THE DIVERSITY AND EVOLUTION OF TRANSPOSABLE ELEMENTS IN THE GENOME OF THE LIZARD *Anolis carolinensis*

by

PETER ANTHONY NOVICK

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.

2010

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 _____

Date Chair of Examining Committee

Laurel Eckhardt _____

Date Executive Officer

Frank T. Burbrink _____

Anthony V. Furano _____

Michael J. Hickerson _____

John Waldman _____

Supervision Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE DIVERSITY AND EVOLUTION OF TRANSPOSABLE ELEMENTS IN THE GENOME OF THE LIZARD *Anolis carolinensis*

by

Peter Anthony Novick

Advisor: Stéphane Boissinot, PhD.

Eukaryotic genomes are littered with repetitive DNA sequences called transposable elements (TEs). Though once considered "junk DNA," these elements can greatly impact their host genomes by influencing genome size, providing novel proteins and promoter sequences, as well as disrupting gene function or causing chromosomal rearrangements. However, the impact TEs have on their host's genome depends on their abundance and diversity, which differ greatly among vertebrate genomes. The genome of teleostean fish contain a very diverse community of elements that are represented only by recent inserts found in very low copy number (<100). On the other hand, most mammalian genomes have a very low diversity of TEs dominated by L1 retrotransposons, yet elements in mammals accumulate to reach extraordinary copy numbers (>100,000). This difference accounts, for the most part, for the difference in genome size between these two groups. Until recently, we did not have a good model to study the transition from the small repeat-poor genome of fish to the larger repeat-rich genome of mammals. The first non-avian reptile genome sequence, the lizard *Anolis carolinensis* (the North American green anole), bridges the large phylogenetic gap between fish and mammals and provides a better understanding of early amniotes genomic evolution. We performed the

first comprehensive analysis of TEs in the anole genome. We found that the anole genome contains an extraordinary diversity of active TEs. This genome contains several concurrently active clades of non-LTR retrotransposons (CR1, L1, L2, RTE, and R4) each represented by multiple families. The vast majority of insertions are very young, suggesting that most elements do not reach fixation and when they do, they decay rapidly. In addition the anole genome is inhabited by multiple superfamilies of DNA transposons (*hAT, Helitron, Maverick and Chapaev*), some of which were laterally transferred to the anole. We conclude that the genomic landscape of the lizard is strikingly similar to the one of fish and shows little resemblance to mammalian genomes.

ACKNOWLEDGEMENTS

First and foremost I would like to thank the Broad Institute Genome Sequencing Platform and Genome Sequencing and Analysis Program, as well as the sequencing team, led by Federica Di Palma, Jessica Alfoldi and Kerstin Lindblad-Toh, for making the data for *Anolis carolinensis* available. Without their hard work sequencing the genome, none of my research could have been completed. I would also like to thank Dr. William Ferguson for his support and training at the beginning stages of my research as well as the rest of the Boissinot lab both past and present, Marc Tollis, Xenia Froelich, Akash Sookdeo, Eryn Blass, Lauren Alvarez, Shira Dvora and Ann Duong. A very special thank you is due to my undergraduate assistant Mark Floumanhaft for his hard work assisting me in my multiple projects. I would like to acknowledge my collaborators who helped me collect these oftentimes difficult to find sequences; Dr. Marcella McClure and Holly Basta (Montana State University, Bozeman) as well as Dr. David Ray and Jeremy Smith (West Virginia University). To Dr. Peter Reddien (MIT), Dr. Michael Steiper (Hunter College), Dr. Mark Batzer (Louisiana State University) and Doug Dellecave from D&J Reptiles, thank you for your generous donations of tissues, whole organisms and DNA that immensely furthered my research. To my family, Linda, Pete, Jackie and Joe: thank you for your support and patience over the past 4 years; I couldn't have done it without you all. Additionally, thank you to the staff of Queens College and the Graduate Center for your help. Thank you so much to my committee, Dr. Frank T. Burbrink, Anthony V. Furano, Michael J. Hickerson and John R. Waldman for their, questions, ideas and time. Lastly, and most importantly, thank you Stéphane Boissinot for your guidance, hard work and assistance in the two chapters that you coauthored and are currently in print, and the third that is under review. I could not have had a better mentor who pushed me, trained me and watched over me. You welcomed me into your lab with open arms and an open mind, and for that I am eternally grateful. This research was supported by PSC-CUNY grant 61542-00-39 to Stéphane Boissinot and Molecular experiments were conducted in part with equipment from the Core Facility for Imaging, Cellular and Molecular Biology at Queens College.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The Diversity, Evolution, Replication and Impact of Transposable Elements in Eukaryotic Genomes

Transposable elements (TEs) are mobile DNA sequences found in all eukaryotes. TEs can be separated into two classes based on their mode of transposition. Class I TEs (retroelements) replicate via a copy-and-paste mechanism that requires an mRNA intermediate and the enzyme reverse transcriptase (Luan et al. 1993). These retroelements can further be subdivided into two monophyletic groups: those that contain Long Terminal Repeats, (LTR)-retrotransposons, and those that do not, non-LTR retrotransposons (sometimes referred to as Long Interspersed Nuclear Elements (LINEs)). Class II transposable elements, or DNA transposons, replicate directly from DNA to DNA via a cut-and-paste mechanism that requires the enzyme transposase.

TEs have played an extremely important role in shaping the size, structure and function of vertebrate genomes. However, the impact TEs have had on the genome of their host varies considerably because of large differences in the number and diversity of elements among vertebrate classes. These differences are still unexplained and may involve differences in the regulation of retrotransposition/transposition by the host, competitive interactions between families for host factors, the intensity of selection against new inserts and the history of the population (Furano, Duvernell, and Boissinot 2003). Until recently, our knowledge of TE diversity has been limited to fish and mammals. This is because the only available vertebrate genomes were predominantly teleostean fish (the zebrafish, *Danio rerio*, the green spotted pufferfish, *Tetraodon nigroviridis*, the Japanese pufferfish, *Takifugu rubripes*, the medaka, *Oryzias latipes* and the threespine stickleback, *Gasterosteus aculeatus*) and mammals (human, *Homo sapiens*, chimpanzee, *Pan troglodytes,* mouse, *Mus musculus*, the Norway rat, *Rattus norvegicus*, cat, *Felis catus*, dog, *Canis familiaris,* and cow, *Bos taurus*). In fact, even today, there are only 3 non-fish and non-mammal vertebrates that have been sequenced (the frog, *Xenopus tropicalis*, the chicken, *Gallus gallus* and the lizard, *Anolis carolinensis*). The poor

representation of amphibians and reptiles can mostly be attributed to researcher's interest as well as the organism's limited use in medical, developmental and disease research.

Fish genomes tend to be small (fugu and tetraodon <400Mb), repeat-poor, but harbor a very high diversity of active TEs (Furano, Duvernell, and Boissinot 2003). Though typically in low copy number, many active clades represented by multiple active families of non-LTR retrotransposons as well as multiple active lineages of DNA transposons have been detected in sequenced fish genomes. For instance, the zebrafish genome contains several active lineages of non-LTR retrotransposons belonging to the CR1, L1, I, RTE, L1 and R2 clades as well as DNA transposons of the *hAT*, *Harbinger*, *Helitron* and *Mariner* families (Basta, Buzak, and McClure 2007; Repbase). Avian genomes are also small (chicken 1.2Gb) but they are highly impoverished in repetitive elements. Though dominated by one clade of non-LTR retrotransposons, CR1 (Chicken Repeat 1), these elements are divergent and fragmented indicating their absence of activity and most likely extinction (Abrusan et al. 2008; Shedlock et al. 2007). Finally, mammalian genomes are large (human/chimp >3Gb), repeat-rich, but typically contain only one active clade, the L1 clade, that evolves as a single lineage. In addition, almost all mammalian genomes lack recent activity from DNA transposons (Lander et al. 2001 and Waterston et al. 2002). The transition from the small genome of fish to the large, repeat-rich genome of mammals constitutes one of the most important transitions in vertebrate evolution (Figure 1). However, this important transition is not understood because of the poor representation of amphibian and reptilian genomes. The sequencing of the first non-avian reptile genome, the *Anolis carolinensis* genome, fills this large Phylogenetic gap between fish and mammals. Analyzing the diversity and evolution of TEs in *Anolis* will therefore yield important information about genomic evolution in amniotes and the evolutionary forces acting upon their copy numbers and diversity.

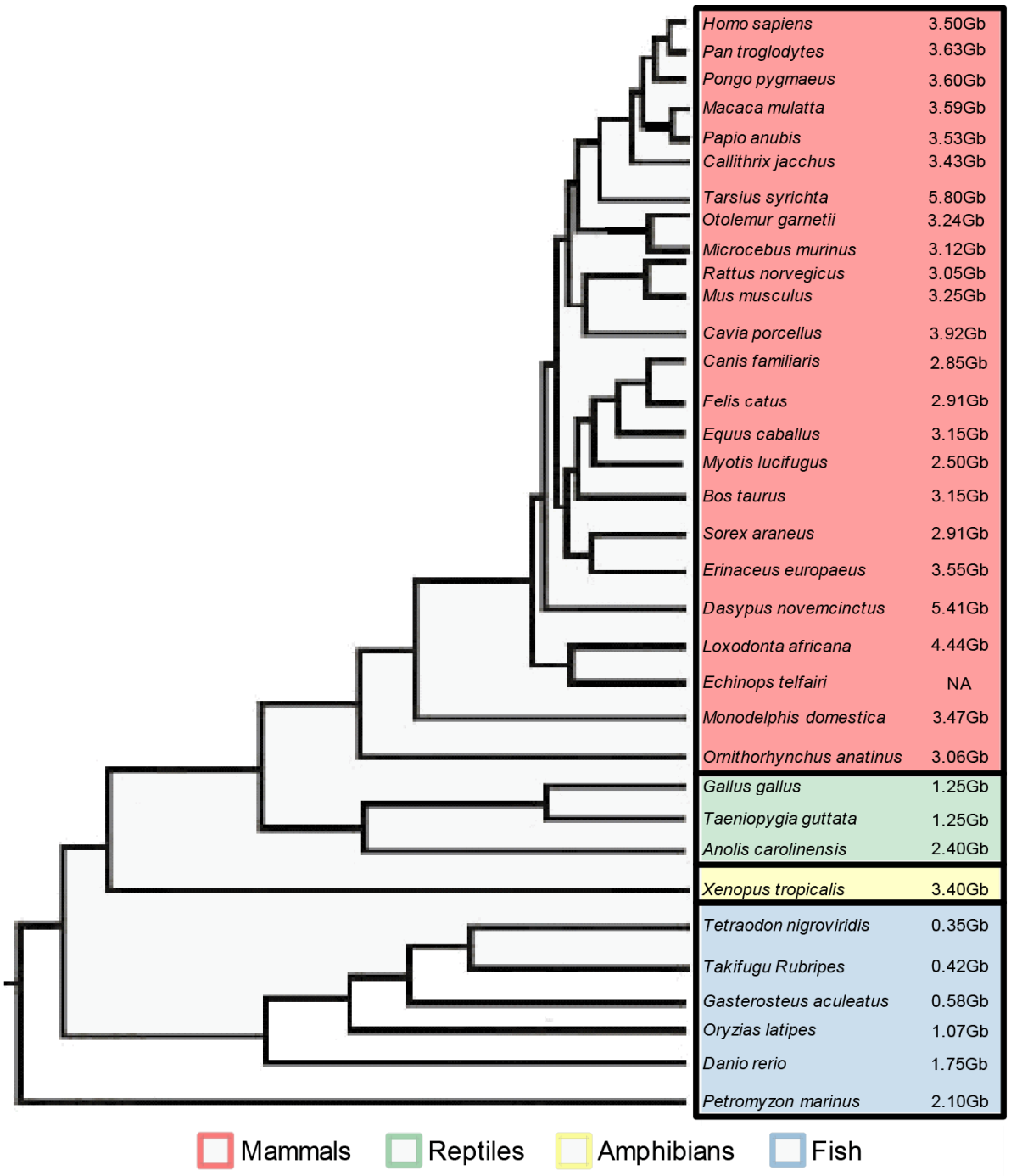

Figure 1: Phylogenetic relationships among vertebrates for which sufficient genome data is available (each class is separated by color). Genome size in billion base pairs (Gb) were collected from the most recent publications on the genome size database (Gregory, 2009) and are delineated to the right of each species name.

Structure and Retrotransposition Process of Non-LTR Retrotransposons

There are 12 monophyletic clades (Figure 2) of non-LTR retrotransposons which diverged before the diversification of eukaryotes (Malik, Burke, and Eickbush 1999; Lovsin, Gubensek, and Kordis 2001). Elements belonging to different clades differ slightly in structure (ie. length and number of coding regions) yet the overall organization is conserved. Most research has been conducted on mammalian LINE-1 (L1) and will be used here as a model. L1 elements are traditionally 6-8kB in length and are composed of a 5' untranslated region (UTR) containing an internal promoter, two open reading frames (ORFs), a 3' UTR and a poly-A tail at their 3' UTR (Figure 3). ORF1 encodes an RNA binding protein that mediates interactions of the protein with the mRNA transcript and contains a coiled-coil domain for protein-protein interactions (Martin and Bushman 2001; Martin, Li, and Weisz 2000). ORF2 contains an endonuclease domain (Feng et al. 1996) and a reverse transcriptase domain (Mathias et al. 1991) to nick the target site and reverse transcribe the mRNA at its new locus.

The retrotransposition process begins with the transcription of the non-LTR retrotransposons from the internal promoter and the resulting mRNA transcript leaves the nucleus. After the L1 proteins are synthesized they associate with the mRNA, and the resulting ribonucleoprotein complex re-enters the nucleus. The self-coded endonuclease then locates a target site, creates a nick in the genomic DNA, and the poly-A tail at the 3' end of the transcript anneals to the exposed target site. The self coded reverse transcriptase then synthesizes the DNA of the new insertion towards the 5' end using the L1 RNA as a template. A second cut is made 5-20bp downstream of the first cut but on the other strand. The first DNA strand is then used as a template by host proteins to complete the insertion. As these elements copy-and-paste, each retrotranspositional event yields a net increase of one element. This process is called Target Primed Reverse Transcription (TPRT) because the reaction of retrotransposition

takes place at the insertion site. It was initially demonstrated for the R2Bm elements in the silkworm *Bombyx mori* and subsequently for L1 (Luan et al. 1993; Cost et. al. 2002).

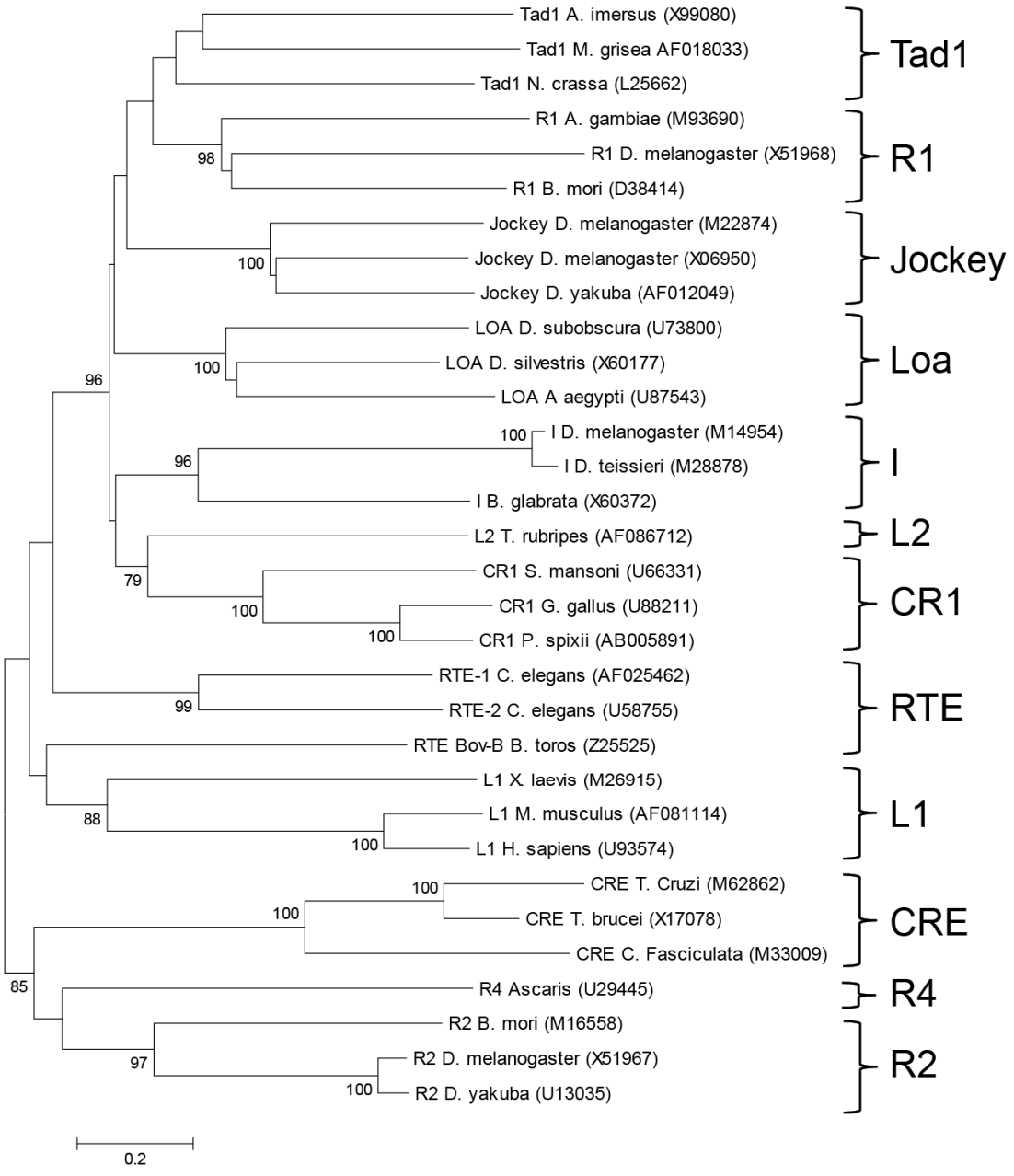

Figure 2: Neighbor joining tree of the 12 clades of non-LTR retrotransposons based upon 300aa of the RT domain in ORF2. Bootstrap values less than 75% have been removed and genbank accession numbers are provided next to the species and clade.

