loci from the February 2007 and May 2010 releases of the *Anolis* genome, both of which are available on the University of California Santa Cruz (UCSC) Genome Browser. We used a consensus sequence query for each *Anolis* L1 family described in Novick et al. (2009) in a BLAT search to retrieve elements from the *Anolis* genome. We aligned the

collected elements to their family consensus sequences and calculated their divergence from family consensus using the Kimura 2-Parameter corrected distance method in MEGA5 [169]. For each insert in the output, we also collected 2500bp of upstream and downstream genomic flank. Flanking regions were submitted to Repeat Masker [158], which screened for single sequence repeats, short tandem repeats, or TEs, which would interfere with PCR primer design. Primers were designed in flanking regions either manually or using Primer3 [156]. For inserts longer than 2kb, we designed family-specific internal primers near the element 3-prime ends from sequence alignments using ClustalW [161]. All primer pairs were tested for specificity using the *in-silico* PCR tool available on the UCSC Genome Browser.

PCR Analysis of L1 Polymorphism

We determined the polymorphism of any novel cloned inserts by using a presence/absence ascertainment with a series of flanking and internal primers. These PCRs were performed on a panel comprised of the individuals whose genomic DNA was originally pooled for the enrichment method. The primers for determining the absence of an element were designed in Primer3 after a BLAT search to locate the insertion site in the *Anolis* genome database and collection of 300bp upstream and downstream of the insertion site. Primers for presence detection were performed with reverse flanking primer and one of the L1 family-specific internal (forward) primers. The specificity of each reaction was verified with the *in-silico* PCR tool on the UCSC Genome Browser. PCRs for presence/absence detection included a 1:00 hold at 94ºC followed by 30 cycles of 0:30 denaturing at 94ºC, 0:30 annealing at 55º-62ºC (depending on the melting temperatures of the primer pairs), and 0:30 extension at 72ºC. To determine the size of these novel elements, we conducted PCRs using genomic DNA from an individual that successfully amplified for element presence with the forward flanking primers and three reverse primers located at various distances from the 5' end of the consensus sequence for each family: 500bp, 1kb, and 2kb (Table 1). These PCRs included a 1:00 hold at 94ºC followed by 30 cycles of 0:30 denaturing at 94º, 0:30 annealing at 55ºC, and 1:30 extension at 72ºC. Successful amplification with these primers allowed us to determine to what extent these novel insertion extended towards their 5' ends.

We also measured the population frequencies of L1 loci retrieved from the database. PCRs were done on an Eppendorf ThermoCycler by the following method: a 1:00 hold at 94ºC followed by 0:30 denaturing at 94ºC, 0:30 annealing at 54-60ºC (depending on the melting temperatures of the primer pairs), and 1:00 extension at 72ºC for 30 cycles, followed by a 4:00 hold at 72º and refrigeration at 4º. Individuals from each green anole population were genotyped according to amplified fragment size after electrophoresis on a 1% agarose gel with ethidium bromide and a Promega BenchTop 1Kb DNA ladder. Insertions were deemed "absent" from their locus when a fragment was amplified equal in length to the expected size from the *in silico* PCR minus the length of the missing element. A short element was deemed "present" when a fragment was amplified equal to the length to the expected size of the element plus the genomic flank. A long element was deemed "present" when a fragment was amplified with the internal primer equal to the expected size of the 3' element end plus the 3' genomic flank.

Population Genetic Analysis

Within each population for each locus, we recorded the total number of present and absent insertions, and population frequencies were calculated as the number of present alleles divided by the number of total chromosomes. We also examined the population frequencies of elements that differ by length categories. To determine if purifying selection is acting against FL insertions in green anole populations, we compared the frequency distribution of FL and TR L1 elements. For this purpose, elements extending all the way from their 3' to 5' ends were counted as FL, while those missing >10% of their 5' ends were counted as TR. Using the Wilcoxon Rank-Sum Test (Mann-Whitney U-Test), we aimed to detect statistically significant differences in allele frequencies between TR and FL loci both within and between populations. We used the Kolmogorov-Smirnov Test to determine if the shape of the frequency distributions between the two insertion types is significantly different.

Results

We collected L1-containing loci from two sources: the *Anolis* genome database and through the direct cloning of inserts from the genomic DNA of individuals. The reasoning behind this two-pronged approach was to minimize ascertainment bias. Since the database was constructed from the sequencing of a single

individual, it may be less likely to contain low-frequency polymorphisms, which are integral to any study of purifying selection. Therefore, the cloning afforded us the opportunity to more closely approximate the amount of genetic variation in natural populations. The five green anole populations we studied for this chapter are treated as distinct entities, and we measured the allelic frequencies of L1 loci within each population separately. This is because it has been shown in previous chapters of this dissertation that these five populations constitute independently evolving lineages with minimal gene flow between them.

All of the L1-containing clones that we were able to map to the database are shown in Table 2 (for L1AC18) and Table 3 (for L1AC20), and the results from the cloning experiments are summarized in Table 5. We sequenced 480 clones, and identified 380 L1 insertions. Using BLAT, we were able to unambiguously identify 265 flanking regions that could be mapped onto the *Anolis* genome database. 47 of these represented insertion sites we sequenced more than once because they were captured multiple times, thus we captured 218 unique L1 insertion sites. Of these, we identified 148 elements from the L1AC18 family and 70 from the L1AC20 family, representing respectively 100.2% and 93% of the copy number estimates of these families from Novick et al. (2009). The remaining cloned L1 either did not contain enough flanking region to allow the determination of the insertion site, or contained repetitive DNA in the flank and thus their insertion sites were ambiguous. Of the 218 unique L1 insertion sites found in the database, 51 (23%) were not occupied by an L1 element. These elements were probably not present in the individual who was sequenced for the Anole Genome Project and are most likely polymorphic in green anole populations. The polymorphism data and the status of novel full-length insertions in green anole populations are also given in Table 5. We were able to successfully measure the polymorphism for 28 of 35 (80%) novel insertion loci from the L1AC18 family and 15 of 18 (83%) novel insertion loci from the L1AC20 family (primers shown in Table 4). We were able to successfully ascertain the size of 18 of 28 (64%) novel L1AC18 inserts, of which 9 were full-length and 9 truncated, and 6 of 15 (40%) novel L1AC20 inserts, of which 2 were full-length and 4 were truncated.

The results from the survey of L1 polymorphism using insertion loci from the database are summarized in Table 6. Three of the truncated insertion loci designed from the database were also captured by our

cloning method, which was not an unexpected result since with that method we were able to retrieve a high proportion of the total copy numbers of the studied L1 families. These loci were L1AC18_128 and L1AC18_223 from L1 family L1AC18, which were fixed across all populations, and L1AC20_150 from L1 family L1AC20, which ranged in population frequency from 88% to total fixation. The high population frequency of these elements is not surprising since they were retrieved from multiple populations during the cloning. Overall, were able to collect population frequency data on 52 insertion loci from 16 of the 20 *Anolis* L1 families described in Novick et al. (2009), including 22 full-length and 30 truncated insertions.

Fixation of L1 Does Occur in Green Anole Populations

Many L1 inserts were fixed in green anole populations. For instance, 22 out of the 30 truncated insertions (73%) collected from the database have reached fixation in at least one of the five green anole populations, and to 5 out of the 22 full-length insertions (23%). The widespread presence of fixed insertions was surprising, because L1 copy number is very low in the *Anolis* genome and ancient insertions are extremely rare. Therefore, we decided to estimate the number of fixed L1 insertions in the genome. To do this, we first looked at the frequencies of L1 inserts with varying levels of divergence in each population. In an attempt to remove the potentially confounding effects of demographic history, we focused at first on only the Gulf-Atlantic and North Carolina populations (Figure 2A and B). This is because the individual that was sequenced for the Anole Genome Project was collected in Aiken, SC and is an admixed individual whose genome is derived from both of these populations. In the Gulf-Atlantic and North Carolina populations, 70% and 66% of L1 inserts that diverge from their consensus by more than 1% are fixed, respectively. This does not mean that only old elements become fixed; the fraction of elements that are both fixed and younger than 1% divergent is somewhat smaller – 10% in the Gulf-Atlantic and 34% in North Carolina – which suggests that at least some elements can reach fixation rather quickly. From the divergence curve, we multiplied the proportion of total elements that are fixed by the 1,006 total L1 copies in the *Anolis* genome reported by Novick et al. (2009) and estimated the total number of fixed inserts to be 342 in the Gulf-Atlantic population and 482 in North Carolina. While these numbers do not comprise a majority of the L1 repertoire in the *Anolis* genome, they do amount to a significant proportion of fixed elements.

It is possible that the unique demographic histories of these populations may be affecting the number of L1 inserts that become fixed. The Gulf-Atlantic and North Carolina populations are relatively young when compared to their conspecifics living on the Florida peninsula, having most likely colonized the Gulf-Coastal Plain and the Mid-Atlantic Seaboard during the Early or Middle Pleistocene. Both of these populations are characterized by lower amounts of genetic variation when compared to their Floridian counterparts, which suggests lower effective population sizes. A small population size can affect the efficiency of purifying selection, and can cause otherwise harmful alleles to drift towards fixation regardless of their selective effect. Therefore, we decided to look at the number of fixed elements in the Central Florida population (Figure 2C), which has been associated with long-term stability since at least the Early Pleistocene and harbors the greatest amount of genetic diversity, suggesting a larger effective population size. We found that 62% of elements diverging from their consensus by more than 1% are fixed in this population, and 11% of elements younger than 1% divergent are fixed as well. This translates into an estimated 335 total fixed L1 inserts in the Central Florida population, which is still an appreciable amount of fixed elements, but is lower than what was estimated for the Gulf-Atlantic population even more so than the North Carolina population.

Purifying Selection Against Full-Length Elements

Our estimates of the amount of fixed L1 elements in green anole populations suggest that non-LTR retrotransposons may be more likely to accumulate in the *Anolis* genome than has been previously suggested [17, 218]. However, this does not necessarily mean that all L1 insertions are selectively neutral. Figure 3 shows the proportion of inserts in each population that are either fixed or polymorphic according to whether they are full-length or truncated. The figure shows that in all populations, truncated elements are much more likely to be fixed than full-length elements, and, conversely, full-length elements are much more likely to be polymorphic than truncated ones. That a much larger proportion of truncated inserts are fixed suggests that full-length elements are prevented from reaching fixation, and perhaps this is because they are subjected to stronger purifying selection. However, 16 full-length insertions (73%) were completely absent in at least one green anole population, compared to 8 truncated (27%), including

16 out of the 22 full-length inserts we screened in the Everglades population. It is difficult to say if purifying selection keeps these inserts at such low population frequencies that we failed to detect them in our sample. Another explanation is that they may have recently inserted into the host genome, sometime after the split in the population histories. Since we learned in previous chapters in this dissertation that the Everglades lineage probably split off from the rest of the species ~2 million years ago, it is perhaps more plausible to assume this latter situation. Therefore, in order to detect purifying selection within each population, we excluded all loci for which we failed to detect presence on a single chromosome. This should not prevent us from detecting selection, since full-length and truncated inserts are generated by the same biological mechanism, and any bias against low frequency alleles will similarly shift the frequency distribution for inserts that are both full-length and truncated.

Table 8 gives the total number of truncated and full-length insertions compared within each population, their average population frequencies, and the statistical significances of the differences in the population frequency means and of the shapes of their distributions. The average frequency of truncated elements was higher than the average frequency of full-length elements in all populations, however this difference was not statistically significant in the Everglades and North Carolina populations. The statistical significance of the difference in population frequency between full-length and truncated elements was significant in the Gulf-Atlantic population ($p<0.01$), and highly significant in the Suwannee and Central Florida populations ($p<0.001$). The shapes of the frequency distributions between truncated and full-length elements were significantly different in all populations that were tested except for North Carolina. The shape of the distribution could not be estimated for full-length elements in the Everglades, because the number of full-length elements in this population was too small to draw any conclusions.

Demographic History Affects the Frequency Distribution of Elements

It is possible that if purifying selection is acting against full-length elements, its efficiency may be different across populations if the effective population sizes of those populations are different. We found that novel full-length inserts were found at low population frequency (<50%) in most instances (Table 5), with two exceptions. One of these exceptions was a greater than 50% population frequency for one of the novel

full-length L1AC20 elements from the Suwannee population. However, this insert was polymorphic (i.e., not fixed) and it is only one of four novel full-length elements cloned from this population, the rest of which were very rare. The other exception was that all of the novel full-length elements from the Gulf-Atlantic population were found at greater than 50% population frequency (Table 6). Two of these elements are completely fixed. This is more likely to be a significant deviation from our earlier conclusion that purifying selection is keeping full-length elements at low population frequencies. Therefore, we decided to use the full data set to compare the frequencies of truncated and full-length insertions within and between each population. As inferred from earlier chapters in this dissertation, the five populations of *A. carolinensis* each have unique demographic histories, and any differential fixation of L1 alleles as a result of this will be able to tell us something about how genetic drift can mediate the fixation of TEs in the genomes of reptiles. The Everglades, Suwannee and Central Florida populations are the oldest green anole populations, they are the most demographically stable, and by every measure contain relatively high neutral genetic diversity ([195], also see Chapter 3 of this dissertation); all of these aspects are associated with large effective population sizes. In contrast, the North Carolina population was estimated to have the smallest effective population size, while the Gulf-Atlantic population experienced a recent and rapid expansion across the continental mainland [195]. This aspect of the Gulf-Atlantic population's history is interesting because there is reason to believe that range expansions promote stronger genetic drift [226, 227].

We found that while the frequencies of truncated elements are not significantly different between any of the populations, the frequencies of full-length elements are highly significantly different between populations of starkly different demographic histories (Table 9). Figure 4 shows the frequency distributions of full-length and truncated L1 elements in each green anole population. From this it is evident that full-length inserts segregate very differently in the Florida populations versus the mainland populations. For instance, the proportion of full-length inserts below 50% population frequency is 83% in the Everglades, 76% in Suwannee, and 80% in Central Florida. Full-length L1 elements are much more common in the mainland populations, with only 29% below 50% population frequency in both North Carolina the Gulf-Atlantic. Since these two populations have either a small estimated effective population

size or have recently experienced a dramatic range expansion, it is likely that relaxed purifying selection due to stronger genetic drift is generating a higher rate of fixation for full-length L1 insertions in these populations.

Discussion

We present here the first study of retrotransposon population dynamics in a reptile, based on a double-sided approach: we collected L1 inserts directly from the genomic DNA of individuals via cloning, and we developed population genetic markers from the *Anolis* genome database. We have three main conclusions: (1) L1 elements are able to reach fixation in *Anolis* more readily than previously thought; (2) truncated elements are more likely to accumulate in the *Anolis* genome while full-length elements are subjected to purifying selection and thus do not accumulate; and (3) the efficiency of purifying selection to remove full-length elements is highly dependent on the demographic history of the population, such that full-length elements are more likely to be fixed in populations of small effective size. Thus, the selective turn over model as it applies to TEs in *Drosophila* cannot fully explain the L1 profile of *Anolis*. In fact, the L1 profile in *Anolis* is remarkably similar to the non-LTR retrotransposon landscape in stickleback fish, in which the accumulation of truncated insertions is offset by a high rate of DNA loss and full-length elements are subjected to purifying selection [82]. Our results are significant as they add to the limited population genetic evidence that L1 evolves differently in mammals than they do in non-mammalian vertebrates.