The nick that is generated by the endonuclease cuts the site so that after the element has been inserted, the target site is repeated on both the 5' and 3' ends of the newly inserted element, forming a Target Site Duplication (TSD). Oftentimes, portions of the element exhibit microhomology to the 3' end of the TSD and anneal prematurely to the target site thus ending the transposition process. As elements are inserted from the 3' end, premature termination will create a truncated (TR) element (lacking regions of the 5' end including ORFs or the 5' UTR), a novel genomic insertion that is not capable of making additional copies. As this is frequently the case, most novel insertions are deemed as "dead-on-arrival;" these elements are no longer capable of retrotransposition and will mutate at the neutral rate. (Martin et al. 2005)

The machinery encoded by non-LTR retrotransposons can act on other transcripts and has been parasitized by non-autonomous elements called Short INterspersed Elements, or SINEs. These elements vary in length from 70 to 500bp and typically possess a region structurally related to tRNA or 7S RNA (Jurka and Zuckerlandl 1991). SINEs may also share very similar TSDs or a 3' end similar to their autonomous LINE families (Ohshima and Okada 2005). Their mobility has been demonstrated for *Alu* SINEs in humans, as well as for SINEs found in the eel (Kajikawa and Okada 2002; Dewannieux, Esnault, and Heidmann 2003).

Structure and Transposition of DNA Transposons

Class II TEs (DNA transposons) can be subdivided into 15 monophyletic superfamiles belonging to three main categories. The most diverse group is the classic "cut-and-paste" transposons consisting of 12 superfamilies (including *hATs* and *Mariners*). Two other groups of transposons, *Helitrons* and *Polintons* have been identified recently with a different, yet not completely understood method of transposition (Feschotte and Pritham 2007).

Autonomous DNA transposons vary in length (1.2-25kb) but always encode the enzyme transposase (Figure 3). These transposase domains are related to ancestral bacterial Insertion Sequences (IS) which behave similarly in bacterial genomes. As the transposase of elements from each superfamily is most closely related to one IS, each of the 12 superfamilies are monophyletic (Feschotte and Pritham 2007). Most DNA transposons also create TSDs upon their insertion at a novel locus but uniquely possess Terminal Inverted Repeats (TIRs), which are portions of the element at the 5' and 3' end that are exactly the same but in the reverse orientation (*Helitrons* lack TIRs) (Pritham, Putliwala, and Feschotte 2007) .

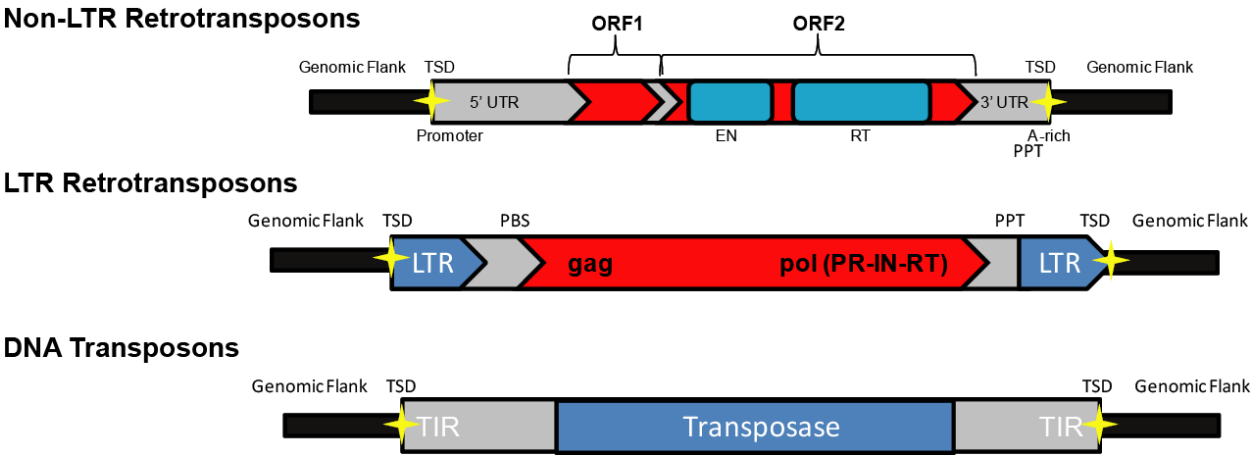

Figure 3: Structure of a typical autonomous non-LTR retrotransposon, LTR-retrotransposon and a copy-and-paste DNA transposon. Abbreviations include: ORF (Open Reading Frame), CC (Coiled-Coil domain), UTR (Untranslated Region), EN (Endonuclease), RT (Reverse Transcriptase), TSD (Target Site Duplication), PBS (Primer Binding Site), PPT (Polypurine Tract), GAG (Group AntiGens), IN (INtegrase), POL (POLyprotein) and PR (protease).

Although the transpositional mechanism is not yet fully understood for all categories of class II elements, it is known that the self encoded transposase of cut-and-paste elements recognizes the TIRs at the 5' and the 3' ends of the element for its excision. Once excised, the resulting transposon/enzyme complex forms a loop and will locate a new target site. A nick is generated at the target site, and the transposon will reinsert beginning from the extreme ends of

the element. The length and sequence of the TSDs and terminal motifs of the TIRs are highly conserved across superfamilies and are useful in categorizing elements (Feschotte and Pritham 2007).

Although these elements “cut-and-paste” themselves (non-replicative), they still have the ability to increase in number in indirect manners. As demonstrated in corn, the transposition of a DNA element can be initiated during the replication of a chromatid and inserted in a novel site (Kunze and Weil 2002). Additionally, the double-strand break created by the excision of the donor DNA transposon may be repaired. If the DNA element is also found on the homologous chromosome, it may be used as a template to repair the excised site from both the 5’ and 3’ ends, yielding a maximum increase of one element for each transpositional event (Hsia and Schnable, 1996). This process has been demonstrated in the fruit fly (Engels, Johnson-Schlitz, 1990).

Oftentimes, gap repair is interrupted and prematurely aborted resulting in the incomplete repair of the donor site and an element that is missing portions of its central region. Additionally, the transposition mechanism of the original donor may be aborted, again resulting in an element that is missing a section of the central sequence (Hsia and Schnable 1996). As the terminal motifs of the TIRs are the only requirement for the recognition of the transposase, these elements can also transpose and yield families of non-autonomous transposons (Craig et al. 2002). In fact, non-autonomous elements oftentimes outnumber their autonomous protein coding counterparts (Feschotte, Zhang, and Wessler 2002). As the genome may only repress autonomous elements, those that are not coding for proteins may go unrecognized and proliferate in the genome. Another reason for their high abundance is that many non-autonomous elements (specifically Miniature Inverted-Repeat Transposable Elements (MITEs)) are rather small (100-600bps) compared to their autonomous counterparts. Smaller elements

are possibly less likely to undergo inter-element recombination, and therefore removal from the genome (Feschotte, Swamy, and Wessler 2003).

Influence of Transposable Elements on their Host Genome

Though once deemed useless (the "junk DNA" view), these "selfish" parasites greatly impact the function and evolution of their host genomes. First observed in 1951, a corn transposon was demonstrated to influence a nearby expressed gene (McClintock 1951). In fact, TEs provide an extraordinary source of evolutionary changes that may be beneficial, neutral or deleterious to their hosts via: (i) the alteration of gene function, (ii) gene transduction and duplication, (iii) genetic deletions due to homologous and non-homologous inter-element recombination, (iv) increase or decrease in genome size, (v) exaptation or domestication of elements forming novel exons or promoter/regulatory regions (Kidwell and Lisch 1997; Feschotte and Pritham 2007).

As TEs insert copies throughout the genome, insertions in coding regions may create a mutant allele which could directly or indirectly cause genetic defects including, in human, cancer, and hemophilia (Chen et al. 2005; Ostertag and Kazazian 2001). TEs are also capable of partial or whole gene transduction or duplication. 3' flanking sequences may be transposed together with a retrotransposon when the transcription of the progenitor fails to stop at the poly-A tail (gene transduction). Genes can also be duplicated when the enzymatic machinery of retrotransposons acts on an RNA transcript that is not associated with a transposable element. Sometimes, these retroprocessed genes can remain functional duplicates leading to novel gene families; other times, the resulting sequences are pseudogenes, a non-functional twin (Dewannieux, Esnault, and Heidmann 2003). Yet, unlike SINEs which generate thousands of copies from parasitizing off their hosts machinery, pseudogenes are generally very low in copy number (Esnault, Maestre, and Heidmann 2000). Retrotransposons are not the only type of

TEs that can increase gene copy number. During the gap repair process of DNA transposons, external portions of genomic DNA, including genes, may be incorporated into the central region of the novel or donor site. For instance, in *Arabidopsis*, over 3000 TEs contained exons from an assortment of genes. It is possible that this process may lead to the creation of novel genes as 20% of these elements contained a chimer of exons from multiple genes (Jiang et al. 2004; Lisch 2002). The aforementioned process has since been demonstrated in other organisms as well as in multiple superfamilies of DNA transposons (Kawasaki and Nitasaka 2004; Zabala and Vadkin 2005).

TEs are also prone to undergo interelement recombination, which in turn may lead to large chromosomal rearrangements and the deletion of genomic DNA. In human, only full length elements are negatively selected (truncated LINEs as well as SINEs were not). Though full length elements contain promoter regions and may contain intact ORFs, the foremost reason they are deleterious is their higher abillity to undergo ectopic recombination (Boissinot et al. 2006; Boissinot and Song 2007). Consequently, full length L1 elements tend to accumulate in low recombining regions, such as the Y chromosome (Boissinot, Entezam, and Furano 2001). This effect is not limited to LINEs as .03% of human genetic illness is related to inter-SINE recombination events resulting in deletions or duplications of exonic regions (Callinan and Batzer 2006).

Since TE's accumulate in some host genomes, they can also greatly influence genome size. L1 elements in mammalian genomes account for about 20% of its size and 50% of the human genome is comprised of TEs of all classes (Lander et. al., 2001). Similarly in plants, the genome size of *Oryza australiensis* (compared to its sister taxa) has doubled due to TE activity (Piegu et al. 2006). However, TE activity may also have an opposite effect on genome size. Genome shrinking has been reported in some flowering plants due to inter-element

recombination at non-homologous sites. The genomes for which this has been reported have demonstrated the rapid removal of TEs and genomic DNA by small, yet frequent deletions from illegitimate recombination (Bennetzen, Ma, and Devos 2005).

At first glance, the many deleterious effects of TEs suggest that they are strictly harmful; however, TEs may also be advantageous to their host. Exaptation of TEs appears to have been relatively common during the evolution of vertebrates and retrotransposon inserts may affect gene expression providing novel promoter and inhibitory sequences to nearby genes (Lunyak et al. 2007; Miskey, Izvak, and Kawakami 2005). Analysis of human TE inserts revealed over 10,000 TE fragments that evolved under purifying selection, especially near genes important in development, and it is estimated that 25% of human promoters contain TE-derived sequences (Lowe, Bejerano, and Haussler 2007; Jordan et al. 2003).

It is also not uncommon for protein coding regions of TEs (mostly transposase) to be domesticated for other genomic functions. For example, the human genome alone contains ≈50 TE derived genes (Lander et al. 2001). Similar domestication has since been observed in other vertebrates as well as drosophila and various species of grass (Zdobnov et al. 2005; Casola et al. 2007; Muehlbauer et al. 2006). In most cases, the domesticated proteins come from DNA transposons and not retrotransposons. This is due to the fact that DNA transposons contain an N-terminal domain that is for DNA binding, and a C-terminal core for the cleavage of the donor element. Both of these functions can be domesticated by the cell for DNA repair, DNA replication and cell death induction (Feschotte and Pritham 2007).

Mode of Transmission and Evolutionary Dynamics of Transposable Elements

The 12 clades of non-LTR retrotransposons diverged from each other more than 600 MYA before the diversification of eukaryotes and are found in plants, animals, fungi and protists

(Malik, Burke, and Eickbush 1999; Volff, Korting, and Schartl 2000). In amniotes, four out of the twelve clades, L1, L2, CR1 and RTE, have amplified to significant levels and have had the most impact on amniote evolution. Elements belonging to different clades can be further divided into families based on shared differences between elements (Furano 2000). Class I elements typically display a strict vertical mode of inheritance.

The L1 clade has been extremely successful at invading the genome of eutherian mammals generating tens of thousands of copies that remain in the genome as DNA fossils (Furano, Duvernell, and Boissinot 2004). The platypus, a monotreme, has had a similar invasion of non-LTR retrotransposons, but instead of L1, another clade, L2, dominates the genome (Warren et al. 2007). The chicken genome as well as other reptilian genomes seems to be dominated by CR1-like elements (Waterston et al. 2002; Shedlock et. al. 2007). A final clade, RTE, has a patchy distribution in fish, reptiles and bovids. Though the vertical inheritance model is predominant, one family of RTE elements, RTE Bov-B shows extreme homology in reptiles and bovids revealing their probable horizontal transmission (Kordis and Gubensek 1998; Zupunski, Gubensek, and Kordis 2001). Opposite to this single species, single clade patterns, the genomes of fish contain multiple active lineages of L1, L2, and CR1, as well as other non-LTR clades like RTE, R2 and R4 (Kordis, Lovsin, and Gubensek 2006). From this view, it is apparent that mammalian genomes operate under different constraints than the genome of non-mammals. It is possible to infer that the ancestral genome of amniotes contained families of elements from each of four clades (L1, L2, CR1 and RTE) (Shedlock et al. 2007), each probably represented by moderately high copy numbers, and that those clades have had variable success in amplifying in the genome of amniotes. The reduction in diversity of retrotransposons and the explosion in copy numbers of some clades represent one of the major transitions in the evolution of amniote genomes.

DNA transposons are also found in all eukaryotic kingdoms; however, this class of elements was once deemed extinct in mammals and their study has mostly been limited to plants, fungi, and insects (Lander et al. 2001; Waterston et al. 2002; Pace and Feschotte 2007). In vertebrates, teleost fish genomes (including salmonoids, fugu and zebrafish) are littered with DNA transposons (Tc1/Mariners) (Krasnov et al. 2005). As for mammals, the initial sequencing and analysis of the human genome revealed that there were no DNA transposons active in the past 37MY (Pace and Feschotte 2007). Moreover, only 1.5% of the human genome is composed of highly degenerate DNA transposons (Pace and Feschotte 2007). Similar analysis of other mammalian genomes, dog, mouse and rat, revealed a similar trend indicating that DNA transposons went extinct independently in those mammalian lineages (Gibbs et al. 2004; Lindbaldh-Toh et al. 2005; Waterston et al. 2002). However, class II transposable elements are not quite extinct in mammals. The recent discovery of active elements in the genus *Myotis* (bats) and the common incidence of horizontal transfer of DNA elements across distantly related animal species have stimulated the idea that these elements may in fact still be active in the genomes of other vertebrates including reptiles (Ray et al. 2007; Ray et al. 2008).

The Iguanian Lizard *Anolis carolinensis*

Comparative genomics has become an ever-growing field with the advent of faster, more efficient methods of whole genome shotgun sequencing; while fourteen mammals have been sequenced, their sister group, the reptiles, lags well behind. Until February of 2007, the only sequenced reptile was an avian reptile, the chicken, (grouped with crocodilians in the Archosaurian lineage) and as its genome was depauperate in TEs (Shedlock et al. 2007) it provided little insight into the evolution and diversification of mammalian genomes. The remaining reptiles may be separated into two groups: the Lepidosauria (Squamates Sphenodons) and Testudines (turtles). Together, these organisms comprise the last major

vertebrate group to be sequenced. More than 7,000 extant species of non-avian reptiles have been described, displaying a wide range of life histories and reproductive modes; found on all continents (except Antarctica), they have colonized virtually all environments (from deserts to rain forests). Mammals diverged from the evolutionary line that led to extant reptiles between 312 and 330MYA (Donoghue and Benton 2007). Therefore, analyzing the genome of a squamate lizard will root the previously sequenced mammalian genomes, fill a large evolutionary gap between fish and mammals, aid in identifying conserved regulatory regions, and provide further insight on gene duplication and the formation of multigene families.

The North American green Anole (also known as the Carolina Anole, Green Anole or Red-Throated Anole), *Anolis carolinensis*, inhabits the Southeastern region of the United States (Texas, Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, Georgia, Florida, South Carolina, North Carolina, Tennessee) and this distribution extends even further as populations were unintentionally introduced to Hawaii, Japan, Guam and Belize most likely as stowaways in cargo ships or via the pet trade (Malone and Davis 1998). It is one of the oldest studied members of its genus; publications date back almost 130 years (Monks 1881). The anole is a small (17-20cm) terrestrial lizard most commonly found in trees and high grass where they readily feed on small insects. Anoles are quite inquisitive and aware of their surroundings, which is important for these territorial lizards. It is quite common to find males fighting/posturing for females by expanding their neck dewlap, bobbing their heads and/or chasing and biting intruders. However, the most striking feature is their incredible ability to change color. Though termed the “green” Anole, this lizard can change colors to various shades of brown and are often incorrectly referred to as “chameleons”. This color change ability, coupled with their low cost, and ease of maintaining has made the Anole a common house pet. (Lovern, Holmes, and Wade 2004)

The genus *Anolis* is the second largest of all tetrapod vertebrates consisting of over 400 described species. Their rapid radiations on the many Caribbean islands have made them a model of choice in evolutionary studies. More specifically, *Anolis carolinensis*, has been a model organism for decades for lab-based and field-based researchers in the fields of development, neuroscience, endocrinology, evolution and behavior. This is because: (1) they reproduce easily in captivity, (2) are abundant in the field and simple to locate. Anoles have been used to study a variety of human diseases, disorders and traits. Due to their "split brain" (lacking a corpus callosum which connects the left and right lobes of mammalian brains), anoles have been used to study seizure disorders, tic disorders, Tourrette's Syndrome and Obsessive Compulsive Disorder (OCD) (Baxter et al. 2001). Other human illnesses such as severe and manic depression have also been studied as the anole's parietal eye is easily manipulated in the lab disrupting their circadian rhythm (Underwood and Calaban 1987).

A little more than two years have passed since the initial genome sequencing, and though not yet anchored to chromosomes, the 7,233 scaffolds are still highly informative. Preliminary research on the anole revealed a high abundance of CR-1 elements (>200,000) but a paucity of other non-LTR retrotransposon families as well as DNA transposons (Shedlock et al. 2007). Other studies of the genome have revealed a variety of non-autonomous t-RNA derived SINE elements (Sauria SINEs) that are mobilized by RTE Bov-B indicating the presence of yet another retrotransposon clade. Active for more than 200MY, these Sauria SINEs display homology to families of elements found in various lineages of snakes, lizards and tuataras and have already demonstrated potential as phylogenetic markers (Piskurek, Austin, and Okada 2006). However, no comprehensive analysis of the mobile elements in the Anole or any other reptile has been performed so far.

In order to get a non-biased view of the dynamics of TEs in the lizard genome, I analyzed a subset of both classes of elements. This allows me to compare the dynamics of two different classes of repetitive elements that coexist in the same genome but amplify using different machinery. First, I will present my research on the evolution and diversity of non-LTR retrotransposons in the lizard genome in order to determine if the dynamics of LINEs in reptiles is more similar to fish or mammals. Next, I surveyed the anole genome for the abundance, time of amplification and evolution of DNA transposon families. Finally, I present an unexpected discovery indicating the repeated horizontal transfer of multiple families of DNA elements into a wide range of genomes including the lizard.

CHAPTER I: NON-LTR RETROTRANSPOSONS

THE EVOLUTIONARY DYNAMICS OF AUTONOMOUS NON-LTR RETROTRANSPOSONS IN THE LIZARD *Anolis carolinensis* SHOWS MORE SIMILARITY TO FISH THAN MAMMALS

Non-LTR retrotransposons, also known as retroposons (International Committee on Taxonomy of Viruses; Hull 2001), are autonomously replicating retroid agents that lack long terminal repeats (LTR). They have considerably affected the size, structure and function of vertebrate genomes. This is exemplified by the fact that at least 30% of the genome of mammals is the result of their activity (Lander et al. 2001; Waterston et al. 2002). Although non-LTR retrotransposons were considered to be among the "junk DNA" class of repetitive elements, research has shown that they have been an extraordinary source of evolutionary novelties. Exaptation of these elements seems to have been relatively common during vertebrate evolution, either as part of coding sequences (i.e. exonization) or as regulatory elements (Makalowski 2000; Nekrutenko and Li 2001; Mikkelsen et al. 2007). In addition, the retrotransposition machinery encoded by non-LTR retrotransposons can act on other transcripts and is responsible for the amplification of SINEs (Short INterspersed Elements) and retroprocessed pseudogenes (Dewannieux, Esnault, and Heidmann 2003; Dewannieux and Heidmann 2005). Some SINEs have also been co-opted as regulatory sequences and have played a major role in the early evolution of tetrapods (Bejerano et al. 2006).

Non-LTR retrotransposons constitute a diverse group of elements that are classified into 12 monophyletic clades (Figure 4) (Burke et al. 1999; Volff, Korting, and Schartl 2000). These 12 clades diverged from each other more than 600 Million years ago and are found in most eukaryotes, including plants, animals, and fungi. Non-LTR retrotransposons encode for the proteins responsible for their own replication. They have a 5' untranslated region (UTR) that acts as an internal promoter and one or two open-reading frames (ORF), depending upon the clade. All non-LTR retrotransposons have a reverse transcriptase domain and 9 of the 12 clades encode for an apurinic/pyrimidinic endonuclease (Martin et al. 1995; Feng et al. 1996; Malik, Burke, and Eickbush 1999). The mechanism of insertion has not been resolved in all

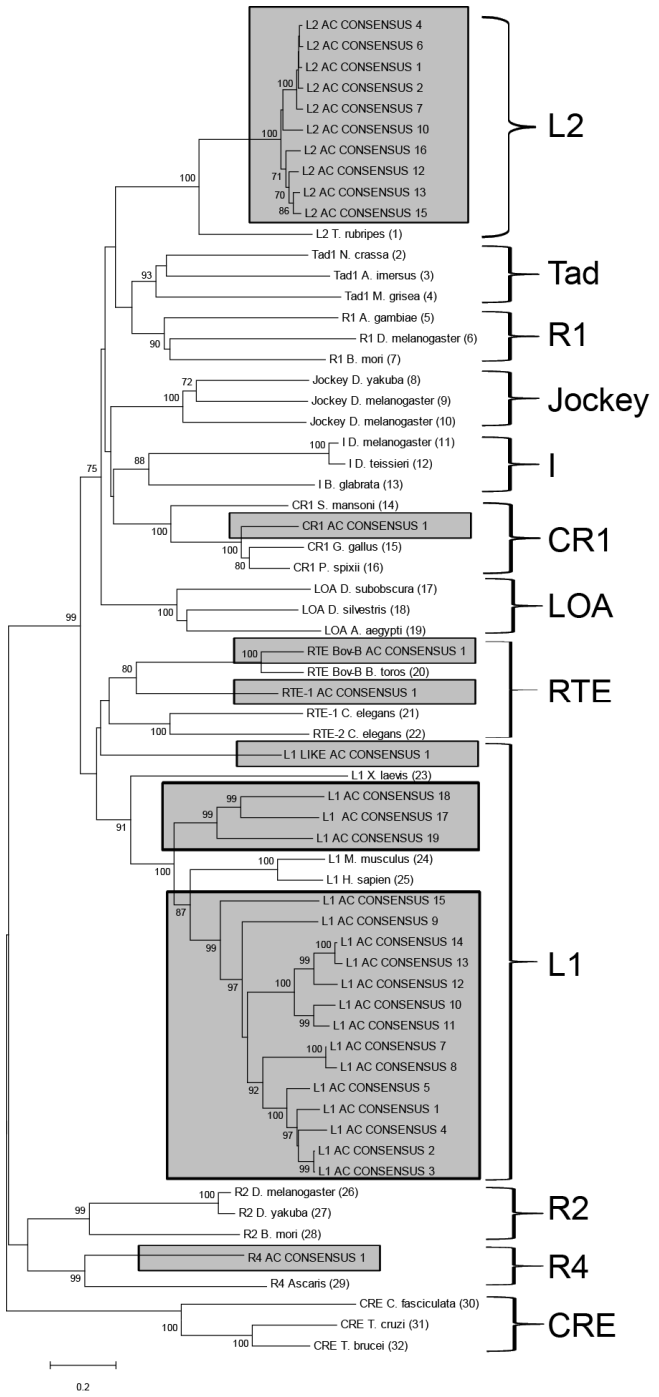

Figure 4: Phylogenetic relationships among the 12 clades of non-LTR retrotransposons. *Anolis* consensus sequences are framed in blue. This NJ tree was constructed from a portion of the translated RT domain. Bootstrap values less than 75 have been removed. Numbers next to non-Anolis sequences correspond to Genbank accession numbers listed in the methods section. The letters “AC” stand for *Anolis carolinensis.*

clades but it is believed that non-LTR retrotransposons use a nick (generated by their own endonuclease) on the host chromosome to prime reverse transcription of the RNA transcript directly into the target site (Luan et al. 1993; Luan and Eickbush 1995; Cost et al. 2002). The majority of new elements produced by this mechanism are truncated at their 5' end and incapable of further retrotransposition ("dead-on-arrival"). Elements belonging to different clades can be further subdivided into families based on shared differences between elements (Furano 2000). As the vast majority of elements do not serve a function for the host, they accumulate mutations at the neutral rate, so that old families of elements are more divergent than younger ones (Voliva et al. 1983; Pascale et al. 1993).

The analysis of complete genome sequences has revealed some striking differences in the abundance and diversity of non-LTR retrotransposons (Lander et al. 2001; Waterston et al. 2002; Volff et al. 2003; Duvernell, Pryor, and Adams 2004; Furano, Duvernell, and Boissinot 2004; Gentles et al. 2007; Warren et al. 2008). In placental mammals, a single clade, LINE-1 (Long Interspersed Nuclear Element, L1), has remained active since the split between eutherians and marsupials. L1 has been very successful in eutherians, reaching extremely high copy numbers (e.g. 500,000 copies in the human genome) (Lander et al. 2001; McClure et al. 2005). Most of these elements are ancient and are the product of past amplifications. Phylogenetic analyses have shown that most modern mammalian genomes contain a single dominant lineage of L1 families, suggesting that the evolution of L1 in mammals is controlled in ways that prevent further diversification (Furano 2000; Furano, Duvernell, and Boissinot 2004). The reason why L1 evolves as a single lineage is unknown but competition between elements for host-encoded transcription factors could explain the unusual evolution of L1 families (Khan, Smit, and Boissinot 2006).

In contrast, the genome of teleostean fish contains several active clades, each represented by divergent families (Volff, Korting, and Schartl 2000; Duvernell, Pryor, and Adams

2004; Furano, Duvernell, and Boissinot 2004; Neafsey 2004; Basta, Buzak, and McClure 2007). For instance, 32 divergent L1 families coexist in zebra-fish, each represented by small numbers (<50) of very similar elements (Duvernell, Pryor, and Adams 2004; Furano, Duvernell, and Boissinot 2004). The differences in diversity and copy number between fish and mammals could result from a number of factors, including differences in the control of retrotransposition by the host, competitive interactions between families of elements, variations in the intensity of selection against new inserts, and the history of populations. The respective importance of these different factors is unclear and remains a matter of debate (Eickbush and Furano 2002; Furano, Duvernell, and Boissinot 2004; Neafsey 2004; Kordis, Lovsin, and Gubensek 2006; Song and Boissinot 2007)**.**

The reduction in retrotransposon diversity and the explosion in copy number of the L1 clade in mammals represent one of the major transitions in the evolution of vertebrate genomes. Until recently, it was difficult to study this important question because of a lack of genomic data in amphibians and reptiles. The recent completion of the lizard (*Anolis carolinensis*) and frog (*Xenopus tropicalis*) genomes bridges a gap between fish and mammals and will provide new insights into the evolution of tetrapod genomes. *Anolis carolinensis*, the green anole, is a small lizard (Squamata: Iguanidae) found in the South-East United States that has become an important model organism in evolutionary and behavioral studies. We found that the anole genome contains at least five active clades of non-LTR retrotransposons that vary considerably in replicative success and diversity. Within most families, elements are very similar to each other suggesting that they have been recently inserted. The lack of elements of intermediate or old age relative to young elements indicates that retrotransposons accumulate in the anole genome at a low rate, possibly because the deleterious impact of retrotransposons is stronger in anoles than it is in mammals.

Materials and Methods

Sequence Acquisition

The results reported here are from the Genome Parsing Suite (GPS) software (McClure et al. 2005) used for identification and classification of *A. carolinensis* retroid content. Of all the retroid components, the reverse transcriptase (RT) is the best conserved through evolutionary time (McClure et al. 1988) and is essential for autonomous transposition. The GPS analysis is therefore centered on the RT, identifying it first and then expanding to other components. The approach of the GPS is radically different from Repeat Masker, which is used to mask out and count repetitive agents using consensus DNA sequences (Smit 2004). Repeat Masker and similar methods suffer from the loss of signal due to mutational saturation because DNA is used to query a genome rather than amino acid sequences. DNA sequence libraries are also often unable to detect new components.

The *A. carolinensis* genome v. 1.0 was downloaded from the University of Santa Cruz Genome Bioinformatics Website (Karolchik et al. 2003). The *A. carolinensis* genome has 1.7 Gb of its expected 2.2 Gb sequenced (about 77%), therefore, any numbers presented in this study are pending further genome refinement. The GPS was populated with 130 Retroid agent queries in this scan of the *A. carolinensis* genome. Although only the non-LTR retrotransposons found in full-length in the *A. carolinensis* genome are analyzed in depth in this study, queries representative of other Retroid families were included to insure correct classification of all Retroid agents. The queries include sequences that are specific to humans, birds, *Anolis*, *Xenopus* and fish, in addition to a set of 30 queries that represent the major families of all Retroid agents. As little has been classified in the *Anolis* genome, novel *Anolis* queries were selected from the GPS results using the following criteria: completeness (presence of the most components), containing the least number of stop codons and frame shifts, containing all the

motifs of the enzymatic core proteins and phylogenetically represent their family. Phylogenetic trees were constructed using all the full-length copies of each family to ensure that outliers were not selected as queries.

The GPS method is divided into two stages: stage 1 GPS deals solely with the RT, while stage 2 GPS classifies full-length (FL) agents. In Stage I GPS, Washington University Basic Local Alignment Search Tool translated nucleotides (WU-tBLASTn) version 2.0 (Gish 1996-2004) was used to query the *A. carolinensis* genome with the following parameters: E=1, -matrix pam70, Q=9, R=1, V=1e7, B=1e7, gapL=0.307, gapK=0.13, gapH=0.7, X=15, gapX=33, gapW=44, gapS2=63, S2=41, hspmax=0, and –span. After WU-tBLASTn scans the *A. carolinensis* genome with the Retroid RT queries, Stage I GPS sorts and filters raw WU-tBLASTn hits, which are redundant and contain false positives, due to: 1) alternative alignments for a given query to a specific region, 2) cross coverage of the queries, and 3) counting as unique, a number of small hits that are actually from the same gene. After sorting by query, chromosome, polarity and reading frame the GPS compounds small hits, and removes false positives due to cross coverage on these compounded hits. The GPS removes redundancy by deleting hits that are completely covered by a longer hit to the same position, thereby preventing overestimation of the amount of potential RT genes. Single contiguous sequences, single compound hits composed of subsequences, and sets of ambiguous hits to the same position and reading direction are all considered unique RT hits. Ambiguous cases are often resolved in Stage II of the GPS. Unique hits are then assessed for quality first by degree of Ordered Series of Motifs (OSM) conservation (McClure 1991), which is made up of six highly conserved motifs that fold to form the active site of the enzyme (Kohlstaedt et al. 1992), and then by presence of frame shifts and stop codons. FL RT hits with neither frame shifts nor stop codons are labeled “perfect”. In Stage 2 GPS, each RT found by Stage 1 has its position in the host’s genome extended 7 kb upstream and downstream, creating a 14 kb (plus the size of the RT) cutout.

Using this RT-outward approach, the GPS is able to construct potential Retroid agent genomes. WU-tBLASTn is used a second time to compare each 14kb+ cutout with the RT hit's corresponding component library. If an RT hit is ambiguous between multiple queries, then each of these queries' component libraries are searched, and the query with the highest score over all components is called as the closest to the new Retroid agent. If the sequence has all the gene components in the query's genomic order it is considered FL. All genomes with one frame shift or stop codon, as well as those that are error free (perfect), are considered to be potentially active, because Retroid agents are known to overcome such mutational errors by translational recoding (http://recode.genetics.utah.edu/) thereby producing functional proteins. Note that some queries themselves may contain frame shifts and stop codons. For a more in depth discussion of the GPS, see McClure et al 2005 and Basta et al 2007. When GPS finds a sequence that has identity to known non-LTR retrotransposon components, but no UTRs, and these regions do not pull out any known UTRs when nBLAST searches are executed (Altschul et al. 1990), a test for novel UTRs is performed. When multiple copies of the sequence are available, 1 kb 5' is aligned to other 5' regions from additional copies, and 3' regions are aligned to other 3' regions, and conserved regions are considered UTRs.

Characterization of Retrotransposon Families

Extracted sequences were aligned using Clustal W in the program BioEdit (Hall 1999) and subsequently categorized into families and subfamilies based upon sequence similarity. Branching patterns of phylogenies created by neighbor joining under the Kimura 2-paramaters' distances in the program MEGA 4.0 (Tamura et al. 2007) was used as well to make sure a family had not been overlooked. A sequence from each family was then submitted to Repeatmasker (www.repeatmasker.org) to verify proper clade classification. In order to determine the ancestral progenitor, FL consensus sequences were constructed for each family.

Consensus sequences have been deposited in Repbase (http://www.girinst.org/repbase). Each consensus sequence was then used in a BLAT search and the resultant output sequences were collected for further analysis. FL copy numbers were determined from the original GPS output. Copy number of truncated elements that contained less than 200 copies was retrieved from the aforementioned BLAT output while estimates for larger families were calculated using the electronic PCR option as well as from the original GPS output. FL consensus sequences from each family were compared to each other using dotmatcher (Rice, Longden, and Bleasby 2000). Additionally, each consensus sequence was analyzed for their individual GC content using the sequence analysis option in DAMBE (Xia and Xie 2001). The genomic environment for at least 20 sequences in each family was also characterized. 50kB of 5' and 3' flank were collected from the genome browser and analyzed for their GC content using DAMBE. In order to determine the relative time of amplification of each family, at least 500bp of the reverse transcriptase domain of at least 20 elements were aligned and analyzed by Mega 4.0 and both pairwise divergence and average from the consensus were calculated using Kimura's 2-parameter's distance.