The question still remains whether or not our conclusions are affected by an ascertainment bias towards middle or high frequency insertions, due to the fact that our database-generated markers result from the sequencing a single individual [228]. This could conceivably skew our estimation of the number of fixed elements in the *Anolis* genome as well as the certainty with which we could detect purifying selection. Indeed, within the Florida populations, novel L1 insertions were found at low population frequencies (Table 5), which might suggest that using the database caused us to miss rare alleles. Yet the frequency distribution of all elements, including those retrieved from the database, shows an over-abundance of rare inserts in Florida. In addition, all of the novel cloned L1 inserts we collected from the Gulf-Atlantic

population were either at very high population frequency (>50%) or were fixed, suggesting that the genetic variation we captured with this method closely mirrors what is in the database. The database was not more likely to yield fixed inserts than the cloning, as our PCR presence/absence study of cloned novel insertions was able to retrieve some elements that were fixed (10%). Therefore, our conclusion that a significant amount of L1 has reached fixation in *Anolis* is accurate and supported by a more unbiased assessment. Even if we were able to completely remove all bias and sample more rare alleles, the frequency distribution of full-length elements, which are rare as suggested by our data, would still be shifted towards zero and would not change the fact that many truncated insertions are fixed; this would actually strengthen our conclusion that purifying selection is acting against full-length elements.

Compared to the human genome, L1 in *Anolis* is relatively low in copy number, and the few elements that are found in the genome are of very recent age [17]. These features of the non-LTR retrotransposon landscape in *Anolis* are reminiscent of what is found in the teleost fish genomes that have been studied so far [15, 16, 80, 82, 229]. To explain these observations in teleost fish, it was originally proposed that TE accumulation was prevented by a high rate of turnover [16] in which the insertion of new elements is offset by the selective loss of insertions, and it was hypothesized that many TEs would exist in populations at low frequencies. This model was initially proposed for and supported by studies of TE dynamics in *Drosophila* [101]. However, the turnover hypothesis was rejected when it was tested in two teleost fish models: stickleback [82] and pufferfish [229]. In pufferfish, the majority of the non-LTR retrotransposon insertions that were studied were found at middle or high population frequencies, which suggests that these elements are not subjected to strong purifying selection. This was a surprising finding since the pufferfish genome is so devoid of non-LTR retrotransposons. However, this analysis only looked at short elements and therefore it may have been biased towards neutral or nearly neutral alleles. In stickleback, all full-length insertions were polymorphic, which suggests that purifying selection acts preferentially against full-length elements in this genome, while a large number of truncated insertions were fixed in populations.

We found that 38 out of 43 novel cloned inserts and 44 out of 52 database-recovered L1-containing loci were polymorphic in at least one green anole population, which is a significant amount of polymorphism that is greater than, for instance, what was observed in the Ta-1 family of L1 inserts of the human genome (86% versus 69%, respectively) [21]. However, five of the novel cloned loci (12%) were fixed in their population of origin, and eight (15%) of the loci from the database were fixed in every population – and therefore in the entire species – which suggests that L1 is quite capable of reaching fixation in the *Anolis* genome. This widespread fixation of L1 elements suggests that, as in stickleback, the turnover model cannot explain the scarcity and young age of L1 elements found in *Anolis*. An alternative explanation would be that L1 has no effect on host fitness, which would be consistent with the conclusion of selective neutrality arrived at by Neafsey et al. (2004) when they found common or fixed non-LTR retrotransposons in their pufferfish population study. We find here that the vast majority of L1 elements that are fixed in *Anolis* are truncated, and that truncated insertions make up the vast majority of older elements. This suggests that at least short L1 insertions may behave as neutral alleles, which would be consistent with the fact that in both *Drosophila* [116] and human [91] truncated elements seem to be neutral or at least are under much weaker selection.

Universal neutrality of L1 in *Anolis* is an unlikely scenario, however, because the data suggest that some elements are subjected to purifying selection. Full-length elements are rare within all *Anolis* L1 families, comprising about 18% of all L1 in the genome [17], and within all natural populations they are found at lower population frequencies relative to truncated elements. The scarcity of full-length elements in *Anolis* is similar to what was found in a study of teleost fish genomes that included zebrafish, Medaka, stickleback and pufferfish [230], and their low frequencies in green anole populations is reminiscent of stickleback as well [82]. This suggests not only that the *Anolis* genome is similar to fish in its autonomous non-LTR retrotransposon repertoire, but also that a similar mechanism is preventing the fixation of full-length elements in all of these genomes. As similar patterns of element decay was reported in stickleback and *Anolis* [17, 82], it is possible that a high rate of DNA loss could account for the scarcity of fixed full-length elements found in both fish and reptiles. However, large DNA deletions would also remove truncated insertions at the same rate, and we now have evidence that truncated elements do become

fixed, therefore it is more likely that the turnover model actually does apply to *Anolis* – but only to full-length elements.

Element length was reported to be the main driver of purifying selection against non-LTR retrotransposons in both *Drosophila*, [117, 222] human [91], and stickleback [82] and the patterns we are reporting for *Anolis* are consistent with that. In both *Drosophila* and human, longer elements are probably more likely than truncated ones to be involved in ectopic recombination, which can cause extremely deleterious chromosomal breaks [223]. Another line of evidence used to support the ectopic recombination model in *Drosophila* and human was that full-length elements accumulate only in genomic regions that are non-recombining. In fact, it has been proposed that an overall low rate of ectopic recombination rate may be a factor that has allowed mammalian genomes to be more tolerant of significant L1 accumulation [94]. However, recombination rates are not yet known in the *Anolis* genome or for reptiles in general, so we cannot rule out a mechanism of purifying selection against L1 other than ectopic recombination.

Another possibility could be that purifying selection acts against full-length elements because, since they contain the open reading frames and promoter necessary for autonomous retrotransposition, there may be a deleterious effect of this process itself [224, 225]. It is clear from our study that in *Anolis* full-length elements are limited not only in genomic copy number but also population frequency; these factors would undoubtedly act to reduce the number of active copies capable of retrotransposition. The mouse genome contains 2,000 – 3,000 potentially active L1 copies [231], which is in stark contrast to the ~90 *Anolis* L1 copies that contain both ORFs and are therefore potentially active [17]. The human genome contains 80 – 100 potentially active L1 copies [232], yet it seems that even though purifying selection acts against full-length elements in the human genome, it is obviously not strong enough to prevent accumulation of active copies. This alone may be enough to explain the differences in L1 copy number between mammals and reptiles. Still a third possibility would be that if full-length elements, which are potentially active, are at very low frequencies in populations, then the transposition rate would be lower than in genomes with more common active elements. The overall result of this would be a relatively low copy number of

elements, which is the case in reptiles and fish. Regardless of the mechanism, the low population frequencies of full-length L1 inserts, especially in conjunction with the fact that the only old and fixed inserts are truncated, strongly suggest that purifying selection is limiting the ability of full-length elements to become fixed in the *Anolis* genome.

Whether or not full-length or truncated L1 elements are being subjected to varying degrees of purifying selection, all TEs are parasites that proliferate within a host genome, and they are therefore dependent on the evolutionary history of their host. Because of this, any change in the effective population size (N_e) of the host will modify the equilibrium between natural selection and genetic drift, which can significantly alter the fate of TEs in the genome. When N_e is large, selection will be more efficient at removing harmful alleles. When N_e is small, otherwise harmful alleles can drift to higher population frequencies. It is therefore reasonable to expect a higher rate of TE fixation in populations that have experienced a bottleneck or founder effect, and this was indeed observed in *Arabidopsis lyrata* [130] and *Drosophila subobscura* [129]. We have also found this to be the case in *Anolis*, as the five populations we have studied all have very different demographic histories that have had a profound influence on the way L1 elements segregate. Full-length elements are much more likely to be fixed in the Gulf-Atlantic and North Carolina green anole populations, both of which meet the criteria for a scenario where genetic drift will be stronger than purifying selection. It is also worth noting that we estimated the largest number of fixed insertions to be in the North Carolina population, which has the smallest effective population size out of the five green anole lineages.

The different fixation rates of full-length L1 insertions in green anole populations with different demographic histories show us how important genetic drift can be for genomic evolution. For instance, if a full-length element is purged from a population via purifying selection, it will be unable to produce new copies. This may result in the removal of harmful alleles, but it might also be the case that the species will potentially lose a source of genetic variation that throughout the history of life, particularly in reptiles [89, 90] has been co-opted in adaptive ways [233]. In a landmark paper, Lynch and Conery (2003) [1] suggested that the origins of eukaryote genome complexity might be a direct result of the shift in the

selection-drift balance that occurred during the evolution of smaller effective population sizes. Therefore, the waxing and waning of effective population sizes, and the fixation or loss of "junk DNA" including TEs, that can occur during lineage diversification can have far-reaching consequences. It leaves the intriguing possibility that large differences in effective population size that may be intrinsic to mammals, reptiles, fish, and insects, respectively, could account for lineage-specific patterns of TE evolution.

While purifying selection seems to be limiting the number of full-length elements, truncated insertions do accumulate readily in the *Anolis* genome. However, there also is a complete absence of L1 insertions that are anywhere near the order of divergence that is typical of some L1 families in the human genome, which can be up to 30% [22]. Thus, unlike the human genome, L1 elements are removed from the *Anolis* genome soon after they become fixed. Novick et al. (2009) analyzed the decay of non-LTR retrotransposons of the RTE clade and reported that large-scale deletions account for the heavily fragmented copies of this group of insert [17]. A similar pattern was found in the Expander non-LTR retrotransposon clade in the stickleback genome [82]. In both of these cases, these elements were much more fragmented than human L1 insertions of similar age, suggesting that DNA loss in the form of large deletions is counteracting the accumulation of retrotransposon copies in fish and reptiles, thus limiting the expansion of the sizes of these genomes. Large deletions are rare in mammals, which may account for the large size of mammalian genomes, and it is possible that DNA loss is a major factor controlling genome size and structure than previously thought. This is a controversial subject, as Petrov (2002) [234] found that small deletions are more common in small insect genomes, and suggested that large deletions are probably too deleterious to be common. However, this may apply only to the compact genomes of insects, as the larger genomes of most vertebrates contain vast intergenic regions that could possibly suffer large deletions without consequence.

Conclusions

We have provided here the first study of TE evolutionary dynamics in reptiles. Contrary to earlier suggestions in which strong purifying selection limits the accumulation of non-LTR retrotransposons in the *Anolis* genome, we find that the L1 retrotransposon actually accumulates readily in this genome. By

studying the population frequencies of L1 inserts collected by direct cloning from genomic DNA and by marker design from the genomic database, we found that truncated L1 insertions are very often fixed in green anole populations, and some appear to be fixed across the entire species. This suggests that short elements behave neutrally in populations and may have little to no effect on host fitness. In contrast, full-length inserts are rare in green anole populations, suggesting that purifying selection is at least acting on long L1 elements. The deleteriousness of full-length L1 elements may stem from their ability to mediate ectopic recombination or their potential for retrotransposition activity. We also found that the demographic history of populations is an important factor that affects the strength of selection against full-length elements. By comparing the frequency spectrum of L1 elements by length in different populations, we found that full-length elements are found at significantly higher frequencies in populations of small effective size where genetic drift is likely to be very strong. Meanwhile, full-length elements are found at significantly lower frequencies in populations of large effective size and long demographic stability, suggesting purifying selection is much more efficient at removing harmful alleles in these populations. The deleterious effect of full-length elements does not appear to completely prevent fixation of L1 elements, yet there are very few ancient elements in the *Anolis* genome. Therefore, we suggest that DNA loss plays a major role in removing L1 insertions after they become fixed. This interplay of selection, demography, and large-scale deletions may account for the differences between the high-copy number L1 profile of mammalian genomes and the low-copy number profile of the genomes of non-mammalian vertebrates.

Figures Chapter Four

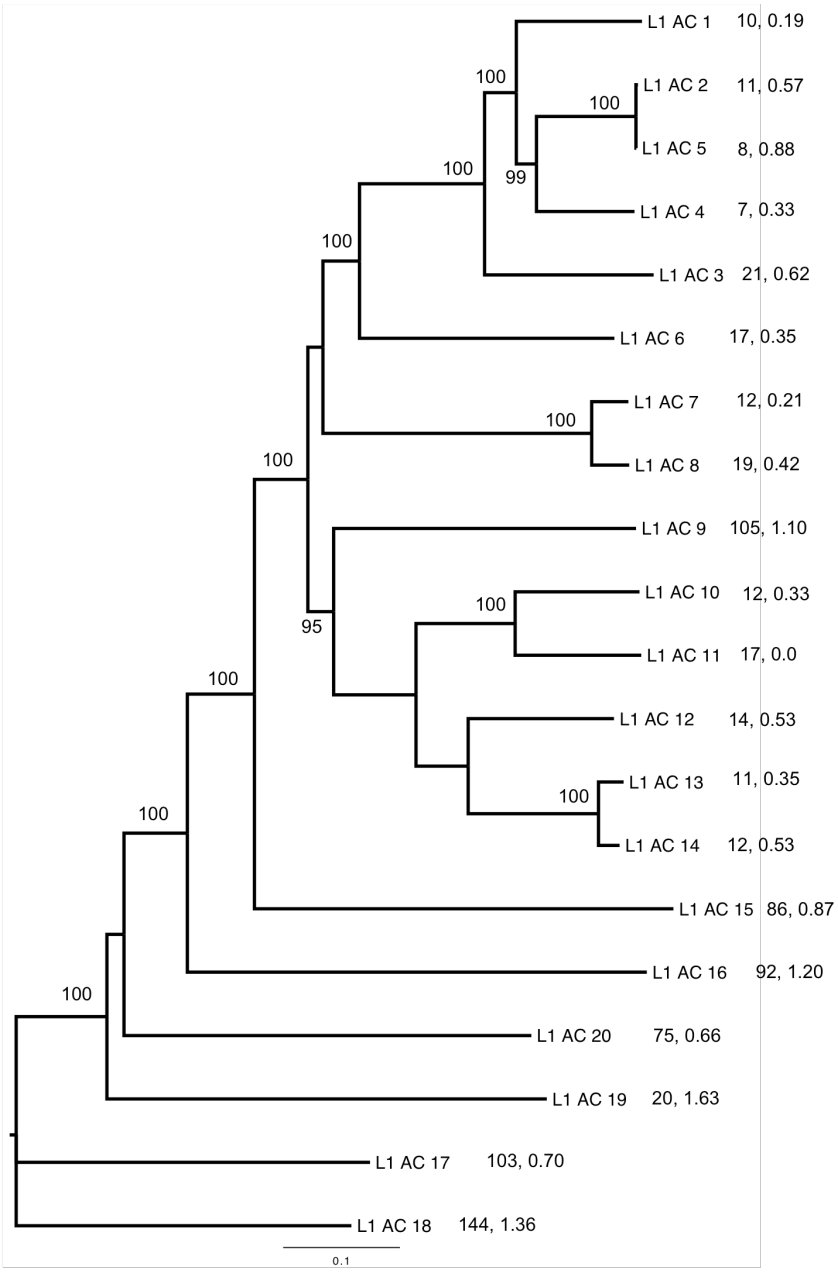

Figure 1. Phylogenetic tree showing the relationships between the ORF2 consensus sequences of the 20 L1 families in the *Anolis* genome constructed with the Neighbor-Joining method. Node support was assessed with 1000 bootstrap replicates (only 95% or greater is shown). Tips are labeled with the L1 family name, copy number, and percent pair wise divergence from consensus sequence as reported in Novick et al. (2009) [17]