Evolutionary Relationships

Consensus sequences for each family were submitted to NCBI's ORF-Finder and Conserved Domains (Marchler-Bauer et al. 2007) to find the location and frame in which the ORFs were located. ORFs were extracted and aligned in nucleic acid and amino acid and the size of each were recorded. Previously published reverse transcriptase domains of each clade were recovered and aligned using Clustal W along with the aforementioned novel proteins. Sequence numbers correspond to those in figure 4 and can be found under the accession numbers: 1-AF086712; 2-L25662; 3-X99080; 4-AF018033; 5-M93690; 6-X51968; 7-D38414; 8-AF012049; 9-X06950; 10-M22874; 11-M14954; 12-M28878; 13-X60372; 14-U66331; 15-U88211; 16-AB005891; 17-U73800; 18-X60177; 19-U87543; 20-Z25525; 21-AF025462; 22-

U58755; 23-M26915; 24-AF081114; 25-U93574; 26-X51967; 27-U13035; 28-M16558; 29-U29445; 30-M33009; 31-M62862; 32-X17078. These sequences were then analyzed phylogenetically using the neighbor joining and maximum likelihood tree using MEGA 4.0 and PHYML online (Guindon et al. 2005), respectively.

Results

The Genome Parsing Suite (GPS) identified a total of 1888 full length (FL) elements in the anole. Further BLAST and BLAT searches of the anole genome using the sequences collected by GPS as probes failed to detect any other FL elements, although we recovered a number of truncated (TR) elements. Representatives from five of the twelve clades of non-LTR retrotransposons were detected: L1, L2, CR1, RTE and R4 (figure 4). These clades differ greatly in copy number and diversity.

Out of the five clades that inhabit the genome of *Anolis carolinensis*, L1 is the least numerous, with 170 FL and 626 TR elements, yet it is the most diverse. A phylogenetic analysis based on the two ORFs (figure 5) reveals the presence of twenty distinct L1 families, represented by very low copy number (7 to 144 elements; table 1). A group of elements (L1-like on figure 4) structurally similar to L1 did not branch with other L1 families, but instead clustered with the RTE clade (with very low bootstrap support). As the structure of these elements suggests an evolutionary affinity to L1, they were provisionally classified as L1-like. L1 families are very divergent from each other and their separation pre-dates the split between reptiles and mammals (figure 4). Within each family, FL elements appear very closely related to each other (figure 5). The phylogenetic tree also shows a near complete absence of internal branches indicating a lack of old L1 inserts in the anole. In addition, repeated BLAST and BLAT searches of the anole genome using the 3' UTR as a probe failed to detect any divergent TR L1 elements. The recent origin of most L1 inserts is confirmed by the very low level of divergence within

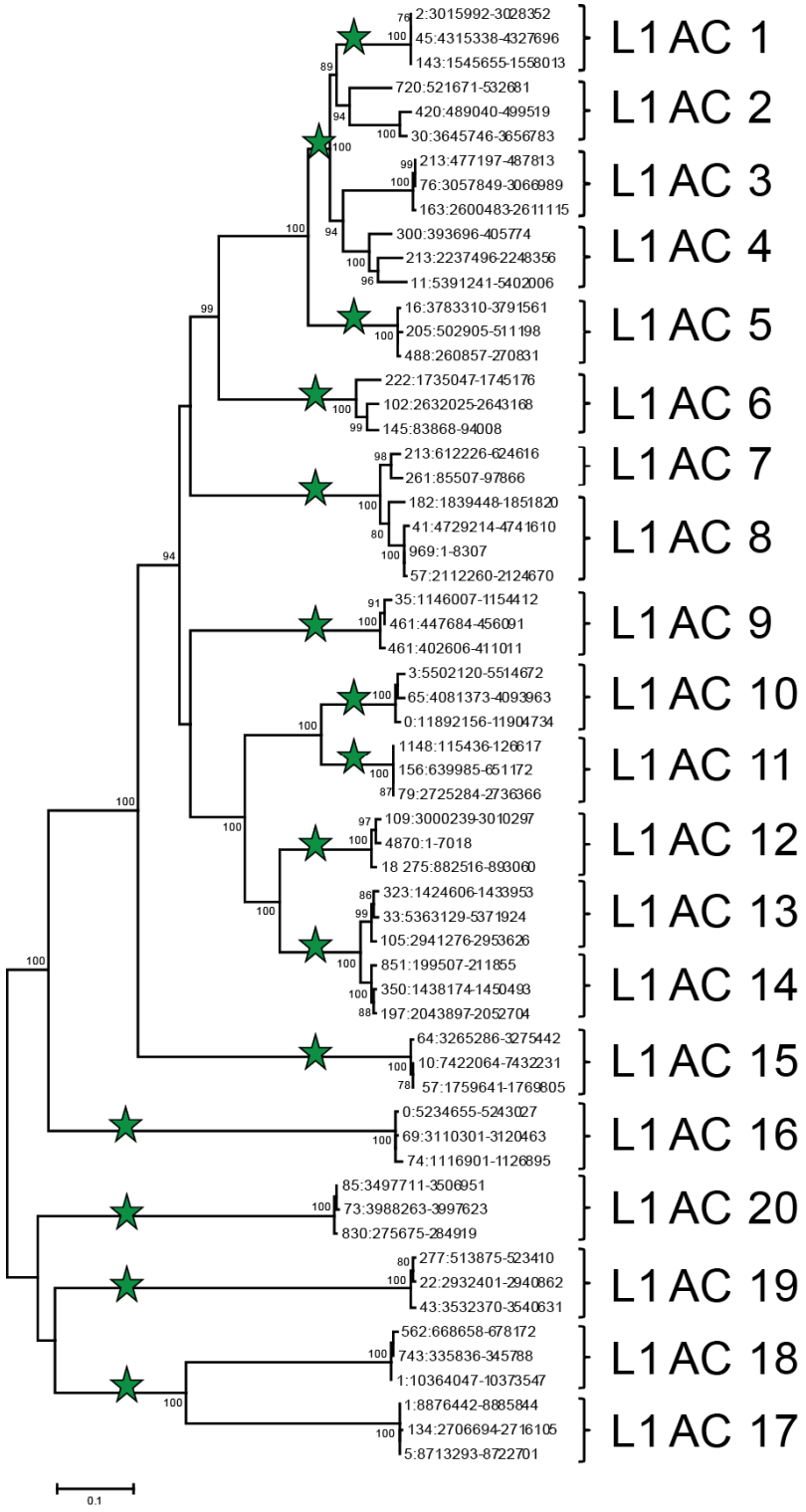

Figure 5: Phylogenetic relationships among the 20 L1 anole families recovered from the GPS output. The tree was constructed using NJ and bootstrap values >75% are shown. Green stars indicate the acquisition of novel 5' UTRs.

Table 1: Characteristics of non-LTR retrotransposons present in the Anole genome separated by family.

Clade	Family	FL Size (bp)	FL Copy Number	Truncated Copy Number	GC Element	GC Environment	Nb. AAs ORF1	Nb. AAs ORF 2	Average Pairwise Divergence (% ± S.E.)[a]	Average Divergence from Consensus (% ± S.E.)[a]
CR1	CR1 AC 1	5300	47	593	46.43	40.55	423	911	0.38 ± 0.07	0.13 ± 0.01
	CR1 AC 2	5843	25	316	46.66	40.21	423	911	0.28 ± 0.05	0.07 ± 0.03
	CR1 AC 3	5708	37	467	46.61	40.38	423	911	1.12 ± 0.12	0.51 ± 0.12
	CR1 AC 4	4653	8	101	46.52	40.33	423	911	2.89 ± 0.34	1.14 ± 0.15
L1	L1 AC 1	6378	5	5	37.19	40.11	353	1269	0.19 ± 0.20	0.10 ± 0.04
	L1 AC 2	6305	7	4	39.02	40.11	353	1335	0.57 ± 0.23	0.46 ± 0.14
	L1 AC 3	6300	11	10	38.14	40.11	353	1335	0.62 ± 0.22	0.39 ± 0.14
	L1 AC 4	6206	3	4	37.58	40.11	353	1267	0.33 ± 0.09	0.12 ± 0.07
	L1 AC 5	6190	3	5	38.08	40.11	353	1259	0.88 ± 0.33	0.00 ± 0.00
	L1 AC 6	6153	7	10	36.47	39.28	353	1250	0.35 ± 0.11	0.17 ± 0.07
	L1 AC 7	6433	4	8	35.69	40.16	354	1250	0.21 ± 0.10	0.13 ± 0.40
	L1 AC 8	6430	4	15	35.41	40.16	354	1265	0.42 ± 0.15	0.18 ± 0.08
	L1 AC 9	6410	7	98	35.80	39.20	341	1251	1.10 ± 0.30	0.50 ± 0.11
	L1 AC 10	6848	4	8	36.05	40.58	359	1246	0.33 ± 0.12	0.25 ± 0.09
	L1 AC 11	6648	6	11	36.54	40.58	366	1256	0.00 ± 0.00	0.00 ± 0.00
	L1 AC 12	6500	5	9	34.80	38.89	353	1270	0.55 ± 0.07	0.27 ± 0.03
	L1 AC 13	6758	3	8	35.29	40.58	361	1251	0.35 ± 0.11	0.17 ± 0.05
	L1 AC 14	6712	4	8	35.71	40.58	361	1251	0.53 ± 0.14	0.41 ± 0.11
	L1 AC 15	6162	11	75	37.51	39.52	349	1243	0.87 ± 0.21	0.42 ± 0.09
	L1 AC 16	6130	17	75	34.31	38.69	349	1260	1.20 ± 0.16	0.54 ± 0.08
	L1 AC 17	5410	18	85	35.86	38.91	351	1252	0.70 ± 0.07	0.35 ± 0.03
	L1 AC 18	5524	24	120	35.60	39.17	374	1240	1.36 ± 0.26	0.63 ± 0.04
	L1 AC 19	5758	4	16	36.49	37.87	351	1245	1.63 ± 0.26	0.76 ± 0.13
	L1 AC 20	5250	23	52	33.53	38.98	380	1243	0.66 ± 0.14	0.35 ± 0.06
	L1-Like AC 1	5211	10	200	41.34	40.29	483	1035	1.79 ± 0.93	0.85 ± 0.48
L2	L2 AC 1	6058	15	230	50.17	38.88	320	1001	0.41 ± 0.07	0.29 ± 0.02
	L2 AC 2	5890	30	172	50.08	38.68	318	1001	1.40 ± 0.23	0.98 ± 0.15
	L2 AC 3	4980	15	155	50.41	38.46	318	785	1.63 ± 0.21	1.27 ± 0.08
	L2 AC 4	6077	15	80	53.51	38.40	341	785	0.50 ± 0.10	0.40 ± 0.08
	L2 AC 5	5410	30	146	52.55	38.27	335	786	1.05 ± 0.11	0.98 ± 0.07
	L2 AC 6	6012	24	180	53.51	39.24	278	785	0.86 ± 0.11	0.48 ± 0.04
	L2 AC 7	5200	30	300	52.11	38.58	317	785	0.43 ± 0.10	0.31 ± 0.05
	L2 AC 8	5890	30	200	53.76	38.77	317	889	0.61 ± 0.11	0.47 ± 0.03
	L2 AC 9	5765	22	50	51.74	39.44	316	855	0.58 ± 0.15	0.31 ± 0.03
	L2 AC 10	5700	21	508	52.89	39.11	342	785	1.47 ± 0.16	1.11 ± 0.07
	L2 AC 11	6288	18	315	51.57	39.17	342	785	0.62 ± 0.09	0.40 ± 0.03
	L2 AC 12	4823	15	220	51.03	38.31	341	792	1.47 ± 0.39	1.33 ± 0.12
	L2 AC 13	5052	49	170	50.59	37.70	343	785	1.03 ± 0.17	1.00 ± 0.06
	L2 AC 14	5110	18	144	51.66	39.18	333	798	1.71 ± 0.16	1.08 ± 0.10
	L2 AC 15	5960	18	400	52.04	38.59	332	919	0.73 ± 0.07	0.43 ± 0.04
	L2 AC 16	5970	18	75	49.76	38.92	NA[b]	696	4.74 ± 0.36	2.86 ± 0.17
	L2 AC 17	5060	12	75	49.88	38.33	362	988	2.59 ± 0.26	2.02 ± 0.14
R4	R4 AC 1	3760	577	1163	42.57	40.14	NA	1137	1.33 ± 0.10	0.41 ± 0.03
	R4 AC 2	3760	417	843	42.57	40.14	NA	1138	1.51 ± 0.12	0.39 ± 0.05
RTE	RTE Bov-B AC 1	3226	61	3203	45.57	39.89	NA	931	3.90 ± 0.36	0.24 ± 0.09
	RTE-1 AC 1	3910	156	96	50.58	40.49	NA	1081	0.18 ± 0.08	0.07 ± 0.02

[a]Divergence was calculated using the Kimura 2-parameters correction in Mega 4.0

[b]Due to the low copy number and divergence of L2 AC 16 elements, the length of ORF1 was unable to be determined.

families (<1% divergence from their consensus; figure 6-A and table 1) for both TR and FL elements. The fact that 50% of FL elements have both ORFs intact (table 2) is also consistent with the young age of these elements. Therefore, it seems that the vast majority of L1 elements inserted very recently in the anole genome. The near-absence of old L1 inserts indicates that L1

elements do not reach fixation in the anole genome, although L1 elements are still active and are probably ancient residents of this genome.

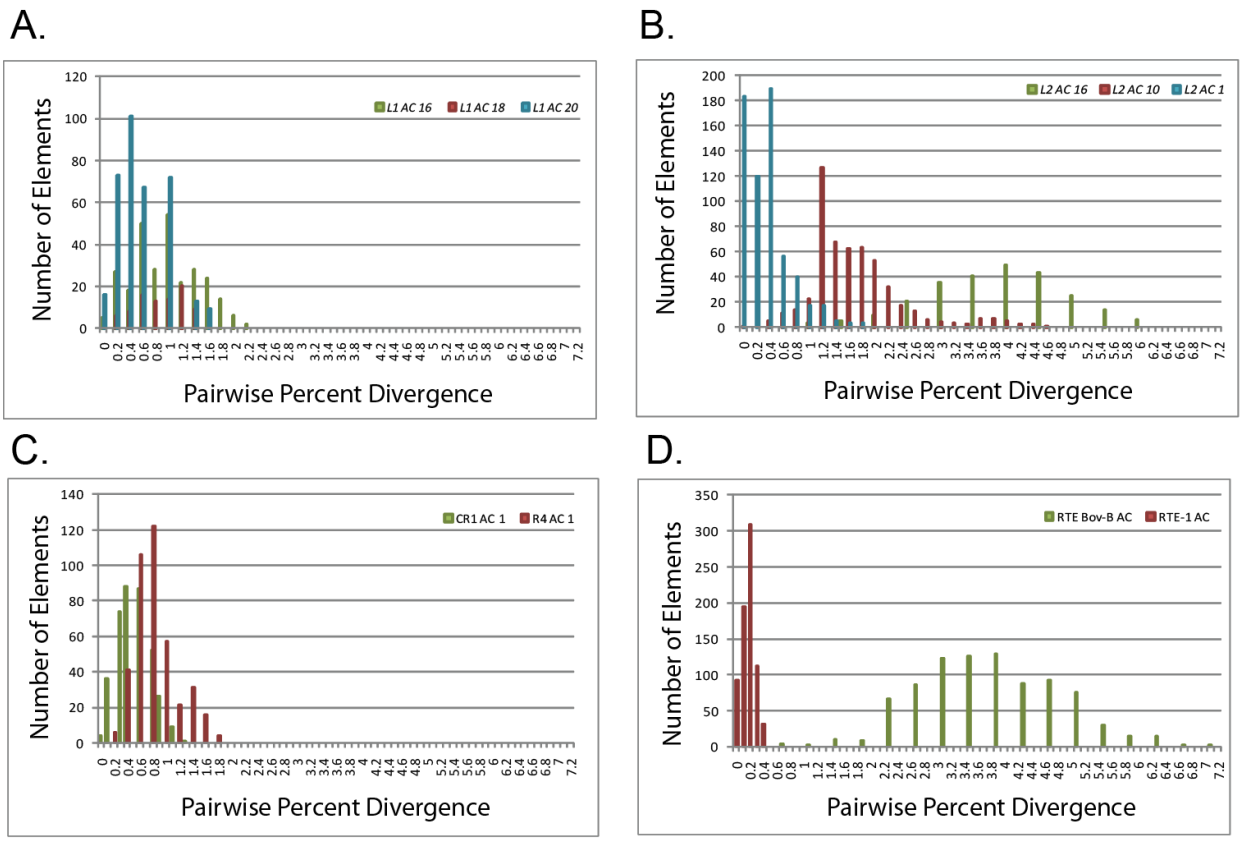

Figure 6: Pairwise divergence distribution of families belonging to all 5 clades recovered from the anole genome; A- L1 families; B- L2 families, C- CR1 and R4 families; D- RTE families.

Table 2: Characteristics of non-LTR clades in the Anole genome.