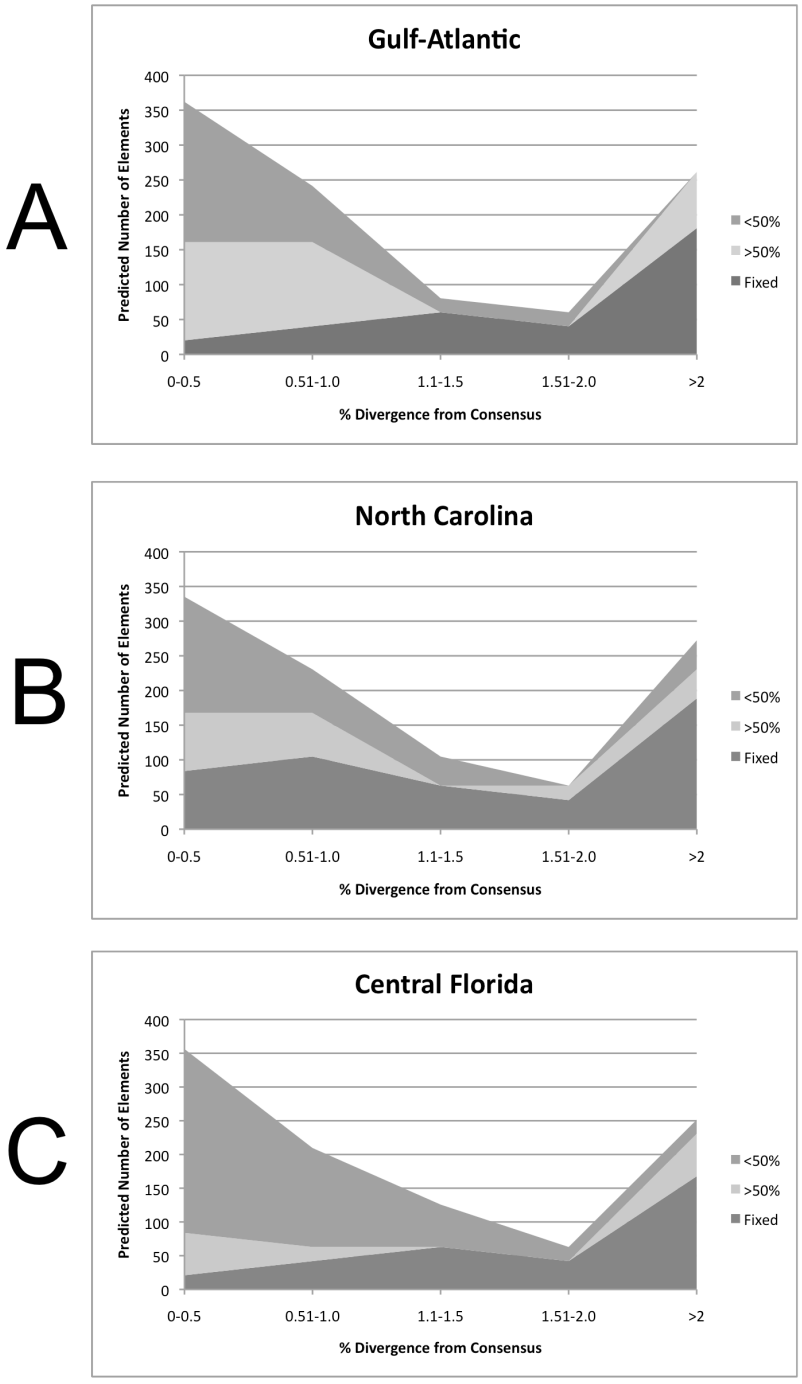

Figure 2. Fraction of fixed and polymorphic L1 elements extrapolated from population data according to their divergence from consensus. The analysis was done separately for the Gulf-Atlantic (A), North Carolina (B) and Central Florida (C).

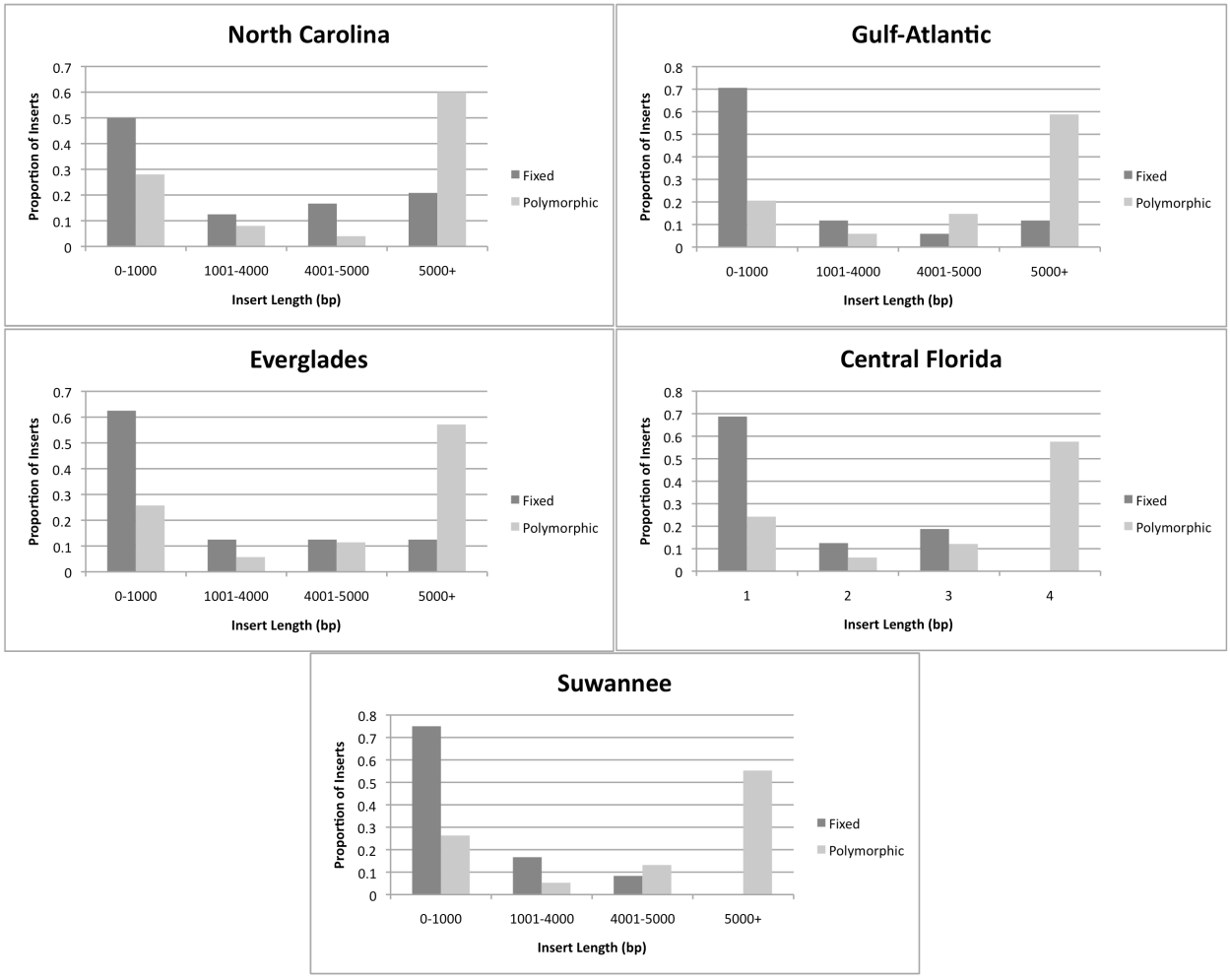

Figure 3. Fraction of polymorphic and fixed L1 elements according to their length in each of five green anole populations. The distribution is based on 52 L1-containing loci retrieved from the *Anolis* genome database.

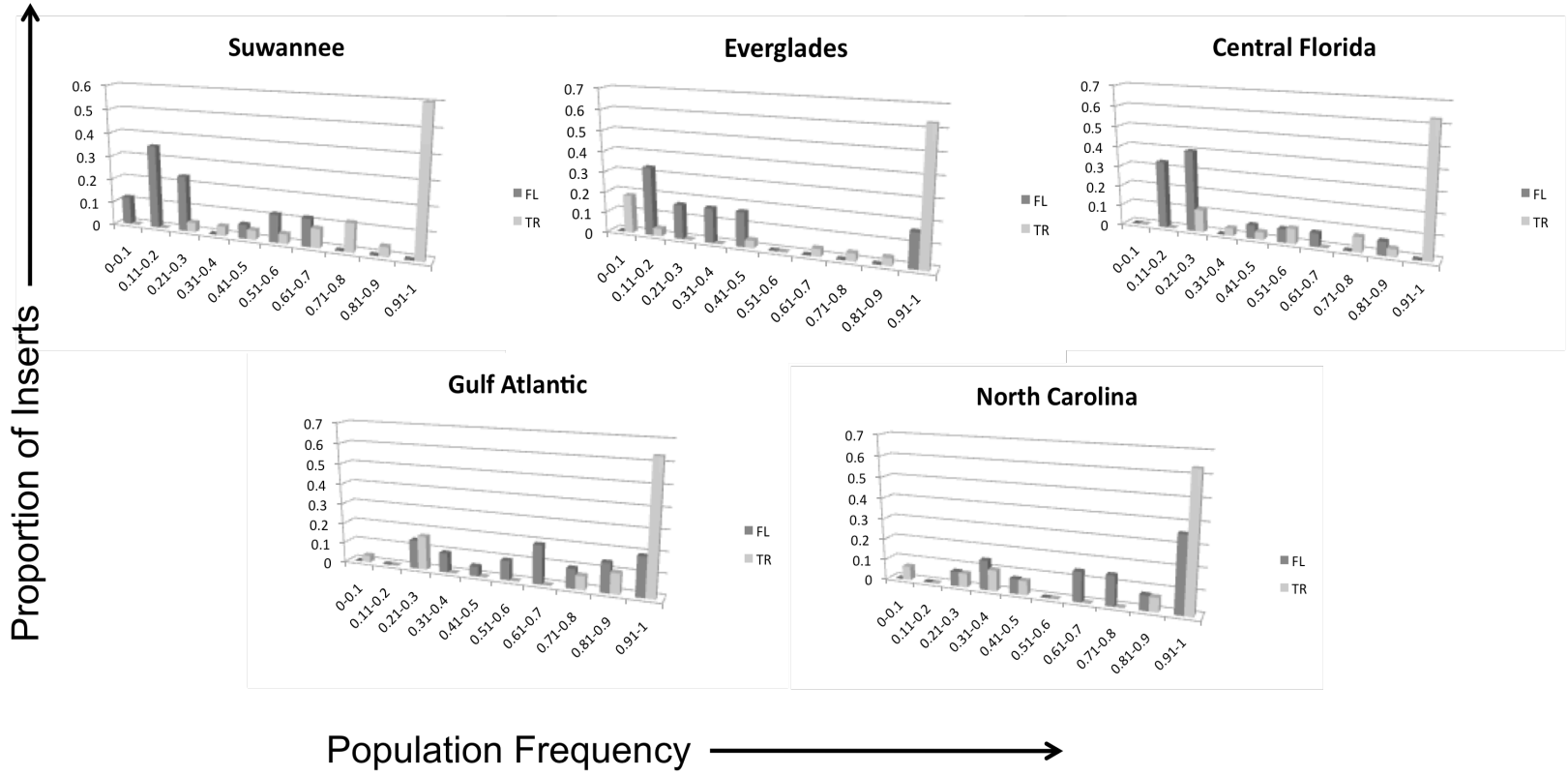

Figure 4. The frequency distributions of full-length (FL) and truncated (TR) L1 elements in five green anole populations.

Tables Chapter Four

Table 1. Primers used for the cloning experiment, including the family-specific biotinylated and nested primers for the enrichment of L1 copies from the L1AC18 and L1AC20 families, and the reverse primers used for size ascertainment.

PRIMER NAME	OLIGO
L1AC18_bio	[BioTEG]TCTGGATCAAGCCATAAG
L1AC18_nest	CAATCCTCGTGTTCGTGG
L1AC18-500r	GGTCAGCAATCTTGTTTAGGAG
L1AC18-1kr	GAAGAGTGTCCTTGGCCTTC
L1AC18-2kr	AGTCCAAATTGAAATCGCCTGC
L1AC20_bio	[BioTEG]CTGGGAAAAATGGAGGAG
L1AC20_nest	TACCTGTGGCTGGATCTGCAAG
L1AC20-600r	TTCTGTGGTGAAGGCCATTA
L1AC20-1kr	AGTCCTTGTTCATTTTAATGGTC
L1AC20-2kr	CTTGCTGAACCCCTATTCTC

Table 2. Cloned L1 inserts from the L1AC18 family that were mapped to the *Anolis* genome.

L1 family	Ligation	Clone (including redundancies)	Population of origin	BLAT coordinates	Status	Size (if empty)
L1-18	1a	1aA11, 1aD11, 1aG7	Everglades	chr1:146415805-146416268	empty	
L1-18	1a	C7	Everglades	chr1:191652357-191652558	occupied	
L1-18	1a	F7	Everglades	chr1:34647095-35075276	empty	
L1-18	1a	A12	Everglades	chr1:75979544-75980487	empty	
L1-18	1a	F8	Everglades	chr1:91926637-91926881	occupied	
L1-18	1a	A7	Everglades	chr2:63269338-63269503	empty	
L1-18	1a	E9	Everglades	chr4:38939682-38940221	occupied	
L1-18	1a	E7, E8	Everglades	chr5:14515642-14516383	empty	
L1-18	1a	A10	Everglades	chr6:53123671-53124162	occupied	
L1-18	1a	F10	Everglades	chr6:88044-88580	empty	
L1-18	1a	B8	Everglades	chrUn_AAWZ02036194:15559-16329	empty	
L1-18	1a	G7	Everglades	chrUn_GL343193:2345208-2345470	occupied	
L1-18	1a	C10	Everglades	chrUn_GL343273:1322407-1323353	empty	
L1-18	1a	B12	Everglades	chrUn_GL343276:125528-125756	occupied	
L1-18	1a	C9	Everglades	chrUn_GL343506:200729-201622	occupied	
L1-18	1a	G8	Everglades	chrUn_GL343701:16274-16655	occupied	
L1-18	1a	B7	Everglades	chrUn_GL343798:55818-56011	occupied	
L1-18	B	F2	CFL	chr2:114208285-114208596	occupied	
L1-18	B	F6	CFL	chr3:188128953-188129040	empty	440
L1-18	B	H7	CFL	chr3:86289985-86290138	empty	443
L1-18	B	E8	CFL	chr6:53123839-53124162	occupied	
L1-18	B	H8	CFL	chrUn_GL343311:1459088-1459243	occupied	
L1-18	B	E12	CFL	chrUn_GL343383:508467-508651	occupied	
L1-18	B	B_E6, B_F1	CFL	chrUn_GL343722:244004-244184	occupied	
L1-18	B	B_G4, J_F1	CFL, Everglades	chr4:34279515-34279907	occupied	
L1-18	B	B_G11, J_D9	CFL, Everglades	chrUn_GL343328:1007484-1007669	occupied	
L1-18	C	A9	CFL	chr1:166172711-166172965	occupied	
L1-18	C	C_A1, C_C4	CFL	chr1:218063991-218064201	occupied	
L1-18	C	H3	CFL	chr1:223912745-223913088	occupied	
L1-18	C	D7	CFL	chr1:233213970-233214234	empty	769
L1-18	C	C3	CFL	chr1:238867833-238868163	occupied	
L1-18	C	B1	CFL	chr2:122246425-122246710	occupied	
L1-18	C	F9	CFL	chr2:45161703-45161839	occupied	
L1-18	C	D10	CFL	chr3:131478411-131478587	occupied	
L1-18	C	E6	CFL	chr3:148956552-148956681	empty	457
L1-18	C	C_B10, C_G3	CFL	chr3:154981433-154981619	occupied	
L1-18	C	H7	CFL	chr5:83332091-83332221	occupied	
L1-18	C	A5	CFL	chrUn_AAWZ02038834:1524-1779	empty	494
L1-18	C	G12	CFL	chrUn_GL343222:2611093-2611207	occupied	
L1-18	C	G2	CFL	chrUn_GL343232:1084448-1084762	occupied	

L1-18	C	D2	CFL	chrUn_GL343322:51045-51155	occupied	
L1-18	C	H5	CFL	chrUn_GL343415:549214-549491	occupied	
L1-18	C	H11	CFL	chrUn_GL343418:136199-136694	occupied	
L1-18	C	G9	CFL	chrUn_GL343428:810250-810386	occupied	
L1-18	C	G6	CFL	chrUn_GL343465:462139-462404	occupied	
L1-18	C	B4	CFL	chrUn_GL343477:385470-385632	occupied	
L1-18	C	B_G7, C_A7	CFL	chrUn_GL343536:170897-171167	occupied	
L1-18	C	E4	CFL	chrUn_GL344022:38357-38496	occupied	
L1-18	C	C_E11, J_G9	CFL, Everglades	chrUn_GL343276:61652-61868	occupied	
L1-18	C	C_G11, J_B5	CFL, Everglades	chrUn_GL343383:508467-508651	occupied	
L1-18	C	C_E1, J_D2	CFL, Everglades	chrUn_GL343762:37525-37629	occupied	
L1-18	C	C_C8, D_H12	CFL, Gulf-Atlantic	chr5:83332091-83332450	occupied	
L1-18	C	C_H2, H_C9	CFL, NC	chrUn_GL343798:55758-56011	occupied	
L1-18	C	C_E9, F_D7	CFL, Suwannee	chrUn_GL343292:499007-499156	occupied	
L1-18	E		Gulf-Atlantic	chr1:101851410-101851624	occupied	
L1-18	E	G11	Gulf-Atlantic	chr1:10925004-10925205	occupied	
L1-18	E	G8	Gulf-Atlantic	chr1:128474959-128475600	occupied	
L1-18	E	G1	Gulf-Atlantic	chr1:191652401-191652558	occupied	
L1-18	E	A9	Gulf-Atlantic	chr2:102527962-102528184	occupied	
L1-18	E	D9	Gulf-Atlantic	chr2:150519974-150520379	empty	717
L1-18	E	H9	Gulf-Atlantic	chr3:119177018-119177280	occupied	
L1-18	E	A2	Gulf-Atlantic	chr3:126206105-126206544	occupied	
L1-18	E	D10	Gulf-Atlantic	chr3:49716213-49716900	empty	797
L1-18	E	E_E5, E_E7	Gulf-Atlantic	chr3:72865281-72865576	empty	548
L1-18	E	E3	Gulf-Atlantic	chr4:39273081-39273240	empty	517
L1-18	E	E_A3, E_H12	Gulf-Atlantic	chr6:61133461-61133640	occupied	
L1-18	E	D2	Gulf-Atlantic	chrLGa:1300945-1301107	occupied	
L1-18	E	E8	Gulf-Atlantic	chrUn_GL343276:125475-125756	empty	515
L1-18	E	E_C8, G_E8	Gulf-Atlantic	chrUn_GL343500:484876-485077	occupied	
L1-18	E	B4	Gulf-Atlantic	chrUn_GL343526:304106-304343	empty	628
L1-18	E	D11	Gulf-Atlantic	chrUn_GL343552:559875-560039	empty	541
L1-18	E	E10	Gulf-Atlantic	chrUn_GL343634:267720-267968	occupied	
L1-18	E	B7	Gulf-Atlantic	chrUn_GL343665:152717-152929	occupied	
L1-18	E	H2	Gulf-Atlantic	chrUn_GL343690:4194-4405	occupied	
L1-18	E	F1	Gulf-Atlantic	chrUn_GL343829:126047-126383	occupied	
L1-18	E	D1	Gulf-Atlantic	chrUn_GL343948:5508-5786	empty	1089
L1-18	E	B10	Gulf-Atlantic	chrUn_GL344250:61780-62248	occupied	
L1-18	E	E_B6, J_D10	Gulf-Atlantic, Everglades	chrUn_GL343386:1058578-1058834	occupied	
L1-18	E	E_H8, G_C6	Gulf-Atlantic, Suwannee	chrUn_GL343208:3955052-3955296	occupied	
L1-18	G	C_B11, E_A10, G_E6	CFL, Gulf-Atlantic, Suwannee	chrUn_GL343209:3588849-3589011	occupied	
L1-18	G	C_F3, G_A10	CFL, Suwannee	chrUn_GL343360:139131-139578	occupied	
L1-18	G	B4	Suwannee	chr1:123802111-123802231	empty	544
L1-18	G	A9	Suwannee	chr1:230890739-230891006	occupied	
L1-18	G	D3	Suwannee	chr1:65533301-65533471	occupied	
L1-18	G	H6	Suwannee	chr1:69198826-69199312	occupied	
L1-18	G	F10	Suwannee	chr1:75883738-75883987	occupied	

L1-18	G	E9	Suwannee	chr2:162513653-162513865	occupied	
L1-18	G	E2	Suwannee	chr3:138645454-138645642	occupied	
L1-18	G	A8	Suwannee	chr3:94885171-94885344	occupied	
L1-18	G	H2	Suwannee	chr4:117739557-117739711	empty	511
L1-18	G	G_A5, G_C1	Suwannee	chr4:75707815-75708068	occupied	
L1-18	G	A4	Suwannee	chrUn_AAWZ02038860:1606-1928	occupied	
L1-18	G	G4	Suwannee	chrUn_GL343277:173687-173973	occupied	
L1-18	G	D4	Suwannee	chrUn_GL343416:674492-674967	occupied	
L1-18	G	B2	Suwannee	chrUn_GL343533:68315-68521	occupied	
L1-18	G	C7	Suwannee	chrUn_GL344606:9891-10093	occupied	
L1-18	G	D11	Suwannee	chrUn_GL344675:6834-7163	occupied	
L1-18	G	G_G9, J_C4	Suwannee, Everglades	chr1:213363770-213363973	occupied	
L1-18	H	D_D2, H_H11	Gulf-Atlantic, NC	chr6:68562576-68562835	occupied	
L1-18	H	F7	NC	chr1:181524430-181524679	occupied	
L1-18	H	H_B12, H_F2	NC	chr2:118048037-118048296	occupied	
L1-18	H	E4	NC	chr3:101766522-101766834	occupied	
L1-18	H	C_D8, H_E6	NC	chr4:16000553-16000902	occupied	
L1-18	H	B8	NC	chr6:941887-942071	occupied	
L1-18	H	A1	NC	chrUn_GL343193:4907440-4907660	occupied	
L1-18	H	A10	NC	chrUn_GL343215:1870401-1870665	empty	
L1-18	H	F10	NC	chrUn_GL343245:1516554-1516656	empty	657
L1-18	H	B2	NC	chrUn_GL343252:680415-680573	occupied	
L1-18	H	H3	NC	chrUn_GL343330:534376-534687	occupied	
L1-18	H	C12	NC	chrUn_GL343352:780026-780263	occupied	
L1-18	H	H1	NC	chrUn_GL343352:780111-780263	occupied	
L1-18	H	D5	NC	chrUn_GL343514:385573-385841	occupied	
L1-18	H	A9	NC	chrUn_GL343554:346423-346668	occupied	
L1-18	H	F12	NC	chrUn_GL343755:51089-51678	empty	704
L1-18	H	H_B11, C_F6	NC, CFL	chr1:74996004-74996506	empty	452
L1-18	H	H_A8, J_C10	NC, Everglades	chr4:55957352-55957506	occupied	
L1-18	H	H_C4, J_A5	NC, Everglades	chrUn_GL343924:16467-16664	empty	611
L1-18	H	B_H6, E_B8, H_B6	NC, Gulf-Atlantic, CFL	chr2:22265314-22265661	occupied	
L1-18	H	G_E3, H_C10	Suwannee, NC	chrUn_GL343578:199594-199822	occupied	
L1-18	J	C_B5, J_F2	CFL, Everglades	chrUn_GL343549:168043-168262	occupied	
L1-18	J	J_C8, J_C6	Everglades	chr1:146415810-146416268	empty	988
L1-18	J	A2	Everglades	chr1:165799580-165799817	occupied	
L1-18	J	C4	Everglades	chr1:213363770-213364060	occupied	
L1-18	J	B8	Everglades	chr1:90830260-90830810	occupied	
L1-18	J	C11	Everglades	chr2:106412030-106412551	occupied	
L1-18	J	D5	Everglades	chr2:136876928-136877794	occupied	
L1-18	J	A7	Everglades	chr2:172988986-172989500	occupied	
L1-18	J	C12	Everglades	chr2:182546700-182546763	empty	564
L1-18	J	H2	Everglades	chr2:28722731-28722938	empty	544
L1-18	J	H3	Everglades	chr3:148806244-148806731	occupied	
L1-18	J	E3	Everglades	chr3:37285687-37286738	occupied	
L1-18	J	F7	Everglades	chr4:13709019-13709536	occupied	

L1-18	J	J_B1, J_E8	Everglades	chr4:38937361-38938167	occupied	
L1-18	J	C9	Everglades	chr5:58265145-58265555	occupied	
L1-18	J	E12	Everglades	chrLGb:2522841-2523724	occupied	
L1-18	J	A12	Everglades	chrUn_GL343198:4806039-4806154	empty	608
L1-18	J	H5	Everglades	chrUn_GL343224:880416-880675	occupied	
L1-18	J	D4	Everglades	chrUn_GL343266:144833-144958	empty	601
L1-18	J	F8	Everglades	chrUn_GL343301:1522650-1523416	occupied	
L1-18	J	D9	Everglades	chrUn_GL343328:1007484-1007776	occupied	
L1-18	J	G3	Everglades	chrUn_GL343330:827129-827502	occupied	
L1-18	J	E5	Everglades	chrUn_GL343343:1070314-1071236	occupied	
L1-18	J	C7	Everglades	chrUn_GL343466:71935-72181	occupied	
L1-18	J	B7	Everglades	chrUn_GL343549:168043-168262	occupied	
L1-18	J	H12	Everglades	chrUn_GL343891:116935-117189	occupied	
L1-18	J	G6	Everglades	chrUn_GL343928:78318-79194	occupied	
L1-18	J	J_G11, C_G10	Everglades, CFL	chr4:133385923-133386442	occupied	
L1-18	J	G_E1, J_D7	Suwannee, Everglades	chr6:47527713-47527842	occupied	

Table 3. Cloned inserts from the L1AC20 L1 family that were mapped to the *Anolis* genome.

L1 family	Ligation	Clone (including redundancies)	Population of origin	BLAT coordinates	Status	Size (if empty)
L1-20	1b	F8	Everglades	chr1:150574818-150575116	occupied	
L1-20	1b	D10, A9	Everglades	chr1:95237017-95237315	empty	
L1-20	1b	E1, G8	Everglades	chr2:149195543-149195989	empty	
L1-20	1b	C7	Everglades	chr4:146516441-146516549	empty	
L1-20	1b	D3, C4	Everglades	chrUn_GL343574:296882-297042	empty	
L1-20	1b	1bE3, 1bA2, 1bC2, 1bH7, 1bH8, 1bA9, 1bB10, I_F10, I_A9, I_B10, A_F9	Everglades, CFL	chr3:32972041-32972608	occupied	
L1-20	1b	E6, F2, H1, F_A7, F_G7, I_H1, D_E8	Everglades, Suwannee, Gulf-Atlantic	chr1:222391511-222391857	occupied	
L1-20	A	D1	CFL	chr1:161965497-161965694	occupied	
L1-20	A	G3	CFL	chr1:234254282-234254521	occupied	
L1-20	A	F6	CFL	chr1:77549284-77549479	empty	586
L1-20	A	G5	CFL	chr2:16183814-16184126	occupied	
L1-20	A	F4	CFL	chr2:171974824-171974933	empty	625
L1-20	A	B8	CFL	chr3:159015587-159015763	occupied	
L1-20	A	F12	CFL	chr3:46090152-46090278	empty	521
L1-20	A	C10	CFL	chr3:95074821-95075368	occupied	
L1-20	A	E5	CFL	chr3:99380188-99380519	empty	588
L1-20	A	A4	CFL	chr4:115499464-115499730	occupied	
L1-20	A	B8	CFL	chr4:144613803-144613922	empty	605
L1-20	A	C1	CFL	chr4:148928149-148928469	occupied	
L1-20	A	H4	CFL	chr4:59990294-59990603	occupied	
L1-20	A	F2	CFL	chrUn_AAWZ02041333:4709-4948	occupied	
L1-20	A	C11	CFL	chrUn_AAWZ02041333:4994-5256	occupied	
L1-20	A	B7	CFL	chrUn_GL343225:1698547-1698971	occupied	
L1-20	A	C4	CFL	chrUn_GL343232:891353-891667	occupied	
L1-20	A	H1	CFL	chrUn_GL343240:1963003-1963556	occupied	
L1-20	A	G2	CFL	chrUn_GL343251:616285-616566	occupied	
L1-20	A	G10	CFL	chrUn_GL343292:1178968-1179331	occupied	
L1-20	A	E6	CFL	chrUn_GL343677:274003-274151	empty	709
L1-20	A	C7	CFL	chrUn_GL344371:15733-15980	occupied	
L1-20	A	A_G9, J_C2, I_G12	CFL, Everglades	chrLGd:167865-168078	occupied	

L1-20	A	A_F3, B_E2, D_E3, I_F7	CFL, NC, Gulf-Atlantic, Everglades	chr1:113202526-113202880	occupied	
L1-20	A	A_H3, F_E8	CFL, Suwannee	chr5:68443673-68443998	occupied	
L1-20	A	A_C3, F_D11	CFL, Suwannee	chrUn_GL343322:85144-85463	occupied	
L1-20	B	A_D8, B_B3	CFL, NC	chr3:127753929-127754449	occupied	
L1-20	B	A_G1, B_D3	CFL, NC	chrUn_GL343231:428175-428563	occupied	
L1-20	B	D4	NC	chr3:166770113-166770339	occupied	
L1-20	B	B4	NC	chr4:17464338-17464696	occupied	
L1-20	B	B1	NC	chr4:31790822-31791189	occupied	
L1-20	B	B2	NC	chr4:82090774-82091049	occupied	
L1-20	B	E1	NC	chr6:24611351-24611526	occupied	
L1-20	B	A4	NC	chr6:78880179-78880582	occupied	
L1-20	B	B8	NC	chrUn_GL343259:610124-610428	occupied	
L1-20	B	A9	NC	chrUn_GL344254:59497-59718	occupied	
L1-20	B	B_C7, D_B10	NC, Gulf-Atlantic	chr3:178322464-178322821	occupied	
L1-20	D	A_B6, D_H4	CFL, Gulf-Atlantic	chrUn_AAWZ02038600:2306-2534	empty	471
L1-20	D	D_F7, D_G10	Gulf-Atlantic	chr1:1753251-1753712	occupied	
L1-20	D	C4	Gulf-Atlantic	chr2:183116474-183117140	occupied	
L1-20	D	B11	Gulf-Atlantic	chr3:48204119-48204516	occupied	
L1-20	D	B1	Gulf-Atlantic	chr5:19967365-19967590	occupied	
L1-20	D	C10	Gulf-Atlantic	chrUn_GL343194:5523354-5523548	empty	403
L1-20	D	H3	Gulf-Atlantic	chrUn_GL343195:2124173-2124349	empty	
L1-20	D	A9	Gulf-Atlantic	chrUn_GL343300:99303-99627	occupied	
L1-20	D	D_A11, C_12	Gulf-Atlantic, NC	chr1:160592253-160592603	occupied	
L1-20	D	D_H2, F_H9, A_D6	Gulf-Atlantic, Suwannee, CFL	chr3:96429452-96429699	occupied	
L1-20	F	A_B9, F_D9	CFL, Suwannee	chrUn_GL343344:277685-277951	occupied	
L1-20	F	E9	Suwannee	chr3:10291510-10291805	occupied	
L1-20	F	C7	Suwannee	chr3:179355854-179356160	occupied	
L1-20	F	A11, C12	Suwannee	chr4:71357780-71357806	empty	500
L1-20	F	A10	Suwannee	chr4:94820807-94820926	empty	579
L1-20	F	A9	Suwannee	chrUn_AAWZ02041207:5497-5671	occupied	
L1-20	F	G10	Suwannee	chrUn_GL343391:890789-891281	occupied	
L1-20	F	F7	Suwannee	chrUn_GL343646:206893-207036	empty	508
L1-20	I	C8	Everglades	chr1:175445903-175446233	empty	653
L1-20	I	B2	Everglades	chr1:223632162-223632587	empty	877
L1-20	I	F10	Everglades	chr3:22195728-22195859	occupied	
L1-20	I	A_H10, I_h11	Everglades	chr4:16610458-16610728	occupied	
L1-20	I	D5	Everglades	chrUn_GL343234:2232684-2232808	occupied	
L1-20	I	C4	Everglades	chrUn_GL343574:296882-297052	occupied	
L1-20	I	D_B6, I_F12	Gulf-Atlantic, Everglades	chrUn_GL343195:943169-943361	occupied	

L1-20	A	A_A10, A_F1, B_B11, B_B2, D_C3, D_C8, D_G5, D_C7, D_E11	CFL, NC, Gulf-Atlantic	chrUn_GL343786:54425-54882	occupied	

Table 4. Primers used for presence/absence ascertainment for L1 loci designed from the *Anolis* genome database. Internal primers for long elements contain the letters "INT" in the primer name. For each forward and reverse primer pair, the expected sizes are given in the row of the forward primer. Presence and absence for truncated (TR) elements was determined using the forward and reverse primers. Presence for full-length (FL) elements was determined with the family-specific internal primer.