Clade	No. FL	No. Total	% ORF 1 [a]	% ORF2 [a]	% ORF 1&2 [a]
L1 AC	180	1006	65.39%	80.77%	50%
L2 AC	380	3800	20.05%	44.61%	9.52%
CR1 AC	117	1594	26.50%	39.32%	14.53%
R4 AC	994	3000	X	31.99%	X
RTE Bov-B AC [b]	61	3264	X	21.34%	X
RTE-1 AC [b]	156	252	X	35.35%	X

[a] Percent of ORFs intact was calculated from orginal GPS output

[b] RTE Bov-B AC and RTE-1 AC were treated as separate clades due to their divergence and differences in evolutionary dynamics

The L2 clade is also very diverse and is represented by 17 monophyletic families (figure 7), ranging in copy number from 71 to 529. L2 families are divergent from each other although not as much as L1 families, the deepest node in the L2 clade corresponding to approximately one fifth of the deepest divergence in the L1 clade (figure 4). The majority of L2 families show evidence of recent activity and are represented by young elements. For instance, the L2AC1 family contains only elements that diverge from each other by 0.41% on average, suggesting that this family might still be active (figure 6-B). We did not detect any L2AC1 element diverging from the family consensus by more than 1%, suggesting that, like L1 elements, L2AC1 elements are not accumulating in the anole genome. In fact, out of 17 L2 families, 15 families have an average divergence from consensus lower than 1%. This is consistent with the fact that 44.6% of all FL L2 elements have an intact ORF2 (table 2). Only two families have an average divergence higher than 2% (L2AC16 and L2AC17). Family L2AC16 has an average divergence of 4.74% and the lowest divergence between two L2AC16 elements is 2.76% (table 1). This indicates that this family has not been recently active and contains elements that are likely to be fixed. Interestingly, the two L2 families that show evidence of accumulation are also the least numerous ones in our collection. The dynamics of amplification of these L2 families is reflected in the distribution of pairwise divergence between elements (figure 6-B): at some point in time these families began accumulating in small numbers in the anole genome and their accumulation coincided or was soon followed by their decline and likely extinction. It is interesting to note that the only two L2 families that show evidence of accumulation are also the only ones that seem to be no longer active. However, these two families are exceptions and the vast majority of L2 elements are very young.

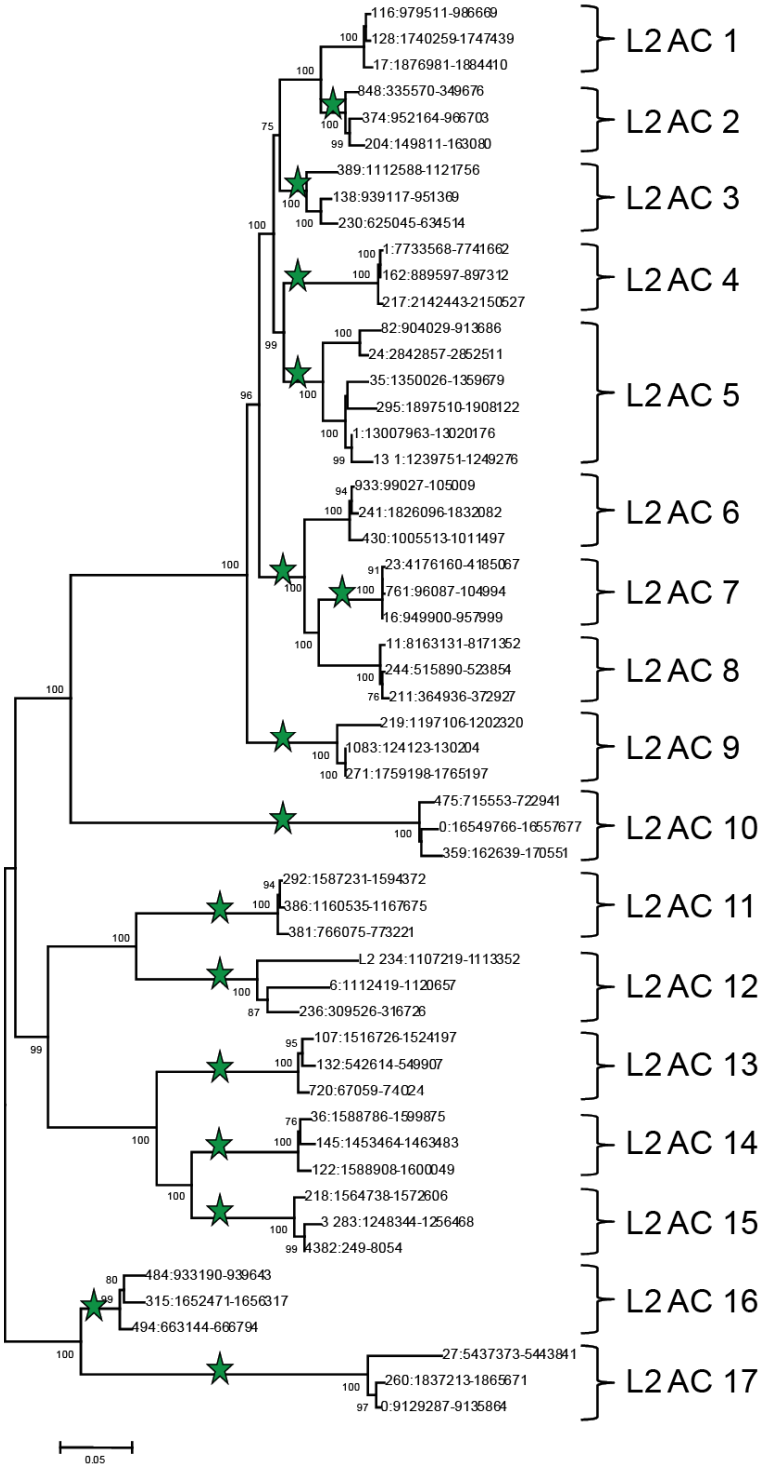

Figure 7: Phylogenetic relationships among the 17 L2 anole families recovered from the GPS output. The tree was constructed using NJ and only bootstrap values >75% are shown. Green stars indicate the acquisition of novel 5' UTRs.

The CR1 and R4 clades show very little family structure compared to L1 and L2, yet they are found in high copy number, reaching ~1,600 and ~3,000 copies, respectively. Only four CR1 families are supported by the phylogenetic analyses (figure 8-A). The R4 clade shows a little more sequence diversity than CR1 but more than half of all R4 elements belongs to two major families (R4AC1 and R4AC2 on figure 8-B). CR1 and R4 families are not as divergent as L1 and L2 families. For instance, the deepest node on the CR1 tree corresponds to 4.3% divergence. Most CR1 and R4 inserts are recent as suggested by the low divergence between elements within each family (<1% from consensus) and by the high proportion of FL elements with intact ORF (table 2). The CR1 copy number reported here contrasts with previous estimates of ~300,000 CR1-derived sequences in the anole genome (Shedlock et al. 2007). Indeed, a screening of the anole genome using GPS revealed a large number of small DNA segments derived from CR1 sequences (table 2). However, most of these elements are highly fractioned and degenerate suggesting they are very ancient.

The RTE clade contains two very divergent lineages (RTE Bov-B AC and RTE-1 AC) that separated before the origin of vertebrates (Zupunski, Gubensek, and Kordis 2001). The Bov-B family is one of the most successful non-LTR retrotransposons in the anole, with 3325 copies, while the RTE-1 family is just under 250 copies. The phylogeny of Bov-B and RTE-1 elements did not reveal any clear subsets (data not shown) like we observed for other clades, suggesting that a small number of closely related progenitors is responsible for the amplification of RTE elements. RTE-1 elements are extremely similar to each other, with an average divergence of 0.21%, suggestive of their young age. In contrast, the average divergence of the Bov-B family is 3.9%. The pairwise divergence distribution of the Bov-B family (figure 6-D) shows that this family has not produced recent insertions and is probably no longer active, although 21% of FL Bov-B elements have an intact ORF. The large divergence of this family suggests that the vast majority of Bov-B inserts is likely to be fixed. The pairwise divergence

distribution of Bov-B (Figure 6-D) is similar to the one of L2AC16 and L2AC17 and provides another example of an extinct family that accumulated in the anole genome.

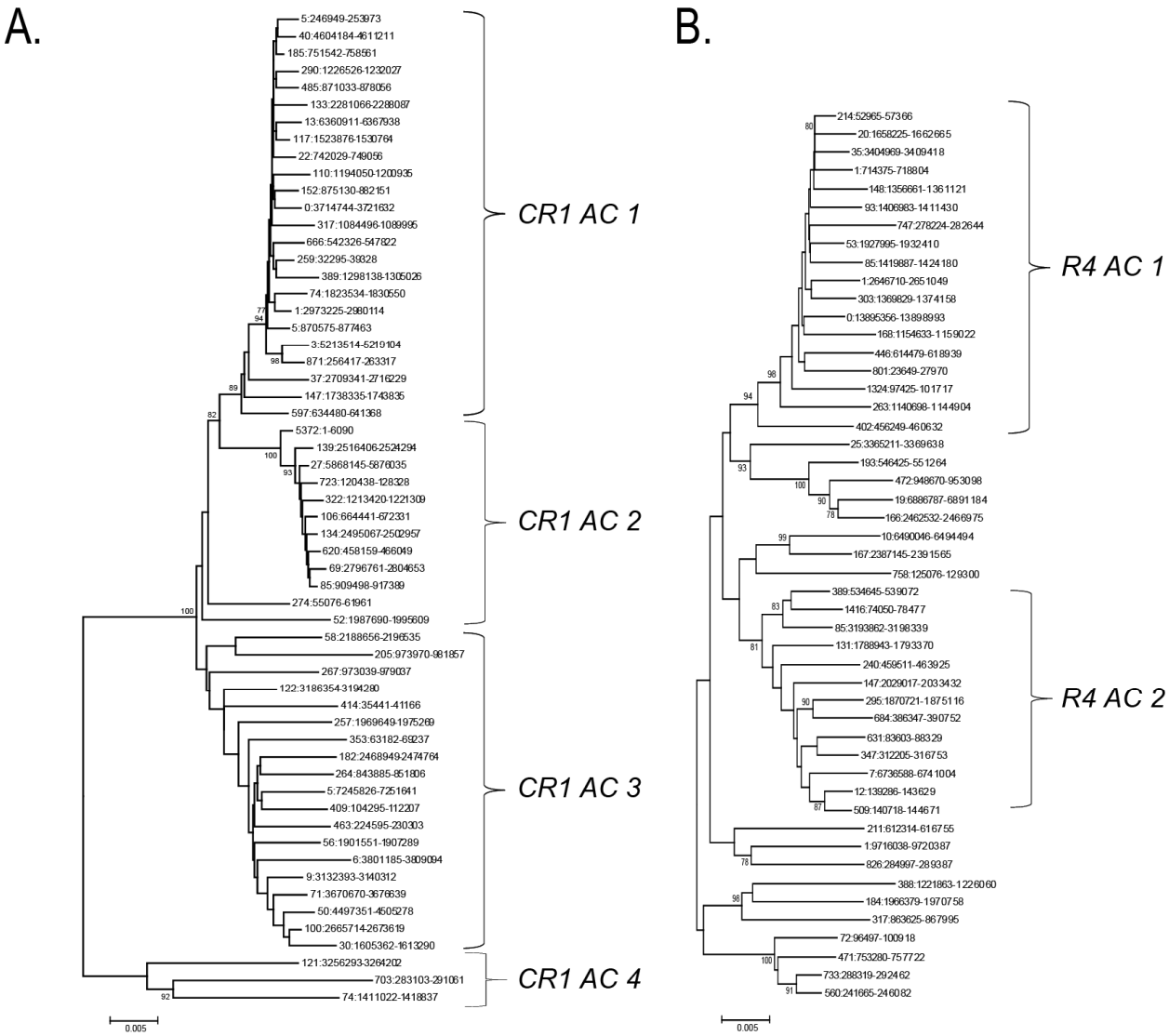

Figure 8: Phylogenetic relationships among CR1 (A) and R4 (B) elements recovered from the GPS output. The tree was constructed using NJ and bootstrap values >75% are shown.

Because the Bov-B family is relatively ancient and numerous, it offers the opportunity to examine the decay of elements in the anole. Using the first 150bp of a Bov-B element as a probe, we performed a BLAT search of the genome. Each hit was retrieved together with 7Kb of

downstream DNA and aligned to the anole Bov-B consensus sequence. As elements with an intact 5' end are presumed to be FL upon insertion, we should be able to find a 3'UTR about 3Kb downstream of each 5' terminus. In fact, we did not find a 3' end downstream of most 5' sequences, as only 28% of elements extended all the way from the 5'UTR to the 3' end (Figure 9). The same analysis performed on the much younger RTE-1 family revealed that 65% of 5'

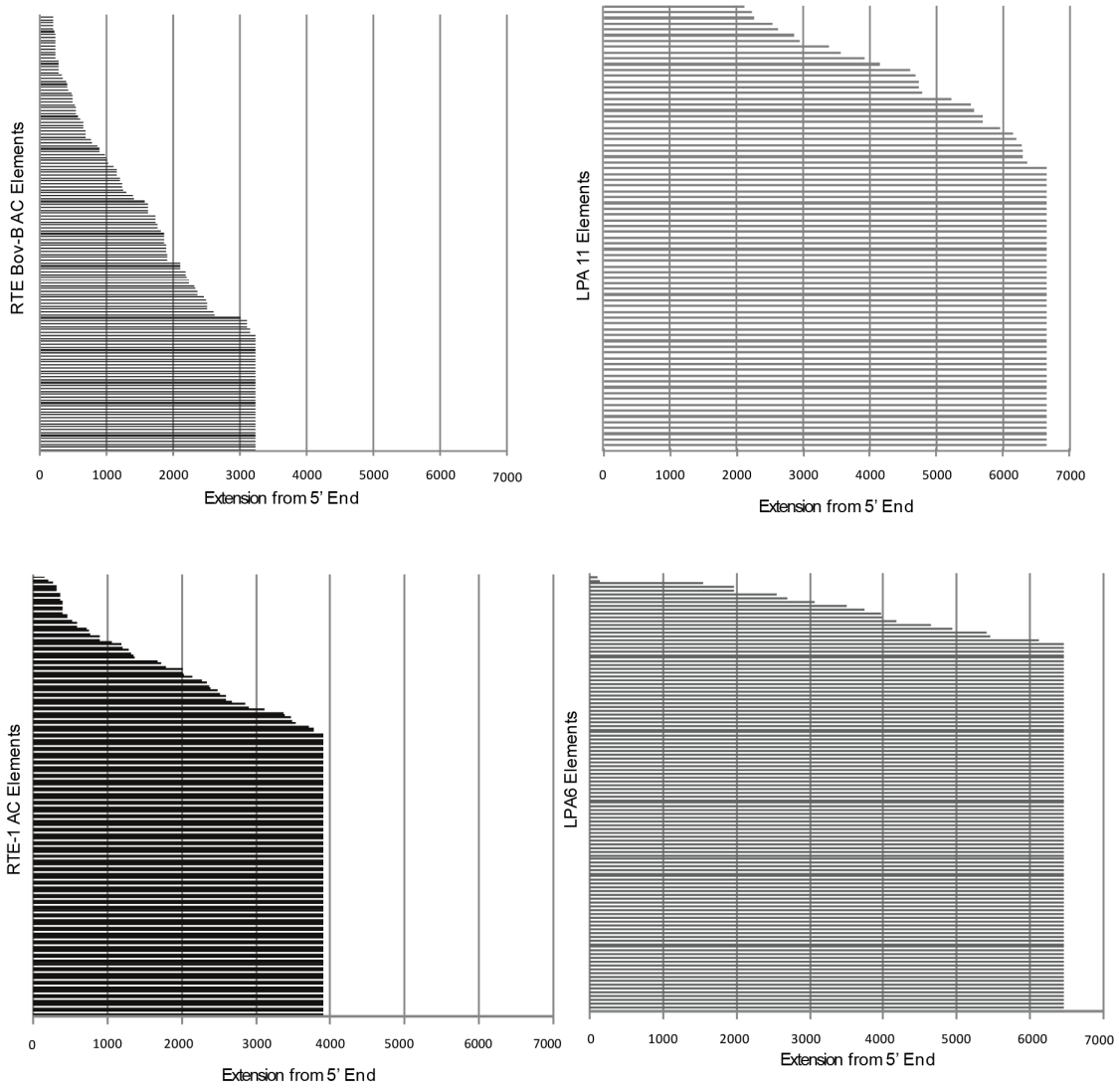

Figure 9: Length of elements as they extend from the 5' end. At least 80 sequences from each of the 4 families, RTE Bov-B AC (A), L1PA11 (B), RTE-1 AC, (C) and L1PA6 (D) were analyzed.

terminal sequences extend all the way to the 3' end. For comparison, we performed the same analysis in human using the 5' end of a L1PA6 and of a L1PA11 element. The average divergence of the LPA6 family is the same as Bov-B AC (~4%) whereas L1PA11 family is much older and is about 16% divergent (Khan, Smit, and Boissinot 2006). Here we found the 3' end corresponding to the 5' end for 86% and 78% of L1PA6 and L1PA11 elements respectively, although L1 elements are about twice as long as RTE elements. This indicates that elements decay much faster in anoles than in humans. FL elements that do not extend to their 3' extremity not only contain small deletions, they completely miss their 3' end. This suggests that the decay of FL elements is probably due to large deletions, possibly mediated by ectopic recombination between elements, instead of DNA loss by small deletions. The rapid decay of elements explains, in part, why we identified very few old (i.e. elements that diverge by more than 5% from their consensus) FL elements.

All the elements we collected are typical members of their clade. However, comparisons of consensus sequences revealed a striking difference between FL elements in the L1 and L2 clades. Although the coding region of these families is quite conserved within each clade and among families within clades, the 5'UTR is not. In fact, we identified 15 and 16 different 5'UTRs in the L1 and L2 clades, respectively. Although the first 20-40bps are shared across L1 families, these 5'UTR sequences show very little homology to each other. For instance, figure 10 shows a dot plot comparison of two L1 and L2 consensus sequences. These consensus sequences differ by less than 10% outside the 5'UTR and can be easily aligned, yet these sequences have completely different, that is non-homologous, 5'UTRs. Interestingly, the family diversity of each clade correlates nicely with the diversity in 5'UTR sequences, each major L1 and L2 lineages have a different 5'UTR (figures 5 and 7). In contrast, clades with low family diversity, like RTE, CR1 and R4, do not show any diversity at their 5' end. This correlation suggests that the ability

to recruit novel promoter sequences in L1 and L2 drives the evolution of simultaneously active families and might be responsible for the diversity of these clades.