PRIMER NAME	OLIGO	PRESENCE (TR)	PRESENCE (FL)	ABSENT
L1AC10-INT1	ATTCCAGAACACGCCCTAAG			
L1AC10.2-f	GCAGGTTGGCACACCTATCT		985	1092
L1AC10.2-r	TGTCCAAAAGCGGTGGTAAT			
L1AC11_2:10-F	GCTCCGGATTGTGTTTTTGT		1224	991
L1AC11_2:10-R	CAGCCAATGTAGCACAGGAA			
L1AC11_INT	GATATTGGCGACAAGAAACAATC			
L1AC11s_6:33-F	CAGGGAATCCTGTGTCCAGT	1153		713
L1AC11s_6:33-R	GGATTGCCAGCCAAGTATGT			
L1AC12s_4:12-F	TGTCGCTGCCAATTCTGTAG	899		482
L1AC12s_4:12-R	GCAGCAGGAACTGAGAAAGG			
L1AC12s_GL3-F	GATTGATGGCCTCTTGTGGT	983		564
L1AC12s_GL3-R	CGCTGCACTCTGTCTACTGC			
L1AC13s_4:27-F	CTGCTTTCAGGGAGTTCTGG	1611		838
L1AC13s_4:27-R	CAAAGCTCCCCAAAATGTGT			
L1AC14_GL-F	AGGCATTTGGTTGGCACTAC		990	832
L1AC14_GL-R	GGTGCAAACTTTTTGGGAAA			
L1AC14_INT	TTCCCAGCGAAACAGAATGG			
L1AC15_2:15-F	GCTGGTGGGACGTTATCTGT		963	1089
L1AC15_2:15-R	TTTCTTTCCCTAGGCCGAAT			
L1AC15_INT1	GTGCGTGCAGGGTTTGTTGT			
L1AC15s_1:87-F	CTGCTGGTCAAGTGTGCTGT	1971		564
L1AC15s_1:87-R	ATCTATTGGGCGTGGACAAC			
L1AC16sGL3-f	CCAGGGAGTCTGCATTTGAT	1231		721
L1AC16sGL3-r	CCCCAAACACTGCTCATTTT			
L1AC16sGL4-f	ATGTGGAGCTTCCCATTTTG	1244		675
L1AC16sGL4-r	ATGGGTCCCAAACGTAATCA			
L1AC16sGL5-f	TGGACAGAGCCTTCTCAGGT	1576		968
L1AC16sGL5-r	CCAAGCATTTCGGATAAGGA			
L1AC17s1:544-f	GGAGGTCCACAAGCATAGGA	1462		546
L1AC17s1:544-r	CCTACCTTTTGAGGCAGCAG			
L1AC17sGLy-f	CCATGTCCTTGGTCATTTCC	1399		511
L1AC17sGLy-r	GGGACTACCACCAAAGGTGA			
L1AC18_chr1:107_f	GCTGGTCTGCATCTTTCACA	1669		589
L1AC18_chr1:107_r	GCCAGTTGGGTGCTATGAAT			
L1AC18_chr1:128_f	CCTTCCTGATCCCCCTAGAC	1436		625
L1AC18_chr1:128_r	TGGCCGATGTCATTATCTGA			
L1AC18_chr1:223_f	ATCCTCCATTCTCGTGCATT	1394		679
L1AC18_chr1:223_r	CCCAAAGAAGGCTCAAAACA			

L1AC18_chr5:543_f	CATGGGTTAATGTGGGGTGT	1799		931
L1AC18_chr5:543_r	TGTTGGACAGGATGGCTAGA			
L1AC19_chr2:144_f	CCACTATTGGCAGGATCGTT	1727		707
L1AC19_chr2:144_r	TTTTGTGTGGCAATTGTGCT			
L1AC19_chr3:139_f	CTGGGTTCCTCAACCAGAAA	1970		1045
L1AC19_chr3:139_r	CCAAGTGCACTCTGCCTGT			
L1AC2.26-F	TTGAGAATCGTTGACGAGCA	1448		180
L1AC2.26-R	ATGGATTGCATCATTCAGCA			
L1AC20_chr1:150_f	TGGACCATGTCAACCACAAC	1330		821
L1AC20_chr1:150_r	CATCCCAGGAGACTCCAAAA			
L1AC20_chr1:257_f	TGGGCAGTGTGTTATCTCCA	1134		509
L1AC20_chr1:257_r	GCACCGAAAAATCACCATGT			
L1AC20_chr3:170_f	ATCCCTGACAGGCAACTCAC		1072	641
L1AC20_chr3:170_r	GACATCAGCTGGTGCTTCAA			
L1AC20_chr4:684_f	CTTCCATGGGTGCCCTACTA	984		627
L1AC20_chr4:684_r	GGAAAGGACGGGGAATCTAC			
L1AC20_chr5:227_f	TGGAAGACAACAAGCAAGGA	1542		830
L1AC20_chr5:227_r	TATCCTGGGTTTCCCTACCC			
L1AC20_chr5:660_f	GACACAGGGGAAAGAAACGA	1649		769
L1AC20_chr5:660_r	GTCATTGTCCTCCAAACATGAA			
L1AC20_INT1	ACAATAAGAGCAGCACATGC			
L1AC3.10-F	AATGACCTTGGAGGGCCTTA		892	570
L1AC3.10-R	CATGCAATGGTCAGCATCAT			
L1AC3.18-F	AACAAAACCCACGACCTGAG		862	760
L1AC3.18-R	CTGGGTTTTTGATGCCAGTT			
L1AC3.21-F	TTTGGGTCTTCGGATTGTTC		600	1092
L1AC3.21-R	TGTGACCAGCACCAAATGTT			
L1AC3.24-F	TCTGCTGAACTCCAAGGTGA		787	720
L1AC3.24-R	CAGTGCTGTGGAAGTCGAAA			
L1AC3.25-F	ATGAGACACGCCACAAATGA		687	420
L1AC3.25-R	TTTGTGTTCCATGGTGCCTA			
L1AC3.3-F	TGGGAAGGGGATGTAATAGC		1722	1800
L1AC3.3-R	GCAGAACCCATCATCACCTTA			
L1AC3.4-F	TGCACTCCAATTAGAGGAAAC		986	714
L1AC3.4-R	CCAGGGTTTGAATCAGCATT			
L1AC3.8-F	TCAGTTTGGGACACATTCCA		1077	1020
L1AC3.8-R	TGTGTGTCCATTGCTTGGTT			
L1AC3.INTa	ATGTGCTTCAGGGTTGGAA			
L1AC4.1-F	GACTGGGAAAAATGCTGGAA		949	650
L1AC4.1-R	TCCTCTGTTGTTGGGGTTTC			
L1AC4.11-F	TCTTGGGGTCATTCATCTCC		1478	1440
L1AC4.11-R	CCACAAACACTCTGCAGGAA			
L1AC4.15-F	TCATCTCAGTGCCTGTTTGC		921	1200
L1AC4.15-R	CAGTTGGCAGCCAGTGTTTA			
L1AC4.17-F	TGTTTGGGAGCCTTGATCTT		1230	1278
L1AC4.17-R	CTCCCATTCATCAAGGTGCT			
L1AC4.18-F	GGAAAATCAGGATTCGGAGA		1103	1470
L1AC4.18-R	TATTCCTTCACCCCAGACCA			
L1AC4.19-F	TGGTTACGAGCTGGCTTCTT		1137	960
L1AC4.19-R	TCAATTACGGGTCATCAGCA			
L1AC4.2-F	TTGGGAAGCCATACTTGAGC		1145	1200

L1AC4.2-R	CCAGCCACCAAAGAGTTGAT			
L1AC4.20-F	CCATGAACCTTCTTCCCAGA		1193	1050
L1AC4.20-R	GCTGGCTGAAGAACCAAGTT			
L1AC4.21-F	CTGCACATATTGGTGGCAAG		1291	1290
L1AC4.21-R	TGGGACCACTGGGAATACAT			
L1AC4.22-F	GCTTGACCTTGACCCCATTA		834	450
L1AC4.22-R	CCCCGACCATTCATACCTA			
L1AC4.25-F	AAGCAGCGAACATTCCTCAT		1124	900
L1AC4.25-R	TATTTCTCCATTGCCCCATC			
L1AC4.26-F	TCCTGCTGGGAACTATTGCT		1292	810
L1AC4.26-R	GTGGGATCTTTCCATGTTGC			
L1AC4.4-F	TGGTTGAGAAAACAGCACCA		1083	960
L1AC4.4-R	TTGTGGGTTTTGAGGAAAGC			
L1AC4.8-F	CCAACAAAGGATTCCCTCAA		1565	1560
L1AC4.8-R	ATCTGGTTGTCTCCCGAATG			
L1AC4INTa	ATGTGGTGGGAGTGTTCTGA			
L1AC8-INT	CTCCACGGCAGCAAGGATTA			
L1AC8.1:108-f	GCTGAGCATTCAAAACAGCA		798	1563
L1AC8.1:108-r	TGCCTGTCCCCTAACTGAAC			
TE_3-F	AGCATCCCAAACCATACCTG	1097		382
TE_3-R	GCTGCCACAGACATGACACT			

Table 5. Primers used for polymorphism ascertainment of novel cloned L1 insertions. Given here are the primers that flanked the empty site in the *Anolis* genome databse, the expected size of the empty site, and the population in which each locus was screened. Presence for each locus was determined by amplification with the reverse primers given here and the forward internal primer from the appopriate L1 family given in Table 1.

PRIMER NAME	OLIGO	Size (if empy)	POPULATION	L1 FAMILY
A_B8F	TCTCTGGAAGTTTGGGTTGG	605	CFL	L1AC18
A_B8R	AGCCATTGCTACCTCCCCTA		CFL	L1AC18
A_E5F	TGTGTACATCTGGCATTTGGA	588	CFL	L1AC18
A_E5R	AAGGCACCATCTCCATCTTG		CFL	L1AC18
A_E6F	TCTGCTTGTGTTGCGTTTTC	709	CFL	L1AC18
A_E6R	GGGCCAGACAAACTGTGAAT		CFL	L1AC18
A_F12F	GCAGAAACACCCTGCAGTTC	521	CFL	L1AC18
A_F12R	GGTGAGCTTCATTCTCAACCA		CFL	L1AC18
A_F4F	TGTCTGCATGGTTTCATGTCT	625	CFL	L1AC18
A_F4R	TATTCCGACCTGTGCTTTGG		CFL	L1AC18
A_F6F	TAGAATGCAGGGCAGAAAGG	586	CFL	L1AC18
A_F6R	TCTGCTTGATGGTGTGAGGA		CFL	L1AC18
B_F6F	GAGGGTCGTTTTGGGATGTA	440	CFL	L1AC18
B_F6R	TAGGGTGAGTTCAGGCTTGG		CFL	L1AC18
B_H7F	GAGGGAAAACCTTGCAACTG	443	CFL	L1AC18
B_H7R	TGCACGTATAGAGGGCACTG		CFL	L1AC18
C_A5F	GGATGGGTTTCCGATTTCTT	494	CFL	L1AC20
C_A5R	GGTGGATTTGGAATTGGTTG		CFL	L1AC20
C_D7F	TGGAGCAAATGGAATTTGAA	769	CFL	L1AC20
C_D7R	GCAAAATGCCGAACAAAAAT		CFL	L1AC20
C_E6F	GGAGCACCTTGAGAAACTGC	457	CFL	L1AC20
C_E6R	TTGCTGATCCAACATTCTGC		CFL	L1AC20
C_F6F	ACGTGGCTGAAAAGCTGAAT	452	CFL	L1AC20
C_F6R	CCAGCAGCATCCTCCTAAAG		CFL	L1AC20
D_C10F	AACTCCGTGCTTTCTGGTTG	403	Gulf-Atlantic	L1AC20
D_C10R	AAAGTCCTGCCCAGATGTGT		Gulf-Atlantic	L1AC20
D_H4F	AATCCCGGGGAACATAGAAC	471	Gulf-Atlantic	L1AC20
D_H4R	GGACCTGTTCAGCCTGTGAT		Gulf-Atlantic	L1AC20
E_B4F	CTTCCATCATTACCCCCAAA	628	Gulf-Atlantic	L1AC18
E_B4R	ATGGCTTTGTTTCCTTGAGC		Gulf-Atlantic	L1AC18
E_D10F	CCTCCCACAAGGATGGTAAA	797	Gulf-Atlantic	L1AC18
E_D10R	TGGACCTGGAATGAAAAAGC		Gulf-Atlantic	L1AC18
E_D11F	GCTTGAGACCGCTAAAGCTC	541	Gulf-Atlantic	L1AC18
E_D11R	CTTATCTGGCTTGGGAAACG		Gulf-Atlantic	L1AC18
E_D1F	AGGAGAAGGTGGAAGCTGAC	1089	Gulf-Atlantic	L1AC18
E_D1R	GGTTCCTCCGTAATCCACCT		Gulf-Atlantic	L1AC18
E_D9F	TGCAGAGCTAGCATTTCCAG	717	Gulf-Atlantic	L1AC18

E_D9R	GTGGTTTCTGGGTGGTATGG		Gulf-Atlantic	L1AC18
E_E3F	TTGTGTCCTGTGGATTTTGC	517	Gulf-Atlantic	L1AC18
E_E3R	CTCCCCCAAACACAGAAGG		Gulf-Atlantic	L1AC18
E_E5F	TGCTCTCAGTCACAGCAGTTG	548	Gulf-Atlantic	L1AC18
E_E5R	CCACCACCATGCCTAGTTCT		Gulf-Atlantic	L1AC18
E_E8F	TGCTGTGTGTGAACTGATTTCT	515	Gulf-Atlantic	L1AC18
E_E8R	CAGGGCACACAGAAGACTGA		Gulf-Atlantic	L1AC18
F_A10F	AGGTTTCAGTCCAGTCGTGA	579	Suwannee	L1AC20
F_A10R	TGCACCTTATGAAAGTTTGGGA		Suwannee	L1AC20
F_A11F	GCCCATGCCTGCTCTAGAT	500	Suwannee	L1AC20
F_A11R	CCTGGACGTTTTGGAGTTGT		Suwannee	L1AC20
F_F7F	AGACTGCTATCCCTCCATCTTG	508	Suwannee	L1AC20
F_F7R	ACCTTGTGTGGCTGTGGAG		Suwannee	L1AC20
G_B4F	TGCTCTCTGGATGTAGATGAACT	544	Suwannee	L1AC18
G_B4F	GTACAAGACACCAGGCTTTGT		Suwannee	L1AC18
G_H2F	ATTTCCCGCACTCTCCCTAG	511	Suwannee	L1AC18
G_H2R	CAGGGCATTTTCCAGCTCTG		Suwannee	L1AC18
H_B11F	ATCAATTGGCACTGTTCCGC	531	NC	L1AC18
H_B11R	ACCAGCAGCATCCTCCTAAA		NC	L1AC18
H_C4F	CCTACGCTTGCACCAAAGAA	611	NC	L1AC18
H_C4R	CAATCCTCGTGTTGGTGGTG		NC	L1AC18
H_F10F	GGCCTGAGGCTGTTAGGAAT	657	NC	L1AC18
H_F10R	TGCACCACTCAGACTTTTGTTT		NC	L1AC18
H_F12F	GAGGGAGCAGGCTTGTTTTC	704	NC	L1AC18
H_F12R	ATGCCCTCAGTGTGGAAGAA		NC	L1AC18
I_B2F	AGGCTGCAGCAGTTTGTTTT	877	Everglades	L1AC20
I_B2R	ACACCTCAGCCACCTCTCTG		Everglades	L1AC20
I_C8R	CTTTTCTTTTGTACGGCTCTGG	653	Everglades	L1AC20
I_C8R	TGTTTATTGGCCTCCGTGTTG		Everglades	L1AC20
J_A12F	GGCAAGAGCTTTTGTGGATT	608	Everglades	L1AC18
J_A12R	CCCTGGGCATTTTCTAGGTT		Everglades	L1AC18
J_C12F	TCCATGTGAGCAATTTTGGA	564	Everglades	L1AC18
J_C12R	AAAGTTGTACAGGGTCACTTGAAA		Everglades	L1AC18
J_C6F	TGGAAGTCTGGCCTCTAGGA	988	Everglades	L1AC18
J_C6R	CCCGAAATTCTCAGTGTGGT		Everglades	L1AC18
J_D4F	TCGAATTCCATCACCATCAA	601	Everglades	L1AC18
J_D4R	TTTTCTAGCTTCCGGCTTCA		Everglades	L1AC18
J_H2F	GAGGAGCTGCAGTCCAAAAA	544	Everglades	L1AC18
J_H2R	TTTGACCATAGCACCCATCA		Everglades	L1AC18

Table 6: Summary of L1 inserts collected during the cloning experiment for each green anole population and information about polymorphism.

		Cloning					
		Everglades	**Suwannee**	**Central Florida**	**North Carolina**	**Gulf-Atlantic**	**Total**
Clones collected and sequenced							480
Clones containing an L1 element							380
Total sequences mapped to database							265
Number of different L1AC18 inserts							
Flanking sequences located in database		55	23	43	22	29	172
Insertion sites occupied in database		42	20	37	17	21	137
Insertion sites empty in database		13	3	6	5	8	35
	Tested by PCR	10	3	5	4	6	28
	Proportion <50% Polymorphism	80%	63%	83%	100%	0	
	No. FL inserts	0	2	2	1	4	9
	Proportion FL >50%	-	0	0	0	100%	
Number of different L1AC20 inserts							
Flanking sequences located in database		18	12	34	14	15	93
Insertion sites occupied in database		12	9	26	14	12	73
Insertion sites empty in database		6	3	6	0	3	18
	Tested by PCR	6	2	5	-	2	15
	Proportion <50% Polymorphism	100%	50%	100%	-	50%	
	No. FL inserts	0	2	0	0	0	2
	Proportion FL >50%	-	50%	-	-	-	

Table 7: Locus-specific information for 52 L1 loci retrieved from the *Anolis* genome and their frequencies in five green anole populations. Asterisk (*) indicates an insert collected from the February 2007 version that we were not able to map onto the May 2010 version.

	Locus	Coordinates	Length (bp)	FL or TR	% Divergence from Consensus	Sample Size (2*N*)	North Carolina	Suwannee	Central Florida	Everglades	Gulf-Atlantic
1	L1AC20_684	chr4:68403974-68404330	357	TR	3.8	397	0.5	0.567	0.838	0.438	0.897
2	L1AC12s_4:12	chr4:126979289-126979702	414	TR	2.5	230	1	0.964	1	1	1
3	L1AC12s_GL3	chrUn_GL343596:105315-105733	419	TR	3.7	186	1	0.75	0.59	0	1
4	L1AC11s_6:33	chr6:33204599-33205038	440	TR	0	178	0.25	0.071	0.026	0	0.28
5	L1AC20_150	chr1:150575039-150575547	509	TR	7.3	390	1	0.969	1	0.938	0.884
6	L1AC16s_GL3	chrUn_GL343395:465703-466212	510	TR	4.3	194	1	1	1	1	1
7	L1AC16s_GL4	chrUn_GL343471:34099-34667	569	TR	1.8	194	1	1	1	1	1
8	L1AC16s_GL5	chrUn_GL344110:24189-24796	608	TR	1.5	194	1	1	1	1	1
9	L1AC20_257	chr1:257099845-257100469	625	TR	12.2	373	1	1	1	0.625	1
10	L1AC20_227	chr5:22764139-22764850	712	TR	11	337	1	1	1	1	1
11	L1AC18_223	chr1:223912127-223912841	715	TR	3.9	292	1	1	1	1	1
12	TE_3	chr1:214783982-214784696	715	TR	1.4	212	0.036	0	0	0	0.214
13	L1AC13s_4:27	chr4:27512892-27513664	773	TR	3	182	0.4	0	0	0	0.905
14	L1AC18_128	chr1:128475510-128476320	811	TR	3.6	419	1	1	1	1	1
15	L1AC18_543	chr5:54332386-54333253	868	TR	5.2	178	1	1	1	1	1
16	L1AC20_660	chr5:66011824-66012703	880	TR	4.7	398	0.375	0.269	0.333	0.125	0.713
17	L1AC17s_1:544	chr1:54502268-54503155	888	TR	2.7	98	0.818	0.667	0.714	1	1
18	L1AC17s_Gly	chrUn_GL343200:1968310-1969771	916	TR	1.4	185	0.269	0	0.026	0	0
19	L1AC19_139	chr3:139851678-139852602	925	TR	1.9	182	1	1	1	1	1
20	L1AC19_2:144	chr2:144963722-144964741	1020	TR	1.7	112	0.875	0.45	0.452	0	0.25
21	L1AC18_107	chr1:107831209-107832288	1080	TR	1.2	380	1	1	1	1	1
22	L1AC2.26	chrUn_GL343239:906,018-907,246	1229	TR	1.1	194	1	1	1	1	1
23	L1AC15s_1:87	chr1:87962249-87963655	1407	TR	0.42	226	0.1	0	0	0	0.025
24	L1AC3.25	chr2:172,762,715-172,767,090	4376	TR	0.45	352	1	0.941	1	1	0.967
25	L1AC3.24	scaffold_24:516,031-520,687*	4657	TR	0.53	268	1	0.706	1	0.833	0.81
26	L1AC3.21	chrUn_GL343497:464,966-469,686	4721	TR	0.45	345	0.5	0.032	0	0	0.071
27	L1AC4.17	scaffold_125:1,567,058-1,571,933*	4876	TR	0.54	202	1	1	1	0.929	1
28	L1AC4.18	scaffold_43:3,503,254-3,508,180*	4927	TR	0.54	302	1	0	0.051	0.071	0.254
29	L1AC4.15	chr3:96,424,624-96,429,616	4993	TR	0.5	108	N.A.	0.7	0.769	1	0.958
30	L1AC3.18	chrUn_GL343280:1,636,141-1,641,248	5108	TR	0.57	108	1	N.A.	N.A.	1	1
31	L1AC4.8	chr5:19,962,314-19,967,530	5217	FL	0.39	62	N.A.	0	N.A.	0	N.A.
32	L1AC4.19	chrUn_GL343243:1,081,190-1,086,410	5221	FL	0.52	200	1	0.5	0.41	1	0.929
33	L1AC4.22	chr2:90,589,288-90,594,512	5225	FL	0.46	18	N.A.	N.A.	N.A.	0	0.313
34	L1AC4.20	scaffold_527:549,438-554,665*	5228	FL	0.85	126	0.9	0.125	0.686	0.125	0.875
35	L1AC4.11	chr3:178,322,659-178,327,899	5241	FL	0.91	106	N.A.	0.563	N.A.	0	0.833
36	L1AC4.2	scaffold_85:3,499,711-3,504,951*	5241	FL	0.35	380	0.308	0.026	0	0	0.539
37	L1AC4.25	chr3:170,477,780-170,483,021	5242	FL	0.46	228	1	0	0	0	0.239
38	L1AC4.26	chr3:159,015,596-159,020,837	5242	FL	0.5	226	0.8	0.167	0.179	0.4	0.654
39	L1AC20_3:170	chr3:170477780-170483022	5243	FL	0.31	194	0	0	0	0	0
40	L1AC4.21	chr3:32,972,264-32,977,686	5243	FL	0.66	178	0.625	0	0	0.25	0.676
41	L1AC4.4	scaffold_30:3,968,578-3,973,822*	5245	FL	0.52	365	0	0.184	0.103	0.5	0.233
42	L1AC4.1	chr2:172,917,348-172,922,593	5246	FL	0.35	250	0	0	0.026	0	0
43	L1AC8_1:108	chr1:108,322,088-108,328,343	5334	FL	1	194	0	0	0	0	0
44	L1AC3.4	chr6:54,998,113-55,004,259	6147	FL	0.47	350	0	0.125	0.026	0	0.382
45	L1AC3.10	chrUn_GL343295:80,571-86,721	6151	FL	0.36	246	1	0.063	0.231	0	0.779
46	L1AC15_5:14	chr5:142569373-142575532	6160	FL	0.1	140	0.8	0.222	0.564	0	0.688
47	L1AC3.8	chr3:168,587,085-168,593,244	6160	FL	0.31	252	1	0.184	0.103	0	0.611
48	L1AC15_2:15	chr2:153639275-153645435	6161	FL	0.31	130	0.7	0.556	0.872	0	1
49	L1AC3.3	scaffold_57:1,761,641-1,767,805*	6165	FL	0.39	289	0.269	0.281	0.026	0.2	0.232
50	L1AC11_2:10	chr2:107077315-107083913	6599	FL	0.06	138	0.417	0.111	0.282	0	0.05
51	L1AC14_GL	chrUn_GL343255:672694-679409	6716	FL	0.7	136	0.917	0	0.231	N.A.	0.85
52	L1AC10.2	chr3:172235202-172242019	6818	FL	0.73	194	0	0	0	0	0

Table 8: Comparison of truncated (TR) and full-length (FL) allele frequencies in five green anole populations. The number of inserts measured (N) for each category is indicated above the average population frequency. The p-value for the Wilcoxon Rank Sum Test is given. D indicates the largest difference between the cumulative distributions of each sample. n.s. = non-significant

Population	**TR**	**FL**	**Wilcoxon Rank-Sum Test**	**Kolmogorov-Smirnov Test**	
				***p*-value**	***D***
Gulf-Atlantic	N=30 0.74	N=21 0.41	<0.01	0.002	0.510
North Carolina	N=30 0.78	N=14 0.73	n.s.	n.s.	0.348
Everglades	N=27 0.81	N=6 0.36	n.s.	-	-
Suwannee	N=25 0.78	N=17 0.73	<0.001	0	0.800
Central Florida	N=27 0.84	N=15 0.30	<0.001	0	0.674

Table 9. Pair wise population comparisons of full-length (FL) and truncated (TR) L1 insertions. P-values from the Wilcoxon Rank-Sum Test are given. n.s. = not significant.

Populations Compared	**Wilcoxon Rank-Sum Test**	
	TR	FL
North Carolina-Suwannee	n.s.	<0.001
North Carolina-Central Florida	n.s.	<0.001
North Carolina-Everglades	n.s.	n.s.
North Carolina-Gulf/Atlantic	n.s.	n.s.
Gulf/Atlantic-Suwannee	n.s.	<0.001
Gulf/Atlantic-Central Florida	n.s.	<0.001
Gulf/Atlantic-Everglades	n.s.	n.s.
Suwannee-Everglades	n.s.	n.s.
Suwannee-Central Florida	n.s.	n.s.
Everglades-Central Florida	n.s.	n.s.

REFERENCES

1. Lynch, M. and J.S. Conery, *The origins of genome complexity.* Science, 2003. **302**(5649): p. 1401-4.
2. Arkhipova, I.R., *Distribution and phylogeny of Penelope-like elements in eukaryotes.* Syst Biol, 2006. **55**(6): p. 875-85.
3. Kapitonov, V.V., S. Tempel, and J. Jurka, *Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences.* Gene, 2009. **448**(2): p. 207-13.
4. Malik, H.S., W.D. Burke, and T.H. Eickbush, *The age and evolution of non-LTR retrotransposable elements.* Mol Biol Evol, 1999. **16**(6): p. 793-805.
5. Luan, D.D., et al., *Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition.* Cell, 1993. **72**(4): p. 595-605.
6. Khazina, E. and O. Weichenrieder, *Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame.* Proc Natl Acad Sci U S A, 2009. **106**(3): p. 731-6.
7. Boissinot, S. and A.V. Furano, *Adaptive evolution in LINE-1 retrotransposons.* Mol Biol Evol, 2001. **18**(12): p. 2186-94.
8. Martin, S.L., et al., *Trimeric structure for an essential protein in L1 retrotransposition.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13815-20.
9. Martin, S.L., *Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells.* Mol Cell Biol, 1991. **11**(9): p. 4804-7.
10. Martin, S.L. and F.D. Bushman, *Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon.* Mol Cell Biol, 2001. **21**(2): p. 467-75.
11. Swergold, G.D., *Identification, characterization, and cell specificity of a human LINE-1 promoter.* Mol Cell Biol, 1990. **10**(12): p. 6718-29.
12. Cost, G.J., et al., *Human L1 element target-primed reverse transcription in vitro.* Embo J, 2002. **21**(21): p. 5899-910.
13. Martin, S.L., et al., *The structures of mouse and human L1 elements reflect their insertion mechanism.* Cytogenet Genome Res, 2005. **110**(1-4): p. 223-228.
14. Ichiyanagi, K. and N. Okada, *Mobility pathways for vertebrate L1, L2, CR1, and RTE clade retrotransposons.* Mol Biol Evol, 2008. **25**(6): p. 1148-57.
15. Duvernell, D.D., S.R. Pryor, and S.M. Adams, *Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover.* J Mol Evol, 2004. **59**(3): p. 298-308.
16. Furano, A.V., D. Duvernell, and S. Boissinot, *L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish.* Trends Genet, 2004. **20**(1): p. 9-14.
17. Novick, P.A., et al., *The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard Anolis carolinensis shows more similarity to fish than mammals.* Mol Biol Evol, 2009. **26**(8): p. 1811-22.
18. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
19. Furano, A.V., *The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons.* Prog Nucleic Acid Res Mol Biol, 2000. **64**: p. 255-94.
20. Adey, N.B., et al., *Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences.* Mol Biol Evol, 1994. **11**(5): p. 778-89.
21. Boissinot, S., P. Chevret, and A.V. Furano, *L1 (LINE-1) retrotransposon evolution and amplification in recent human history.* Mol Biol Evol, 2000. **17**(6): p. 915-28.
22. Khan, H., A. Smit, and S. Boissinot, *Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates.* Genome Res, 2006. **16**(1): p. 78-87.
23. Warren, W.C., et al., *Genome analysis of the platypus reveals unique signatures of evolution.* Nature, 2008. **453**(7192): p. 175-83.
24. Wei, W., et al., *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol, 2001. **21**(4): p. 1429-39.
25. Dewannieux, M., C. Esnault, and T. Heidmann, *LINE-mediated retrotransposition of marked Alu sequences.* Nat Genet, 2003. **35**(1): p. 41-8.