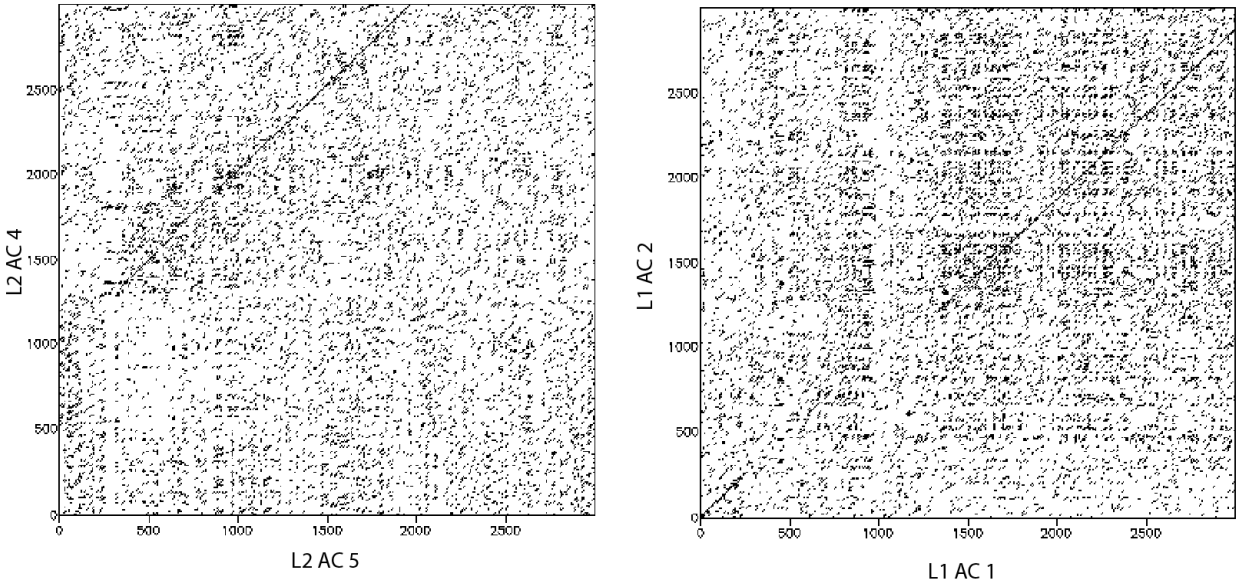

Figure 10: Dotmatcher results comparing the 5' UTR of two closely related L2 (Left) and L1 (Right) consensus sequences.

Discussion

The anole genome is the first non-avian reptilian genome whose complete sequence is available. It bridges a large phylogenetic gap and provides a unique opportunity to comparatively investigate the evolution of amniote genomes. We found that the anole genome contains an extraordinary diversity of non-LTR retrotransposons. Five clades show signs of very recent activity and two of these clades, L1 and L2, contain numerous families, some of which have been simultaneously active since the separation between mammals and reptiles (315 MYA). We identified at least 46 recently active families of non-LTR retrotransposons in the anole and we estimate the number of potentially active elements (i.e. FL elements with both ORF intact) to be about ~500 copies. This situation is reminiscent of the diversity of non-LTR retrotransposons in teleostean fish. Fish genomes usually contain several active clades of non-

LTR retrotransposons, sometimes represented by a large diversity of families (Volff et al. 2003; Duvernell, Pryor, and Adams 2004; Furano, Duvernell, and Boissinot 2004; Neafsey 2004; Basta, Buzak, and McClure 2007). For instance the pufferfish (*Tetraodon nigroviridis*) and the zebrafish (*Danio rerio*) genomes contain 6 and 7 active clades of non-LTR retrotransposons, respectively (Volff et al. 2003; Basta, Buzak, and McClure 2007). In zebrafish the L1 clade is represented by 32 active families that diverged before the origin of vertebrates (Furano, Duvernell, and Boissinot 2004). This situation contrasts greatly with mammalian genomes that are dominated by a single clade, L1 in eutherians and marsupials (Furano, Duvernell, and Boissinot 2004; Gentles et al. 2007) and L2 in monotremes (Gilbert and Labuda 2000; Warren et al. 2008). L1 usually evolves in mammals as a single lineage, so that only a unique family of closely related elements is active at a given time (Furano 2000; Furano, Duvernell, and Boissinot 2004). For instance, a single family, called Ta, is currently active in human (Skowronski, Fanning, and Singer 1988). This family contains only two subsets of active elements, Ta-0 and Ta-1, which are considerably less divergent (<1%) from each other than anole's L1 families are (Boissinot, Chevret, and Furano 2000).

The much greater diversity of non-LTR retrotransposons in the anole and in teleostean fish does not translate into larger genome sizes. In fact, mammalian genomes are significantly larger than those genomes. The human and mouse genomes are 3.2 and 2.8Gb, respectively; in comparison, the *Anolis* genome is only 2.2Gb while teleostean genomes vary between 0.4Gb (in *Tetraodon nigroviridis*) and 1.7Gb (in *Danio rerio*) (Volff et al. 2003). The larger size of mammalian genomes relative to other vertebrates is directly related to the abundance of L1 elements in mammals. It was estimated that the human genome contains ~500,000 copies of L1 elements that account for ~17% of our DNA (Lander et al. 2001). In contrast, we identified only ~16,000 non-LTR retrotransposons (FL and TR), accounting for ~1.3% of the anole genome. In addition to the recent elements we characterized, the anole genome contains a large number of

small DNA fragments derived principally from ancient CR1 activity (Table 2) (Shedlock et al. 2007). When these old elements are taken into account, the total fraction of the anole genome derived from non-LTR retrotransposons is still largely inferior (~10% of the genome) to that of mammals. This is consistent with the fact that the total number of reverse transcriptase hits detected by GPS in anole is about 50% of that in human (table 2; McClure et al. 2005). Therefore, the accumulation of non-LTR retrotransposons found in mammals is truly specific of this class of vertebrates and suggests a major difference between mammalian and non-mammalian vertebrates in the way host genomes interact with their parasitic non-LTR retrotransposons.

The scarcity of divergent elements and the abundance of very young inserts indicate that the vast majority of non-LTR retrotransposons do not reach fixation in the anole genome. This suggests a rapid turnover of elements, in which the insertion of new elements is offset by the loss of element-containing loci. This mode of evolution is similar to the turn-over of elements in *Drosophila*, where selection against element-containing loci limits copy number (Charlesworth and Langley 1989; Eickbush and Furano 2002). Similarly, the turn-over of elements in the anole suggests that selection against retrotransposon inserts must be strong enough to prevent their accumulation. L1 inserts, in particular long ones, are also negatively selected in mammals (Boissinot, Entezam, and Furano 2001; Boissinot et al. 2006; Song and Boissinot 2007), yet they do accumulate in mammalian genomes. Therefore, the cost that non-LTR elements impose on the fitness of their host is likely to be higher in the anole than it is in mammals. The deleterious effect of retrotransposons can result from their ability to mediate ectopic recombination (Langley et al. 1988), effects on gene activity (Charlesworth and Charlesworth 1983) or the retrotransposition mechanism *per se* (Boissinot, Entezam, and Furano 2001). Because the five clades are all characterized by a high rate of turn-over, non-LTR

retrotransposons must have a common deleterious effect that is independent of their clade and that affect equally FL and truncated elements.

A deleterious effect due to the retrotransposition process is unlikely because selection should prevent the fixation of retrotransposition-competent FL elements but not of TR elements because they are incapable of further retrotransposition. Gene inactivation is also an unlikely mechanism because the anole genome is sufficiently large and contains regions with low gene density where non-LTR retrotransposons could accumulate. Chromosomal rearrangements and large DNA deletions caused by ectopic (i.e. non-allelic) recombination predicts that selection should act primarily against long elements because they are more likely to be involved in ectopic exchange than shorter ones. Experimental and genomic evidence in mammals suggest that ectopic recombination occurs rarely if the length of sequence homology is shorter than 1.2Kb (Cooper, Schimenti, and Schimenti 1998; Song and Boissinot 2007). As short elements are also rare in the anole genome, it is doubtful that this mechanism plays a major role, unless ectopic recombination in reptiles requires a much shorter length of homology than in mammals or unless the rate of recombination is much higher in anole than in mammals (see below). However, this model fails to explain why many elements belonging to some specific families (L2AC16, L2AC17 and Bov-B) have reached fixation in the anole genome, as there is no reason to believe that these elements were less likely to mediate ectopic recombination. Recent experimental works suggest a possible extension of the ectopic exchange model. In yeast, a defense mechanism limits inter-element recombination by changing the conformation of the chromatin at the insertion site and in neighboring sequences (Ben-Aroya et al. 2004). If chromatin modification has a negative impact on the function of the genome, all elements, whatever their length or their clade, would be deleterious and therefore eliminated by selection. If some families lack sequence motifs recognized by the surveillance machinery of the host, they could temporarily evade host defense and accumulate. This model is still speculative

because a defense mechanism, designed to prevent inter-element recombination, has yet to be found in vertebrates.

Our analysis of the Bov-B family emphasizes another mechanism that could account for both the smaller size of the anole genome and for the scarcity of old elements. We found that elements in the anole decay much faster than their mammalian counterpart and that their decay results from the loss of their ends presumably caused by inter-element recombination. This pattern is very similar to the decay of CR1 elements in chicken (Abrusan et al. 2008) where the low abundance of elements could be the consequence of their high recombination rate. Although very little is known about the recombination rate in squamate reptiles, our data suggests that the rate of ectopic recombination might be higher in the anole (and birds) than in mammals. It has been previously suggested that one of the conditions that makes mammalian genomes permissive to the amplification of L1 is a low rate of ectopic recombination (Eickbush and Furano 2002). The decay of L1 in mammals results from the accumulation of neutral mutation and small indels and not from the loss of element ends, as observed in anole and chicken. This difference in the decay of copies underlies a fundamental difference in the frequency of ectopic exchange between mammals and non-mammalian vertebrates and provides an explanation for the contrasted diversity and abundance of non-LTR retrotransposons among vertebrates.

CHAPTER II: DNA TRANSPOSONS

THE EVOLUTION AND DIVERSITY OF DNA TRANSPOSONS IN THE GENOME OF THE LIZARD *Anolis carolinensis*

Introduction

Transposable elements (TEs) are mobile DNA sequences that are typically classified into two categories: the class I elements that require an RNA intermediate for their transposition (*e.g*. retrotransposons) and the class II elements that transpose as single or double strand DNA (Craig et al. 2002). Class II transposons can be divided into three categories that differ in their mode of transposition: the classical cut-and-paste DNA transposons, the rolling circle transposons or *Helitrons*, and the *Mavericks* or *Polintons* (Feschotte and Pritham 2007). The cut-and-paste group is the most diverse and is subdivided into 10 superfamilies that diverged before the diversification of eukaryotes. Cut-and-paste transposons contain an open-reading frame (ORF) that encode for the enzyme transposase. The transposase specifically recognizes the terminal inverted repeats (TIRs) of the element, excises the transposon and inserts it elsewhere in the host's genome (Robertson 2002). Upon insertion in the genome, target site duplications (TSDs) are produced. The length and sequence of the TSDs and terminal motifs of the TIRs are highly conserved across superfamilies and are useful in categorizing elements. Following the excision of an element, the donor site may be repaired via homologous recombination. However, the gap repair process is oftentimes interrupted resulting in shorter elements with internal deletions (Engels et al. 1990). These shorter copies still possess TIRs that can be recognized by the transposase encoded by complete elements and consequently they retained their mobility (Hartl, Lozovskaya, and Lawrence 1992). These non-autonomous elements compete with their progenitors for the transposase and often outnumber their autonomous relatives.

TEs have dramatically affected the size, structure and function of the genomes they inhabit (Lander et al. 2001). Although most TE insertions are either neutral or deleterious, the domestication by the host of TE-encoded sequences can occur and is responsible for the

evolution of fundamental biological processes such as light-sensing in plants and V(D)J recombination in vertebrates (Jiao, Lau, and Deng 2007; Jones and Gellert 2004; Matthews 2006). However, it is likely that the impact of class I and class II elements varies among species because their abundance and diversity greatly differ among groups of organisms (Eickbush and Furano 2002; Furano, Duvernell, and Boissinot 2006; Pritham, Feschotte, and Wessler 2005). For instance, fish genomes contain a diversity of active DNA transposons that coexist with a multitude of retrotransposon families (Duvernell, Pryor, and Adams 2004). In contrast, mammalian genomes are dominated by class I elements and it was believed until recently that mammals completely lack active class II elements, although DNA transposons were once diverse and very active in early mammalian evolution (Lander et al. 2001; Lindblad-Toh et al. 2005). However, recent analyses have shown that vertebrate genomes, including mammalian genomes, can be re-colonized by laterally transferred DNA transposons and that these transfers seem to occur relatively frequently (Pace et al. 2008; Novick et al. 2009b).

The evolution of DNA transposons and their impact on vertebrate genomes is incompletely understood in part because the most studied vertebrate genomes are mammalian genomes that lack active class II elements (Lander et al. 2001; Lindblad-Toh et al. 2005). Here we present the first analysis of class II elements in the North American green anole, *Anolis carolinensis.* The green anole is the first non-avian reptile to have its genome sequenced, bridging a large phylogenetic gap between fish and mammals. We discovered that DNA transposons are represented in the anole genome by 10 autonomous families belonging to four superfamilies (*hAT*, *Mariner*, *Chapaev*, and *Helitron*). These autonomous elements are responsible for the amplification of a multitude of non-autonomous families that largely outnumber their autonomous counterparts. The age distribution of DNA transposons suggests that novel insertions can readily reach fixation, yet the near absence of ancient elements

indicates that some post-insertional mechanism(s) limits the accumulation of DNA transposons in the anole genome.

Materials and Methods

Acquisition of class II elements in the anole genome

An exhaustive search of the anole genome for class II transposons was completed with three different methods. Our initial analysis was accomplished using the program PILER (Edgar and Myers 2005). We used this program to find matching sequences of a minimum repeat length of 100bp and a minimum consensus of 95%. From the output subgroups, only those with a minimum of ten copies were analyzed. These novel families were then extracted and assembled into contigs using Seqman II (part of the DNAStar package), combining any redundant output into a single family. The resulting alignments were collected to form an initial library of TEs from the anole genome. This library was then used as the basis of a RepeatMasker search of the genome to find additional copies of the putative elements. Hits of sufficient length, typically at least 100bp, were extracted from the genome along with a minimum of 500bp of flanking sequence using custom PERL scripts. The extracted sequence subgroups were again aligned using MUSCLE and consensus sequences were generated. The process was repeated until the full length sequence of each putative element was obtained.

A second search of the genome, using the Repeatscout program (Price et al. 2005), was performed to identify any previously undiscovered elements (lmer = 12, thresh = 100). The resulting putative TEs were assembled into contigs using Seqman II and aligned using MUSCLE. Any previously identified putative elements were not processed, while new elements were used to create a library for use in a Repeatmasker search and the output as described above.

Lastly, we performed a BLASTX search of the genome using amino acid sequences derived from a known transposon library available from Repbase (v13.01). The resulting hits of at least 100bp, with a maximum score of 2e-150, at multiple loci were extracted along with 500bp of flanking sequence (using a modified version of the previously used PERL scripts) and aligned.

Classification of elements

Elements were separated into superfamilies and further subdivided into families based upon size and sequence similarity. A consensus sequence for each family was created. The pairwise divergence between elements and the average divergence from the consensus sequence were calculated using Kimura's 2-parameter method in Mega 4.0 (Tamura et al. 2007). As these elements are not annotated, we estimated copy number for each family by using the BLAST option on NCBI. In order to only count non-fragmented elements, we limited our percent identity score to 90%.

Finally, elements were scanned for the presence/absence of TIRs, TSDs, ORFs and similarity to elements in other genomes. TIRs were discovered by aligning the 5' in the positive orientation with the reverse complement of the 3' of elements belonging to the same family. TSDs were determined by collecting 20 base pairs downstream of the 3' end and 20 bases upstream of the 5' end for at least 20 sequences per family. Percentages for each of the four possible nucleotides were then calculated for each position. As repeat masker oftentimes did not recognize some of these novel non-autonomous elements, the sequence and length of TSDs and TIRs were used to categorize each family to its proper autonomous superfamily. ORF finder and Conserved Domains Database (Marchler-Bauer et al. 2007) were used in tandem to find the length of ORFs and types of proteins encoded by autonomous elements. Finally, in order to identify possible events of horizontal transfer, consensus sequences from

each family were submitted to the BLAST option on the NCBI website and a multitude of sequenced organisms were screened.

Results

The genome of *A. carolinensis* contains 68 families (10 autonomous and 58 non-autonomous) of class II transposons, representing four of the previously recognized superfamilies: *hAT*, *Mariner*, *Helitron* and *Chapaev*. These superfamilies differ drastically in abundance and diversity in anole.