26. Ohshima, K., et al., *The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements.* Mol Cell Biol, 1996. **16**(7): p. 3756-64.
27. Piskurek, O., H. Nishihara, and N. Okada, *The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian genome of Anolis carolinensis.* Gene, 2009. **441**(1-2): p. 111-8.
28. Piskurek, O., C.C. Austin, and N. Okada, *Sauria SINEs: Novel short interspersed retroposable elements that are widespread in reptile genomes.* J Mol Evol, 2006. **62**(5): p. 630-44.
29. Gentles, A.J., et al., *Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica.* Genome Res, 2007. **17**(7): p. 992-1004.
30. Eickbush, T.H. and V.K. Jamburuthugoda, *The diversity of retrotransposons and the properties of their reverse transcriptases.* Virus Res, 2008. **134**(1-2): p. 221-34.
31. Havecker, E.R., X. Gao, and D.F. Voytas, *The diversity of LTR retrotransposons.* Genome Biol, 2004. **5**(6): p. 225.
32. Malik, H.S. and T.H. Eickbush, *Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses.* Genome Res, 2001. **11**(7): p. 1187-97.
33. Kim, A., et al., *Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster.* Proc Natl Acad Sci U S A, 1994. **91**(4): p. 1285-9.
34. Song, S.U., et al., *An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus.* Genes Dev, 1994. **8**(17): p. 2046-57.
35. Ribet, D., et al., *An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus.* Genome Res, 2008. **18**(4): p. 597-609.
36. Gifford, R. and M. Tristem, *The evolution, distribution and diversity of endogenous retroviruses.* Virus Genes, 2003. **26**(3): p. 291-315.
37. Andersson, M.L., et al., *Diversity of human endogenous retrovirus class II-like sequences.* J Gen Virol, 1999. **80 (Pt 1)**: p. 255-60.
38. Stuart-Rogers, C. and A.J. Flavell, *The evolution of Ty1-copia group retrotransposons in gymnosperms.* Mol Biol Evol, 2001. **18**(2): p. 155-63.
39. Goodwin, T.J. and R.T. Poulter, *The DIRS1 group of retrotransposons.* Mol Biol Evol, 2001. **18**(11): p. 2067-82.
40. Jiang, N., et al., *Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice.* Genetics, 2002. **161**(3): p. 1293-305.
41. Feschotte, C. and E.J. Pritham, *DNA transposons and the evolution of eukaryotic genomes.* Annu Rev Genet, 2007. **41**: p. 331-68.
42. Kapitonov, V.V. and J. Jurka, *Helitrons on a roll: eukaryotic rolling-circle transposons.* Trends Genet, 2007. **23**(10): p. 521-9.
43. Kapitonov, V.V. and J. Jurka, *Rolling-circle transposons in eukaryotes.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8714-9.
44. Kapitonov, V.V. and J. Jurka, *Self-synthesizing DNA transposons in eukaryotes.* Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4540-5.
45. Pritham, E.J., T. Putliwala, and C. Feschotte, *Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses.* Gene, 2007. **390**(1-2): p. 3-17.
46. Fischer, M.G. and C.A. Suttle, *A virophage at the origin of large DNA transposons.* Science, 2011. **332**(6026): p. 231-4.
47. Hartl, D.L., E.R. Lozovskaya, and J.G. Lawrence, *Nonautonomous transposable elements in prokaryotes and eukaryotes.* Genetica, 1992. **86**(1-3): p. 47-53.
48. Novick, P.A., et al., *The evolution and diversity of DNA transposons in the genome of the Lizard Anolis carolinensis.* Genome Biol Evol, 2011. **3**: p. 1-14.
49. Yang, G., et al., *Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE.* Science, 2009. **325**(5946): p. 1391-4.
50. Kordis, D., N. Lovsin, and F. Gubensek, *Phylogenomic analysis of the L1 retrotransposons in Deuterostomia.* Syst Biol, 2006. **55**(6): p. 886-901.
51. Waters, P.D., et al., *Evolutionary history of LINE-1 in the major clades of placental mammals.* PLoS One, 2007. **2**(1): p. e158.

52. Kordis, D. and F. Gubensek, *Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10704-9.
53. Schaack, S., C. Gilbert, and C. Feschotte, *Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution.* Trends Ecol Evol, 2010. **25**(9): p. 537-46.
54. Bartolome, C., X. Bello, and X. Maside, *Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes.* Genome Biol, 2009. **10**(2): p. R22.
55. Thomas, J., S. Schaack, and E.J. Pritham, *Pervasive horizontal transfer of rolling-circle transposons among animals.* Genome Biol Evol, 2010. **2**: p. 656-64.
56. Pace, J.K., 2nd, et al., *Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods.* Proc Natl Acad Sci U S A, 2008. **105**(44): p. 17023-8.
57. Gilbert, C., et al., *Rampant Horizontal Transfer of SPIN Transposons in Squamate Reptiles.* Mol Biol Evol, 2011.
58. Novick, P., et al., *Independent and parallel lateral transfer of DNA transposons in tetrapod genomes.* Gene, 2010. **449**(1-2): p. 85-94.
59. Gilbert, C., et al., *A role for host-parasite interactions in the horizontal transfer of transposons across phyla.* Nature, 2010. **464**(7293): p. 1347-50.
60. Piskurek, O. and N. Okada, *Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals.* Proc Natl Acad Sci U S A, 2007. **104**(29): p. 12046-51.
61. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
62. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
63. Warren, W.C., et al., *Genome analysis of the platypus reveals unique signatures of evolution.* Nature, 2008. **453**(7192): p. 175-U1.
64. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
65. Mikkelsen, T.S., et al., *Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences.* Nature, 2007. **447**(7141): p. 167-77.
66. Hellsten, U., et al., *The genome of the Western clawed frog Xenopus tropicalis.* Science, 2010. **328**(5978): p. 633-6.
67. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
68. Wallis, J.W., et al., *A physical map of the chicken genome.* Nature, 2004. **432**(7018): p. 761-4.
69. Warren, W.C., et al., *The genome of a songbird.* Nature, 2010. **464**(7289): p. 757-62.
70. Gibbs, R.A., et al., *Evolutionary and biomedical insights from the rhesus macaque genome.* Science, 2007. **316**(5822): p. 222-34.
71. Wade, C.M., et al., *Genome sequence, comparative analysis, and population genetics of the domestic horse.* Science, 2009. **326**(5954): p. 865-7.
72. Pontius, J.U., et al., *Initial sequence and comparative analysis of the cat genome.* Genome Res, 2007. **17**(11): p. 1675-89.
73. Jaillon, O., et al., *Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.* Nature, 2004. **431**(7011): p. 946-57.
74. Wolfsberg, T.G., *Using the NCBI Map Viewer to browse genomic sequence data.* Curr Protoc Hum Genet, 2011. **Chapter 18**: p. Unit18 5.
75. Flicek, P., et al., *Ensembl 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D800-6.
76. Fujita, M.K., S.V. Edwards, and C.P. Ponting, *The Anolis lizard genome: an amniote genome without isochores.* Genome Biol Evol, 2011. **3**: p. 974-84.
77. Alfoldi, J., et al., *The genome of the green anole lizard and a comparative analysis with birds and mammals.* Nature, 2011. **477**(7366): p. 587-91.
78. Losos, J.B., *Lizards in an evolutionary tree : the ecology of adaptive radiation in anoles*2009, Berkeley: University of California Press. xx, 507 p.
79. Kidwell, M.G. and D.R. Lisch, *Transposable elements and host genome evolution.* Trends Ecol Evol, 2000. **15**(3): p. 95-99.
80. Volff, J.N., et al., *Diversity of retrotransposable elements in compact pufferfish genomes.* Trends Genet, 2003. **19**(12): p. 674-8.

81. Furano, A.V., D.D. Duvernell, and S. Boissinot, *L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish.* Trends Genet, 2004. **20**(1): p. 9-14.
82. Blass, E., M. Bell, and S. Boissinot, *Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback.* Genome Biol Evol, 2012. **4**(5): p. 687-702.
83. Donoghue, P.C. and M.J. Benton, *Rocks and clocks: calibrating the Tree of Life using fossils and molecules.* Trends Ecol Evol, 2007. **22**(8): p. 424-31.
84. Kordis, D., *Transposable elements in reptilian and avian (sauropsida) genomes.* Cytogenet Genome Res, 2009. **127**(2-4): p. 94-111.
85. Boissinot, S., et al., *The insertional history of an active family of L1 retrotransposons in humans.* Genome Res, 2004. **14**(7): p. 1221-31.
86. Kojima, K.K., V.V. Kapitonov, and J. Jurka, *Recent expansion of a new Ingi-related clade of Vingi non-LTR retrotransposons in hedgehogs.* Mol Biol Evol, 2011. **28**(1): p. 17-20.
87. de la Chaux, N. and A. Wagner, *BEL/Pao retrotransposons in metazoan genomes.* BMC Evol Biol, 2011. **11**(1): p. 154.
88. Abrusan, G., et al., *Biased distributions and decay of long interspersed nuclear elements in the chicken genome.* Genetics, 2008. **178**(1): p. 573-81.
89. Di-Poi, N., J.I. Montoya-Burgos, and D. Duboule, *Atypical relaxation of structural constraints in Hox gene clusters of the green anole lizard.* Genome Res, 2009. **19**(4): p. 602-10.
90. Di-Poi, N., et al., *Changes in Hox genes' structure and function during the evolution of the squamate body plan.* Nature, 2010. **464**(7285): p. 99-103.
91. Boissinot, S., et al., *Fitness cost of LINE-1 (L1) activity in humans.* Proc Natl Acad Sci U S A, 2006. **103**(25): p. 9590-4.
92. Boissinot, S., A. Entezam, and A.V. Furano, *Selection against deleterious LINE-1-containing loci in the human lineage.* Mol Biol Evol, 2001. **18**(6): p. 926-35.
93. Song, M. and S. Boissinot, *Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination.* Gene, 2007. **390**(1-2): p. 206-13.
94. Eickbush, T.H. and A.V. Furano, *Fruit flies and humans respond differently to retrotransposons.* Curr Opin Genet Dev, 2002. **12**(6): p. 669-74.
95. Wiens, J.J., et al., *Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa.* Syst Biol, 2010. **59**(6): p. 674-88.
96. Fry, B.G., et al., *Early evolution of the venom system in lizards and snakes.* Nature, 2006. **439**(7076): p. 584-8.
97. Castoe, T.A., et al., *Discovery of highly divergent repeat landscapes in snake genomes using high throughput sequencing.* Genome Biol Evol, 2011.
98. Bingham, P.M., M.G. Kidwell, and G.M. Rubin, *The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family.* Cell, 1982. **29**(3): p. 995-1004.
99. Schaefer, R.E., M.G. Kidwell, and A. Fausto-Sterling, *Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Morphological and Cytological Studies of Ovarian Dysgenesis.* Genetics, 1979. **92**(4): p. 1141-52.
100. Pasyukova, E.G., et al., *Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness.* J Hered, 2004. **95**: p. 284-290.
101. Biemont, C., et al., *Population dynamics of the copia, mdg1, mdg3, gypsy, and P transposable elements in a natural population of Drosophila melanogaster.* Genetical Research, 1994. **63**(3): p. 197-212.
102. Charlesworth, B. and D. Charlesworth, *The population dynamics of transposable elements.* Genetical Research, 1983. **42**: p. 1-27.
103. Charlesworth, B. and C.H. Langley, *The population genetics of Drosophila transposable elements.* Annu Rev Genet, 1989. **23**: p. 251-287.
104. Charlesworth, B., A. Lapid, and D. Canada, *The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. II. Inferences on the nature of selection against elements.* Genetical Research, 1992. **60**(2): p. 115-30.
105. Nuzhdin, S.V. and T.F. Mackay, *The genomic rate of transposable element movement in Drosophila melanogaster.* Mol Biol Evol, 1995. **12**(1): p. 180-1.
106. Biemont, C., et al., *Maintenance of transposable element copy number in natural populations of Drosophila melanogaster and D. simulans.* Genetica, 1997. **100**(1-3): p. 161-6.

107. Montgomery, E., B. Charlesworth, and C.H. Langley, *A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster.* Genetical Research, 1987. **49**(1): p. 31-41.
108. Langley, C.H., J.F. Brookfield, and N. Kaplan, *Transposable elements in mendelian populations. I. A theory.* Genetics, 1983. **104**(3): p. 457-71.
109. Boissinot, S., et al., *Fitness cost of LINE-1 (L1) activity in humans.* Proc Natl Acad Sci U S A, 2006. **103**: p. 9590-9594.
110. Bartolome, C., X. Maside, and B. Charlesworth, *On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster.* Mol Biol Evol, 2002. **19**: p. 926-937.
111. Rizzon, C., et al., *Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome.* Genome Res, 2002. **12**(3): p. 400-7.
112. Hill, W.G. and A. Robertson, *The effect of linkage on the limit to artificial selection.* Genetical Research, 1966. **8**: p. 269-294.
113. Callinan, P.A. and M.A. Batzer, *Retrotransposable elements and human disease.* Genome Dyn, 2006. **1**: p. 104-15.
114. Han, K., et al., *L1 recombination-associated deletions generate human genomic variation.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19366-71.
115. Gasior, S.L., et al., *The Human LINE-1 retrotransposon creates DNA double-strand breaks.* J Mol Biol, 2006. **357**(5): p. 1383-1393.
116. Petrov, D., et al., *Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila.* Mol Biol Evol, 2003. **20**(6): p. 880-892.
117. Petrov, D.A., et al., *Population genomics of transposable elements in Drosophila melanogaster.* Mol Biol Evol, 2011. **28**(5): p. 1633-44.
118. Dolgin, E.S. and B. Charlesworth, *The effects of recombination rate on the distribution and abundance of transposable elements.* Genetics, 2008. **178**(4): p. 2169-77.
119. Goodier, J.L. and H.H. Kazazian, Jr., *Retrotransposons revisited: the restraint and rehabilitation of parasites.* Cell, 2008. **135**(1): p. 23-35.
120. Hua-Van, A., et al., *The struggle for life of the genome's selfish architects.* Biol Direct, 2011. **6**: p. 19.
121. Muotri, A.R., et al., *The necessary junk: new functions for transposable elements.* Hum Mol Genet, 2007. **16 Spec No. 2**: p. R159-67.
122. Oliver, K.R. and W.K. Greene, *Transposable elements: powerful facilitators of evolution.* Bioessays, 2009. **31**(7): p. 703-14.
123. Bejerano, G., et al., *A distal enhancer and an ultraconserved exon are derived from a novel retroposon.* Nature, 2006. **441**(7089): p. 87-90.
124. Aminetzach, Y.T., J.M. Macpherson, and D.A. Petrov, *Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila.* Science, 2005. **309**(5735): p. 764-7.
125. Gonzalez, J., et al., *High rate of recent transposable element-induced adaptation in Drosophila melanogaster.* PLoS Biol, 2008. **6**(10): p. e251.
126. Pardue, M.L. and P.G. Debaryshe, *Retrotransposons that maintain chromosome ends.* Proc Natl Acad Sci U S A, 2011.
127. Lyon, M.F., *X-chromosome inactivation: a repeat hypothesis.* Cytogenet Cell Genet, 1998. **80**(1-4): p. 133-137.
128. Whitney, K.D., et al., *A role for nonadaptive processes in plant genome size evolution?* Evolution, 2010. **64**(7): p. 2097-109.
129. Garcia Guerreiro, M.P., et al., *Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of Drosophila subobscura.* BMC Evol Biol, 2008. **8**: p. 234.
130. Lockton, S., J. Ross-Ibarra, and B.S. Gaut, *Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata.* Proc Natl Acad Sci U S A, 2008. **105**(37): p. 13965-70.
131. Gonzalez, J., et al., *Inferring the strength of selection in Drosophila under complex demographic models.* Mol Biol Evol, 2009. **26**(3): p. 513-26.
132. Dolgin, E.S., B. Charlesworth, and A.D. Cutter, *Population frequencies of transposable elements in selfing and outcrossing Caenorhabditis nematodes.* Genet Res (Camb), 2008. **90**(4): p. 317-29.