The hAT superfamily

The *hAT* superfamily is the most abundant and diverse in the anole genome. It is represented by five autonomous and 32 non-autonomous families (table 3). All these families display the typical structural features typical of the *hAT* superfamily, including 8bp TSDs and terminal motifs of sequence YARNG. Four of the five autonomous *hAT* families are found in distantly related animals and result from independent events of horizontal transfer (*hAT-HT1_AC HT*, *hAT-HT2_AC*, *hAT-HT3_AC* and *SPIN_AC*) (Pace et al. 2008; Novick et al. 2009b). The *hAT-4_AC* family is the only one for which there is no evidence of lateral transfer because we failed to find elements similar to *hAT-4_AC* in other genomes. The four laterally transferred *hAT* elements are more closely related to mammalian *Charlie* elements whereas *hAT-4_AC* belongs to the *hobo* family of *hAT*. Each of these autonomous families produced non-autonomous (from 1 to 15) families with similar TSD and TIR sequences. As non-autonomous families are typically more numerous than autonomous families, non-autonomous copies outnumber autonomous copies in the anole genome by two orders of magnitude (~330 autonomous copies vs ~24,000 non-autonomous copies). For instance, the most abundant autonomous family, *hAT-4_AC* (290 copies), is responsible for the mobility of at least 15 non-

Table 3: Characteristics and nomenclature of all families of autonomous and non-autonomous hAT DNA transposons in the lizard *Anolis carolinensis*

Name	Copy Number >90% Identity	Length in bp	TSD	Length of TIR	TIR	% Divergence ± S.E.	% Divergence from consensus ± S.E.
hAT-1_AC	5 (5)[a]	2968	8bp[c]	16	CAGTGATGGSSAACCT	5.82 ± 0.40	3.88 ± 0.35
hAT-1N1_AC	868	585	NTCTAGAN	16	CARTGATGGSCAACCT	4.75 ± 0.64	2.55 ± 0.42
hAT-2_AC	5 (4)[a]	2246	8bp[c]	15	CAGGGGTCCCCAAAC	0.14 ± 0.07	0.00 ± 0.00
hAT-2N1_AC	28	1485	NHCTAGRN	16	CAGGGGTCCCCAAACT	1.34 ± 0.31	0.73 ± 0.15
hAT-2N2_AC	43	1065	NTCTAGAN	16	CAGGGGTCCYCAAACT	3.09 ± 0.27	1.52 ± 0.22
hAT-2N3_AC	1372	780	NTNTANAN	16	CAGGGGTCCYCAAACT	4.76 ± 0.43	1.75 ± 0.18
hAT-3_AC	<25 (13)[a]	2754	8bp[c]	14	CAGTGRTTCCCAAA	5.62 ± 0.29	3.22 ± 0.22
hAT-3N1_AC	344	326	NTCTAGAN	14	CAGTGRTTCCCAAA	3.98 ± 0.37	1.78 ± 0.23
hAT-3N2_AC	132	833	NYYTARRN	16	CAGGGGTCCCCAAACT	8.44 ± 0.77	4.04 ± 0.35
hAT-3N3_AC	682	799	NTCTAGAN	14	CAGTGRTTCCCAAA	9.66 ± 0.54	6.31 ± 0.42
hAT-4_AC	25 (290)[a]	≈15,000	8bp[c]	11	TAGGCTTGMTC	1.93 ± 0.09	1.12 ± 0.07
hAT-4N1_AC	300	1854	NTRNNYAN	11	TAGGCTTSATC	1.80 ± 0.30	0.90 ± 0.15
hAT-4N2_AC	201	2359	NTRNNYAN	11	TAGGCTTSATC	1.71 ± 0.40	1.08 ± 0.30
hAT-4N3_AC	489	2192	NTRNNYAN	11	TAGGCTTSATC	0.90 ± 0.01	0.50 ± 0.01
hAT-4N4_AC	60	2042	NTRNNYAN	11	TAGGCTTSATC	1.07 ± 0.09	0.49 ± 0.06
hAT-4N5_AC	34	2440	NTRNNYAN	11	TAGGCTTSATC	0.88 ± 0.09	0.27 ± 0.04
hAT-4N6_AC	188	1995	NTRNNYAN	11	TAGGCTTSATC	0.82 ± 0.11	0.28 ± 0.14
hAT-4N7_AC	295	2766	NTRNNYAN	11	TAGGCTTGAGC	1.64 ± 0.13	0.46 ± 0.04
hAT-4N8_AC	264	2262	NTRNNYAN	11	TAGGCTTGAGC	2.28 ± 0.21	0.73 ± 0.08
hAT-4N9_AC	284	2465	NTRNNYAN	11	TAGGCTTGAGC	1.90 ± 0.21	1.04 ± 0.13
hAT-4N10_AC	71	2646	NTRNNYAN	11	TAGGCTTGAGC	2.43 ± 0.18	0.69 ± 0.11
hAT-4N11_AC	120	2686	NTRNNYAN	11	TAGGCTTGAGC	1.92 ± 0.13	0.68 ± 0.19
hAT-4N12_AC	26	3413	NTRNNYAN	11	TAGGCTTGAGC	1.81 ± 0.13	0.91 ± 0.08
hAT-4N13_AC[b]	16	>3.5kB	NA	11	TAGGCTTGAGC	2.23 ± 0.23	1.15 ± 0.15
hAT-4N14_AC	235	1911	NTANNTAN	11	TAGGCTTGMKC	1.06 ± 0.14	0.56 ± 0.08
hAT-4N15_AC	689	2490	NTANNTAN	11	TAGGCTTGAKC	0.56 ± 0.10	0.23 ± 0.04
SPIN_AC[c]	<5 (1)[a]	NA	-	-	CAGYGGTTCTCAACCT	NA	NA
SPIN_NA11_AC[c]	12138	273	-	-	-	NA	NA
SPIN_NA1_AC	181	745	NYYTARRN	16	CAGTGKTTCTCAACCT	12.3 ± 0.69	7.04 ± 0.74
hAT-N1_AC	>1000	135	NYYTARRN	57	CAGTGGTTCTCAACCTGTGG[e]	4.51 ± 0.66	2.36 ± 0.37
hAT-N2_AC	<20	188	NYYTARRN	15	CAGSYTTYTYMAMCM	49.9 ± 4.49	18.5 ± 2.26
hAT-N3_AC	163	600	NYYYRRRN	16	CAGTGGTTCCCAACCT	10.8 ± 0.77	6.00 ± 0.51
hAT-N4_AC	>1000	732	NYYTARRN	16	CAGGGGTCCYCAAACT	3.01 ± 0.24	1.50 ± 0.12
hAT-N5_AC	16	495	NTNTANAN	16	CAGGSRTGTCCAACCT	33.7 ± 2.97	15.2 ± 1.93
hAT-N6_AC	964	810	NTTRYAAN	12	CAGAGSCGGYCC	0.60 ± 0.09	0.30 ± 0.04
hAT-N7_AC	>1000	288	NNTNNANN	12	CAGAGCCGGYCC	2.39 ± 0.36	1.68 ± 0.51
hAT-N8_AC	>1000	329	NNTNNANN	12	CAGAGSCGGYCC	1.77 ± 0.35	1.25 ± 0.27

[a]Number of ORFs found in the genome regardless of the 5' and 3' ends

[b]Few elements from this family were collected lacking 5' ends deeming it impossible to calcluate length or TSDs. TIRs were hypothesized from the 3' end.

[c]Not enough elements were retreived to construct a TSD pattern beyond the number of bps.

[d]Data from Pace II et. al. 2008

[e]Only the first 20bp of the TIR are presented here.

autonomous families, totaling 3272 copies. Most non-autonomous families correspond to deleted versions of autonomous elements; consequently determining their evolutionary affinities is relatively easy. Yet, we identified 8 non-autonomous families that did not show any similarity with a known autonomous family beyond the TIR. For instance, *hAT-N4_AC* elements have TIRs that are indistinguishable in length and sequence from *hAT-2_AC*, yet they do not share homology with autonomous and non-autonomous *hAT-2* outside the TIRs (table 3). Although

the origin of this family remains unclear, the similarity of its TIRs with the TIRs of autonomous *hAT-2* elements suggests that *hAT-N4_AC* elements are mobilized by *hAT-2*. Out of the 8 "orphan" families, we found similarity with the TIR of an autonomous family for 5 of them suggesting that a known autonomous family is responsible for their mobilization. The remaining three (*hAT-N6*, *hAT-N7* and *hAT-N8*) have identical TIR but the sequence of their TIR is different from the TIR of the 5 autonomous families. The transposase of one of the autonomous copies could be mobilizing these elements despite their lack of similarity in the TIR. Alternatively, we cannot exclude that another autonomous *hAT* family exists in anole and either was missed by our search if its copy number is extremely low or is so new in anole that it is polymorphic in populations and absent from the individual used for the genome sequencing.

Although autonomous *hAT* elements in anole are typical members of their superfamily, they differ considerably in length. The *hAT-HT1_AC*, *hAT-HT2_AC*, *hAT-HT3_AC* and *SPIN_AC* families are all between 2 and 3Kb long but *hAT-4_AC* elements are much longer, ranging from 9 to 15Kb (Figure 11). This unusual length results from the incorporation in *hAT-4* elements of a considerable amount of extraneous DNA, including a number of partial transposon insertions from class II (such as multiple fragments of *Chapaev3-1_AC*) and class I (such as *RTE-Bov B*, *CR1* and *Sauria-SINE*). As these elements belong to families older than *hAT-4_AC*, it is likely that their presence in the sequence of *hAT-4_AC* elements result from the incorporation of large fragments of genomic DNA that fortuitously contained ancient TE fragments and not from insertion events in *hAT-4_AC* elements. The incorporation of genomic DNA has drastically increased the length of *hAT-4_AC*, yet it has not altered the replicative ability of this family. The phylogenetic tree in figure 11 recapitulates the evolution of the *hAT-4_AC* family. It is based on a 3Kb fragment common to all full-length *hAT-4_AC* elements that includes the transposase domain. Autonomous *hAT-4* elements are very similar in this 3Kb region, as suggested by the

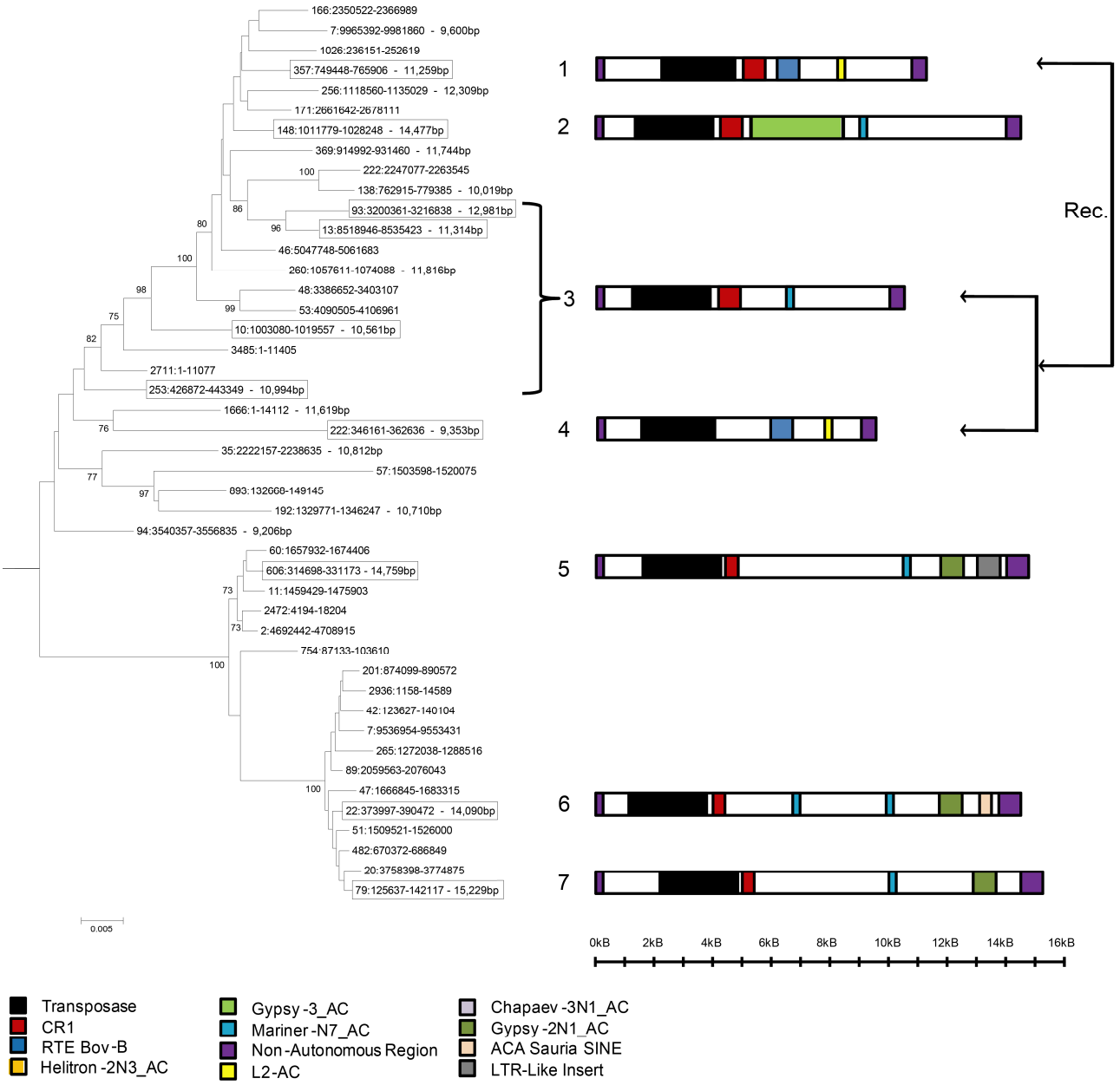

Figure 11- NJ tree of autonomous *hAT-4_AC* elements based on the transposase domain. Boxed sequences were analyzed for their length and nested TEs (see key) by locally running repeatmasker with a library of repetitive sequences found in the anole genome. Seven different patterns of nested elements were recovered and are detailed to the right of each sequence (structure 3 corresponds to elements 93, 13, 10 and 253 delineated by the bracket). Though all 45 elements are extremely similar to each other, they can still greatly differ in their length (indicated by the ruler) and mosaic of nested elements. The arrows on the right indicate the recombination of elements 3 and 4 resulting in element 1.

short length of the branches, yet they cluster in several distinct lineages that differ significantly in structure due to the frequent insertion or loss of DNA sequence (figure 11). The extreme structural variation in the relatively young *hAT-4_AC* family demonstrates that DNA transposons in anole can gain or lose DNA sequences at a very high rate. Additionally, we found that some elements were composite of other autonomous copies. For instance, the 5' half of element 1 (figure 11) is very similar to element 3 but its 3' half is identical in structure to element 4. This indicates that inter-element recombination can generate novel elements, thus increasing the structural diversity of the *hAT-4_AC* family.

The dynamic nature of *hAT-4_AC* evolution is also apparent in non-autonomous copies. *hAT-4_AC* is responsible for the amplification of at least 15 families (based upon sequence homology) of non-autonomous elements ranging in length from 1.3 to 3.4Kb. These non-autonomous families can be separated into two groups that contain slightly different TIRs: group A that contains families *hAT-4N1_AC* to *hAT-4N6_AC* and group B with families *hAT-4N7_AC* to *hAT-4N13_AC* (figure 12). The similarity between the TIRs of group A with the TIRs of autonomous *hAT-4_C* elements suggests that these elements result from deletions of the autonomous elements found in the anole genome. In contrast, group B elements are likely to have evolved from a subset of *hAT-4_AC* elements with different TIRs that is apparently no longer present in the anole genome. Within group A and B, elements have similar ends but differ drastically in structure due to insertions, deletions and the incorporation of genomic DNA, often containing TE from other classes or superfamilies. Using the presence of TE fragments embedded within non-autonomous families as markers, we were able to decipher the evolutionary history of these families (depicted on figure 13). The ancestral group A elements contained a 520bp *Penelope* insertion and a 333bp *ACA SINE*. A partial deletion (250bp) of the *Penelope* element occurred yielding families *hAT-4N2*, *3* and *5*. Independently, a partial deletion (176bp) of the *ACA SINE* occurred and is shared by families *hAT-4N1* and *6*. A recombination