133. Lockton, S. and B.S. Gaut, *The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata.* BMC Evol Biol, 2010. **10**: p. 10.
134. Wright, S.I., et al., *Population dynamics of an Ac-like transposable element in self- and cross-pollinating arabidopsis.* Genetics, 2001. **158**(3): p. 1279-88.
135. Wright, S.I., N. Agrawal, and T.E. Bureau, *Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana.* Genome Res, 2003. **13**(8): p. 1897-903.
136. Duret, L., G. Marais, and C. Biemont, *Transposons but not retrotransposons are located preferentially in regions of high recombination rate in Caenorhabditis elegans.* Genetics, 2000. **156**(4): p. 1661-9.
137. Janes, D.E., et al., *Genome evolution in Reptilia, the sister group of mammals.* Annu Rev Genomics Hum Genet, 2010. **11**: p. 239-64.
138. Eckalbar, W.L., et al., *Somitogenesis in the anole lizard and alligator reveals evolutionary convergence and divergence in the amniote segmentation clock.* Dev Biol, 2012. **363**(1): p. 308-19.
139. Schneider, C.J., *Exploiting genomic resources in studies of speciation and adaptive radiation of lizards in the genus Anolis.* Integr Comp Biol, 2008. **48**(4): p. 520-6.
140. Conant, R. and J.T. Collins, *A field guide to reptiles & amphibians : eastern and central North America*. 3rd ed. The Peterson field guide series1998, Boston: Houghton Mifflin. xviii, 616 p.
141. Lovern, M.B., M.M. Holmes, and J. Wade, *The green anole (Anolis carolinensis): a reptilian model for laboratory studies of reproductive morphology and behavior.* ILAR J, 2004. **45**(1): p. 54-64.
142. Glor, R.E., J.B. Losos, and A. Larson, *Out of Cuba: overwater dispersal and speciation among lizards in the Anolis carolinensis subgroup.* Mol Ecol, 2005. **14**(8): p. 2419-32.
143. Holman, J.A., *Pleistocene amphibians and reptiles in North America*1995: Oxford University Press. 243 p.
144. Vance, T., *Morphological variation and systematics of the green anole, Anolis carolinensis (Reptilia: Iguanidae).* Bulletin of the Maryland Herpetological Society, 1991. **27**: p. 43-89.
145. Gratz, R., *A statistical analysis of variation of scutellation in Anolis carolinensis carolinensis Voigt.* Unpub. M.S. thesis, Univ. of Notre Dame, IN, 1972.
146. Wade, J., A. Echternacht, and G. McCracken, *Genetic Variation and Similarity in Anolis carolinensis (Sauria: Iguanidae).* Copeia, 1983. **1983**(2): p. 523-529.
147. Soltis, D.E., et al., *Comparative phylogeography of unglaciated eastern North America.* Mol Ecol, 2006. **15**(14): p. 4261-93.
148. Jackson, N.D. and C.C. Austin, *The combined effects of rivers and refugia generate extreme cryptic fragmentation within the common ground skink (scincella lateralis).* Evolution, 2010. **64**(2): p. 409-28.
149. Wood, D.A., et al., *Refugial isolation and divergence in the Narrowheaded Gartersnake species complex (Thamnophis rufipunctatus) as revealed by multilocus DNA sequence data.* Mol Ecol, 2011. **20**(18): p. 3856-78.
150. Burbrink, F.T., R. Lawson, and J.B. Slowinski, *Mitochondrial DNA phylogeography of the polytypic North American rat snake (Elaphe obsoleta): a critique of the subspecies concept.* Evolution, 2000. **54**(6): p. 2107-18.
151. Fontanella, F.M., et al., *Phylogeography of Diadophis punctatus: extensive lineage diversity and repeated patterns of historical demography in a trans-continental snake.* Mol Phylogenet Evol, 2008. **46**(3): p. 1049-70.
152. Rosenberg, N.A. and M. Nordborg, *Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.* Nat Rev Genet, 2002. **3**(5): p. 380-90.
153. Brito, P.H. and S.V. Edwards, *Multilocus phylogeography and phylogenetics using sequence-based markers.* Genetica, 2009. **135**(3): p. 439-55.
154. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
155. Li, C., J.J. Riethoven, and L. Ma, *Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes.* BMC Evol Biol, 2010. **10**: p. 90.

156. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.
157. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
158. Smit, A., R. Hublley, and P. Green, *RepeatMasker Open-3.0.* 1996-2010.
159. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
160. Drummond AJ, et al., *Geneious v5.5, Available from http://www.geneious.com.* 2010.
161. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
162. Hall, T., *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.* Nucleic Acids Symposium Series, 1999. **41**: p. 95-98.
163. Stephens, M., N.J. Smith, and P. Donnelly, *A new statistical method for haplotype reconstruction from population data.* Am J Hum Genet, 2001. **68**(4): p. 978-89.
164. Hudson, R.R. and N.L. Kaplan, *Statistical properties of the number of recombination events in the history of a sample of DNA sequences.* Genetics, 1985. **111**(1): p. 147-64.
165. Librado, P. and J. Rozas, *DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.* Bioinformatics, 2009. **25**(11): p. 1451-2.
166. Woerner, A.E., M.P. Cox, and M.F. Hammer, *Recombination-filtered genomic datasets by information maximization.* Bioinformatics, 2007. **23**(14): p. 1851-3.
167. Vaidya, G., D. Lohman, and R. Meier, *SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information.* Cladistics, 2011. **27**(2): p. 171-180.
168. Stamatakis, A., T. Ludwig, and H. Meier, *RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.* Bioinformatics, 2005. **21**(4): p. 456-63.
169. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
170. Drummond, A.J. and A. Rambaut, *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC Evol Biol, 2007. **7**: p. 214.
171. Macey, J.R., et al., *Phylogenetic relationships among Agamid lizards of the Laudakia caucasia species group: testing hypotheses of biogeographic fragmentation and an area cladogram for the Iranian Plateau.* Mol Phylogenet Evol, 1998. **10**(1): p. 118-31.
172. Glor, R.E., et al., *Partial island submergence and speciation in an adaptive radiation: a multilocus analysis of the Cuban green anoles.* Proc Biol Sci, 2004. **271**(1554): p. 2257-65.
173. Kolbe, J.J., et al., *Multiple sources, admixture, and genetic variation in introduced anolis lizard populations.* Conserv Biol, 2007. **21**(6): p. 1612-25.
174. Rambaut, A. and A.J. Drummond, *Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer.* 2007.
175. Huelsenbeck, J.P., P. Andolfatto, and E.T. Huelsenbeck, *Structurama: bayesian inference of population structure.* Evol Bioinform Online, 2011. **7**: p. 55-9.
176. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.
177. Falush, D., M. Stephens, and J.K. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.* Genetics, 2003. **164**(4): p. 1567-87.
178. Earl, D.A. and B.M. Vonholdt, *STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.* Conservation Genetics Resources, 2011.
179. Evanno, G., S. Regnaut, and J. Goudet, *Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.* Mol Ecol, 2005. **14**(8): p. 2611-20.
180. Orozco-terWengel, P., J. Corander, and C. Schlotterer, *Genealogical lineage sorting leads to significant, but incorrect Bayesian multilocus inference of population structure.* Mol Ecol, 2011. **20**(6): p. 1108-21.
181. Kalinowski, S.T., *The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure.* Heredity (Edinb), 2011. **106**(4): p. 625-32.
182. Excoffier, L. and H.E. Lischer, *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.* Mol Ecol Resour, 2010. **10**(3): p. 564-7.

183. Tajima, F., *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics, 1989. **123**(3): p. 585-95.
184. Fu, Y.X., *Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection.* Genetics, 1997. **147**(2): p. 915-25.
185. Harpending, H.C., *Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution.* Hum Biol, 1994. **66**(4): p. 591-600.
186. Drummond, A.J., et al., *Bayesian coalescent inference of past population dynamics from molecular sequences.* Mol Biol Evol, 2005. **22**(5): p. 1185-92.
187. Heled, J. and A.J. Drummond, *Bayesian inference of population size history from multiple loci.* BMC Evol Biol, 2008. **8**: p. 289.
188. Hewitt, G.M., *Genetic consequences of climatic oscillations in the Quaternary.* Philos Trans R Soc Lond B Biol Sci, 2004. **359**(1442): p. 183-95; discussion 195.
189. Slatkin, M., *Isolation by Distance in Equilibrium and Nonequilibrium Populations.* Evolution, 1993. **47**(1): p. 264-279.
190. Sokal, R.R. and F.J. Rohlf, *Biometry : the principles and practice of statistics in biological research*. 3rd ed1995, New York: W.H. Freeman. xix, 887 p.
191. Avise, J.C., *Phylogeography : the history and formation of species*2000, Cambridge, Mass.: Harvard University Press. viii, 447 p.
192. Church, S.A., et al., *Evidence for multiple Pleistocene refugia in the postglacial expansion of the eastern tiger salamander, Ambystoma tigrinum tigrinum.* Evolution, 2003. **57**(2): p. 372-83.
193. Zamudio, K.R. and W.K. Savage, *Historical isolation, range expansion, and secondary contact of two highly divergent mitochondrial lineages in spotted salamanders (Ambystoma maculatum).* Evolution, 2003. **57**(7): p. 1631-52.
194. Walker, D. and J. Avise, *Principles of phylogeography as illustrated by freshwater and terrestrial turtles in the southeastern united states.* Annual Review of Ecology and Systematics, 1998. **29**: p. 23-58.
195. Tollis, M., et al., *Multi-locus phylogeographic and population genetic analysis of Anolis carolinensis: historical demography of a genomic model species.* PLoS One, 2012. **7**(6): p. e38474.
196. Campbell-Staton, S.C., et al., *Out of Florida: mtDNA reveals patterns of migration and Pleistocene range expansion of the Green Anole lizard (Anolis carolinensis).* Ecol Evol, 2012. **2**(9): p. 2274-84.
197. Lane, E., *Florida's geological history and geological resources.* Special publication (Florida Geological Survey (1989)) ; no. 35.1994, Tallahassee, FL: Published for the Florida Geological Survey,.
198. Petuch, E.J., *Cenozoic seas : the view from eastern North America*2004, Boca Raton: CRC Press. 308 p.
199. Guiher, T.J. and F.T. Burbrink, *Demographic and phylogeographic histories of two venomous North American snakes of the genus Agkistrodon.* Mol Phylogenet Evol, 2008. **48**(2): p. 543-53.
200. Yang, Z. and B. Rannala, *Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds.* Mol Biol Evol, 2006. **23**(1): p. 212-26.
201. Weir, J.T. and D. Schluter, *Calibrating the avian molecular clock.* Mol Ecol, 2008. **17**(10): p. 2321-8.
202. Knowles, L.L. and B.C. Carstens, *Estimating a geographically explicit model of population divergence.* Evolution, 2007. **61**(3): p. 477-93.
203. Ronquist, F., et al., *MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.* Syst Biol, 2012. **61**(3): p. 539-42.
204. Guindon, S., et al., *Estimating maximum likelihood phylogenies with PhyML.* Methods Mol Biol, 2009. **537**: p. 113-37.
205. Hudson, R.R., M. Slatkin, and W.P. Maddison, *Estimation of levels of gene flow from DNA sequence data.* Genetics, 1992. **132**(2): p. 583-9.
206. Heled, J. and A.J. Drummond, *Bayesian inference of species trees from multilocus data.* Mol Biol Evol, 2010. **27**(3): p. 570-80.
207. McCormack, J.E., et al., *Calibrating divergence times on species trees versus gene trees: implications for speciation history of Aphelocoma jays.* Evolution, 2011. **65**(1): p. 184-202.

208. Zheng, Y., et al., *Exploring patterns and extent of bias in estimating divergence time from mitochondrial DNA sequence data in a particular lineage: a case study of salamanders (order Caudata).* Mol Biol Evol, 2011. **28**(9): p. 2521-35.
209. Ho, S.Y., et al., *Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.* Mol Biol Evol, 2005. **22**(7): p. 1561-8.
210. Peterson, G.I. and J. Masel, *Quantitative prediction of molecular clock and ka/ks at short timescales.* Mol Biol Evol, 2009. **26**(11): p. 2595-603.
211. Neill, W., *Historical Biogeography of Present-Day Florida*. Bulletin of the Florida State Museum, ed. W. Riemer. Vol. 2. 1957, Gainesville, FL: University of Florida.
212. Barton, N.H., *The role of hybridization in evolution.* Mol Ecol, 2001. **10**(3): p. 551-68.
213. Staubach, F., et al., *Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus).* PLoS Genet, 2012. **8**(8): p. e1002891.
214. Allendorf, F.W. and G. Luikart, *Conservation and the genetics of populations*2007, Malden, MA ; Oxford: Blackwell Pub. xix, 642 p.
215. Gompert, Z., *Population genomics as a new tool for wildlife management.* Mol Ecol, 2012. **21**(7): p. 1542-4.
216. Gompert, Z., T.L. Parchman, and C.A. Buerkle, *Genomics of isolation in hybrids.* Philos Trans R Soc Lond B Biol Sci, 2012. **367**(1587): p. 439-50.
217. Rosenberg, N.A., *Distruct: A program for the graphical display of population structure.* Molecular Ecology Notes, 2004(4): p. 137-138.
218. Tollis, M. and S. Boissinot, *The transposable element profile of the anolis genome: How a lizard can provide insights into the evolution of vertebrate genome size and structure.* Mob Genet Elements, 2011. **1**(2): p. 107-111.
219. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
220. Charlesworth, B., P. Sniegowski, and W. Stephan, *The evolutionary dynamics of repetitive DNA in eukaryotes.* Nature, 1994. **371**(6494): p. 215-20.
221. Le Rouzic, A. and G. Deceliere, *Models of the population genetics of transposable elements.* Genet Res, 2005. **85**(3): p. 171-81.
222. Petrov, D.A., et al., *Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila.* Mol Biol Evol, 2003. **20**(6): p. 880-92.
223. Langley, C.H., et al., *On the role of unequal exchange in the containment of transposable element copy number.* Genet Res, 1988. **52**(3): p. 223-35.
224. Nuzhdin, S.V., E.G. Pasyukova, and T.F. Mackay, *Positive association between copia transposition rate and copy number in Drosophila melanogaster.* Proc Biol Sci, 1996. **263**(1372): p. 823-31.
225. Brookfield, J.F. and R.M. Badge, *Population genetics models of transposable elements.* Genetica, 1997. **100**(1-3): p. 281-94.
226. Lohmueller, K.E., et al., *Proportionally more deleterious genetic variation in European than in African populations.* Nature, 2008. **451**(7181): p. 994-7.
227. Slatkin, M. and L. Excoffier, *Serial founder effects during range expansion: a spatial analog of genetic drift.* Genetics, 2012. **191**(1): p. 171-81.
228. Clark, A.G., et al., *Ascertainment bias in studies of human genome-wide polymorphism.* Genome Res, 2005. **15**(11): p. 1496-502.
229. Neafsey, D.E., J.P. Blumenstiel, and D.L. Hartl, *Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies.* Mol Biol Evol, 2004. **21**(12): p. 2310-8.
230. Basta, H.A., A.J. Buzak, and M.A. McClure, *Identification of novel retroid agents in Danio rerio, Oryzias latipes, Gasterosteus aculeatus and Tetraodon nigroviridis.* Evol Bioinform Online, 2007. **3**: p. 179-95.
231. Akagi, K., et al., *Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition.* Genome Res, 2008. **18**(6): p. 869-80.
232. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
233. Bowen, N.J. and I.K. Jordan, *Exaptation of protein coding sequences from transposable elements.* Genome Dyn, 2007. **3**: p. 147-62.

234. Petrov, D.A., *DNA loss and evolution of genome size in Drosophila.* Genetica, 2002. **115**(1): p. 81-91.