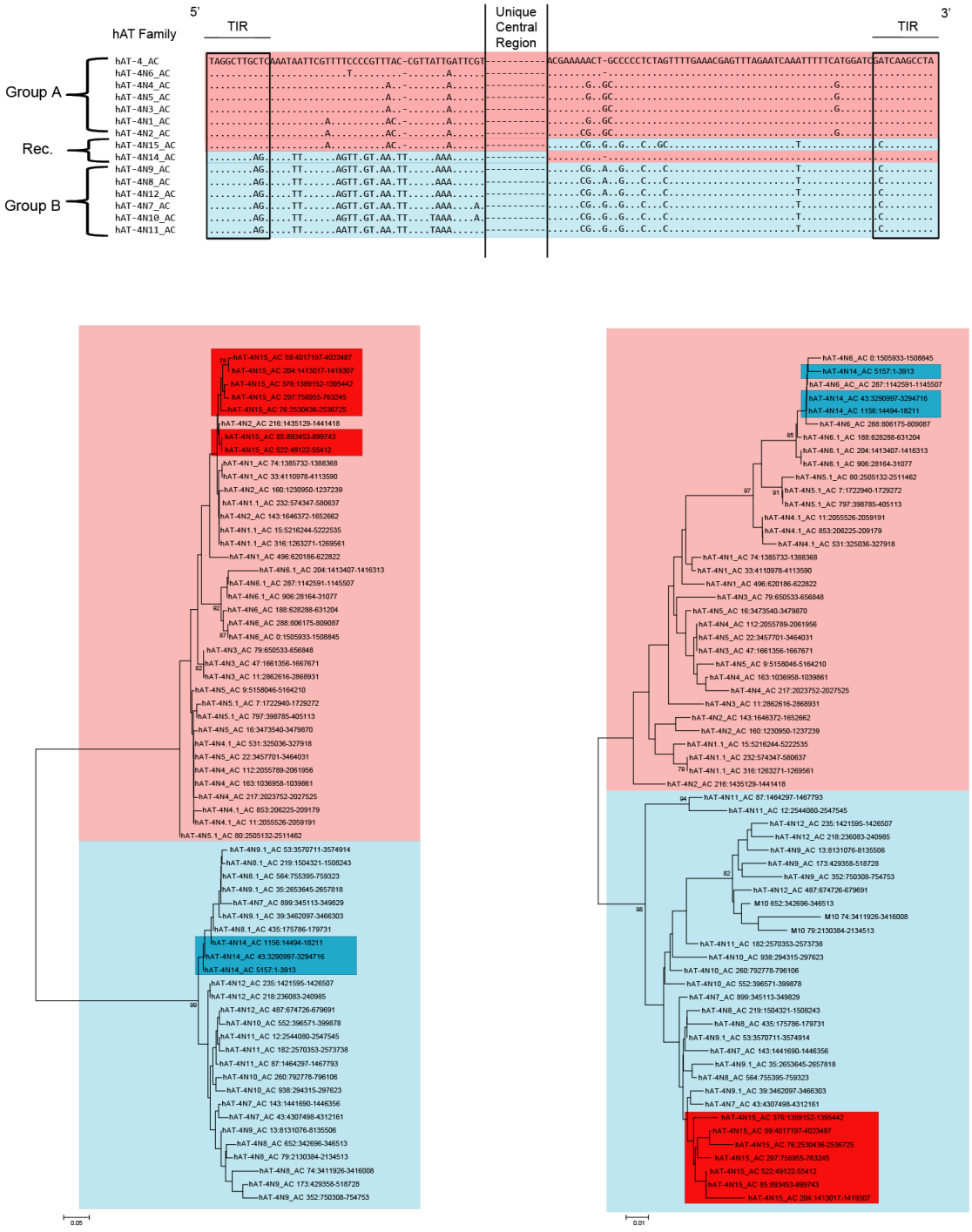

Figure 12: (A) 5' and 3' termini of consensus sequences of *hAT-4_AC* non-autonomous families and their autonomous *hAT-4_AC*. TIRs are boxed, and recombinant elements *hAT-4N14_AC* and *hAT-4N15_AC* are barred for emphasis. (B) NJ tree of 150bps of the 5' region and (Left) 300bps of the 3' region (Right) of non-autonomous *hAT-4_AC* elements. Bootstrap values less than 75% have been removed. At least three elements from each family are included. Boxes around elements reveal the group swap of *hAT-4N14_AC* (blue) and *hAT-4N15_AC* (red) in group A (light red) and group B (light blue).

event between an element containing the deleted version of *Penelope* and one containing the deleted *SINE* resulted in family *hAT-4N4*. Ancestral group B elements contain a *CR1* and a *Chapaev* insertion and are represented in the anole genome by families *hAT-4N7*, *8* and *9*. Two independent recombination events occurred between group B and group A elements: one resulted in a family that is identical to a typical group A element but with a 3' end similar to a group B element (*hAT-4N15_AC*) and the other produced a family which is similar to group B over most of its length but has a group A 3' end (figure 12 and 13). A fourth recombination event between family *hAT-4N15* and a group B element produced three families (*hAT-4N10*, *11* and *12*) with termini typical of group B but a central portion similar to group A.

The vast majority of *hAT* elements are relatively young as the average divergence from consensus is less than 10% for 85% of the families (figure 14). As expected, there is a relatively good concordance between the age of autonomous families and their non-autonomous counterparts. This is true of families *hAT-HT1_AC*, *hAT-HT 3_AC* and *hAT-4_AC*. For instance, the average divergence from consensus of family *hAT-4_AC* is 1.12% while the divergence of its non-autonomous relatives ranges from 0.49% to 1.15%. However, this might not be true of family *hAT-HT2_AC*. Autonomous *hAT-HT2* elements are extremely young and in fact their mean divergence from consensus is 0.00%. In contrast, their non-autonomous counterparts have divergence ranging from 0.73% to 1.75% and, thus, seem to pre-date their autonomous progenitor. It is plausible that these non-autonomous copies resulted from a previous wave of lateral transfer of the *hAT-2* family that would have produced non-autonomous families but failed to establish a resident population of autonomous copies.

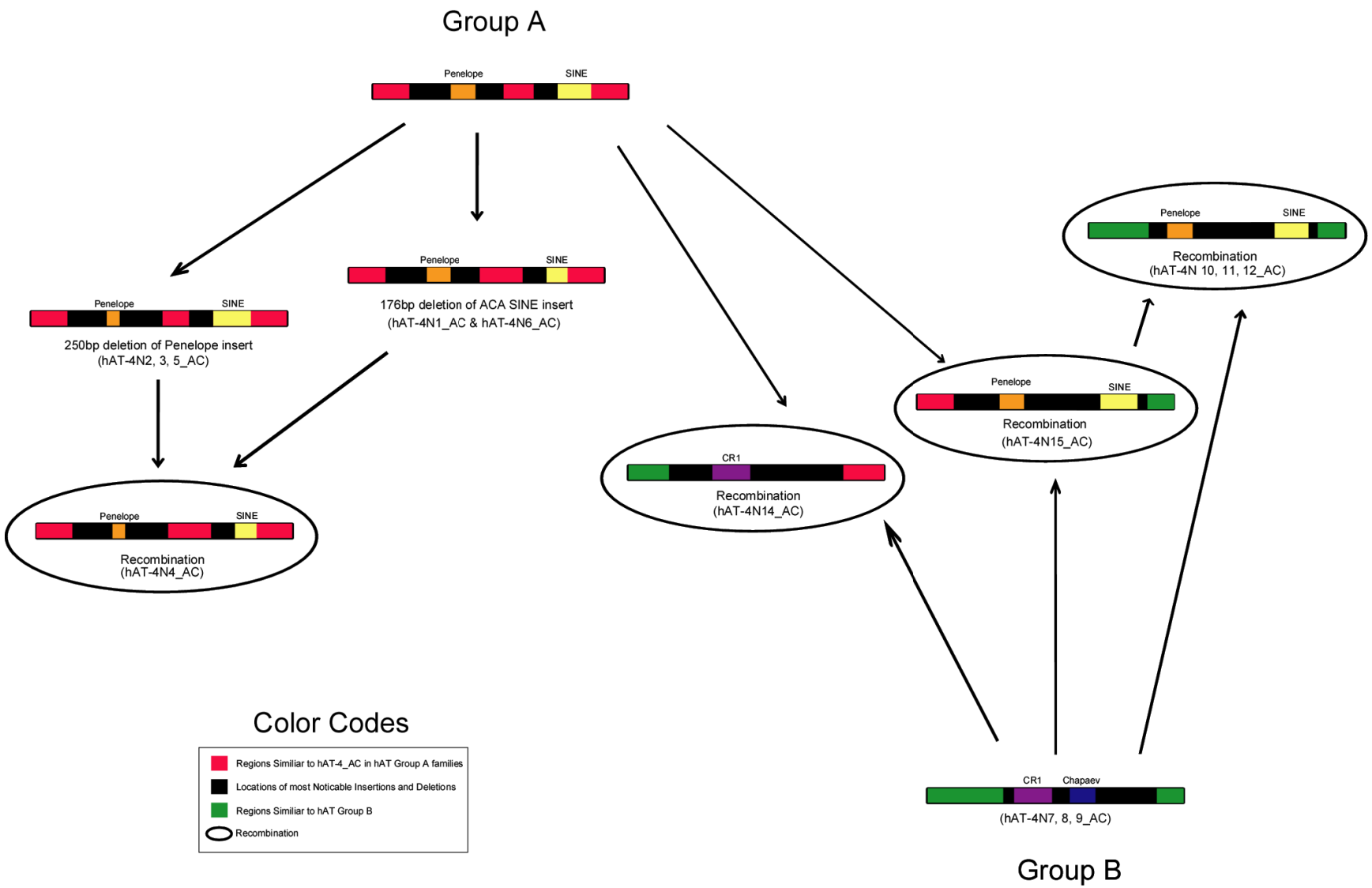

Figure 13: Structure and evolution of the 15 non-autonomous *hAT-4_AC* related families.

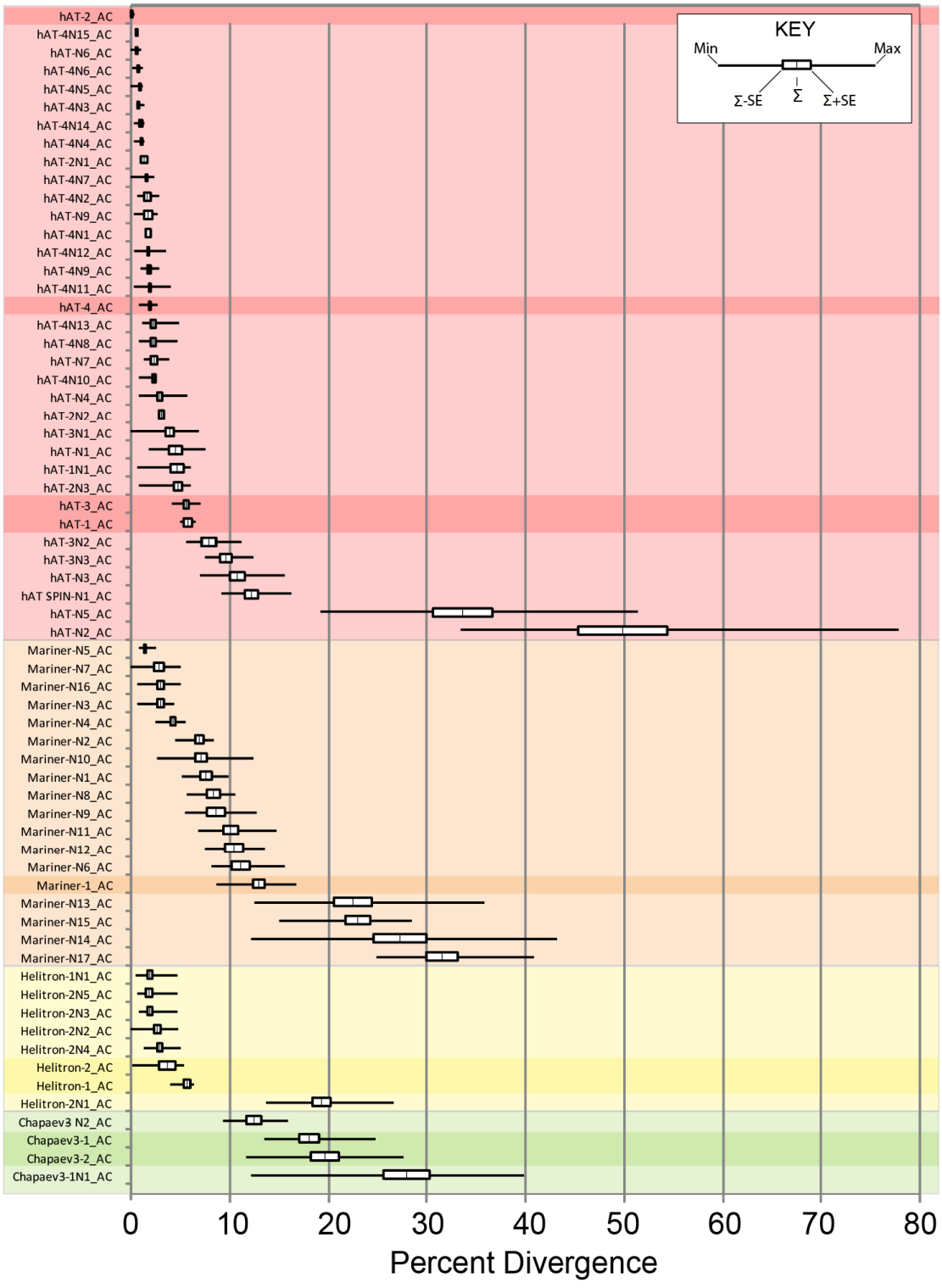

Figure 14: Divergence plot of *hAT* (red), *Mariner* (orange), *Helitron* (yellow), and *Chapaev* (green) families found in the genome of the lizard. Values were calculated using Kimura's 2-parameter method in Mega 4.0. Autonomous families are emphasized with darker bars.

The Mariner superfamily

The second most diverse superfamily of DNA transposons, the *Mariner* superfamily, consists of one autonomous family and 17 non-autonomous families (Table 4). These families are characterized by a TA TSD and TIRs that vary considerably in length and sequence. The single autonomous family, *Mariner-1_AC*, contains 58 copies and is related to the *Tc1* subset of *Mariner* elements. The 17 non-autonomous *Mariner*-like families range from 14 to ~5,000 copies for a total of ~17,500 copies. Surprisingly, none of the non-autonomous elements share any similarity with *Mariner-1_AC*. *Mariner-1_AC* is relatively ancient as elements diverge from each other by 13.0% on average and is probably inactive. Thus, it is unlikely that *Mariner-1_AC* is responsible for the recent burst of activity of several families (*Mariner-N3_AC*, *Mariner-N4_AC*, *Mariner-N5_AC*, *Mariner-N7_AC* and *Mariner-N16_AC*) which amplified to considerable numbers and have very low average divergence (less than 5%). It is also unlikely that *Mariner-1_AC* is responsible for the amplification of ancient families that predates its origin (such as *Mariner-N13*, *14*, *15* and *17*). The lack of overlap in divergence between the single autonomous family and the 17 non-autonomous families indicates that *Mariner-1_AC* is not responsible for their amplification and suggests that other *Mariner* families are or have been active in the anole genome. Despite numerous attempts to identify such autonomous families in the anole genome, we failed to find any other autonomous families. It is plausible that some of these elements never reached high copy numbers and because of the fast decay of TEs in anole (Novick et al. 2009a), are no longer identifiable. Unlike *hAT* elements, most *Mariner*-like families are relatively ancient and no longer active. In fact, only 5 families have an average pair-wise divergence lower than 5% and only two diverge by less than 3%. The rarity of autonomous copies and the age distribution of *Mariner*-like families suggest that mariner elements could be frequently invading the anole genome by lateral transfer, produce abundant non-autonomous families but fail to become stable residents of this genome.

Table 4: Characteristics and nomenclature of DNA transposons in the lizard *Anolis carolinensis*

Superfamily	Name	Copy Number >90% Identity	Length in bp	TSD	Length of TIR	TIR Sequence	% Divergence ± S.E.	% Divergence from consensus ± S.E.
Helitron	*Helitron-1_AC*	<5	8771	TA	-	-	5.72 ± 0.35	3.04 ± 0.19
	Helitron-1N1_AC	84	1436	TA	-	-	1.90 ± 0.25	1.02 ± 0.14
	Helitron-2_AC	<5	NA	TA	-	-	3.66 ± 0.87	1.95 ± 0.40
	Helitron-2N1_AC	66	1290	TA	-	-	19.4 ± 0.96	11.3 ± 0.66
	Helitron-2N2_AC	608	1930	TA	-	-	2.69 ± 0.34	1.58 ± 0.27
	Helitron-2N3_AC	>1000	2000	TA	-	-	1.98 ± 0.29	0.85 ± 0.10
	Helitron-2N4_AC	860	550	TA	-	-	2.96 ± 0.32	2.55 ± 0.30
	Helitron-2N5_AC	>1000	2000	TA	-	-	1.91 ± 0.36	1.78 ± 0.35
Mariner	*Mariner-1_AC*	58	1306	TA	39	HGADGGGGCGTTCATTAAAG[a]	13.0 ± 0.62	7.73 ± 0.38
	Mariner-N1_AC	901	489	TA	203	CGAGGGCTATCCAGAAAGTA[a]	7.62 ± 0.55	4.39 ± 0.43
	Mariner-N2_AC	>1000	783	TA	40	CGAGGGTTGAATGAAAAGTA[a]	6.93 ± 0.44	3.87 ± 0.27
	Mariner-N3_AC	>1000	427	TA	24	CCSTGTTTCCCCGAAAATAA[a]	3.04 ± 0.41	1.45 ± 0.19
	Mariner-N4_AC	>1000	661	TA	25	CYGTATATACTCGAGTATAA[a]	4.32 ± 0.27	2.32 ± 0.18
	Mariner-N5_AC	604	1396	TA	23	CCSTGTTTCCCCGAAAATAA[a]	1.42 ± 0.12	0.79 ± 0.08
	Mariner-N6_AC	889	405	TA	23	CCGTATATACTCGAGTATAA[a]	11.1 ± 0.90	6.62 ± 0.57
	Mariner-N7_AC	>5000	323	TA	24	CAGTAGAGTCTCACTTATCC[a]	2.91 ± 0.52	1.52 ± 0.38
	Mariner-N8_AC	>1000	468	TA	25	CAGTAGAGTCTCACTTATCC[a]	8.39 ± 0.66	4.83 ± 0.41
	Mariner-N9_AC	105	339	TA	15	CAGTGYYCCCTCRCT	8.65 ± 0.91	5.54 ± 0.57
	Mariner-N10_AC	287	664	TA	311	CGAGGGCTATCCAGAAAGTT[a]	7.14 ± 0.55	5.18 ± 0.53
	Mariner-N11_AC	348	401	TA	15	CAGTGTTCCCTCRCT	10.1 ± 0.74	5.18 ± 0.53
	Mariner-N12_AC	952	321	TA	20	GAGTCTCRCTTATCCAACMT	10.5 ± 0.87	5.74 ± 0.57
	Mariner-N13_AC	36	168	TA	22	CYGTATTTCTTCAATTSTAA	22.6 ± 1.90	10.5 ± 0.88
	Mariner-N14_AC	14	240	TA	31	GATTGTAGCTACAGTATGAC[a]	27.3 ± 2.74	15.2 ± 1.67
	Mariner-N15_AC	498	524	TA	18	CASRKGTGTCAAACTCAA	23.1 ± 1.24	12.4 ± 0.76
	Mariner-N16_AC	3771	350	TA	15	CAGTGTTCCCTCACT	3.02 ± 0.34	1.62 ± 0.19
	Mariner-N17_AC	137	356	TA	26	CAGGTTGAGYATCCCTTATC[a]	31.6 ± 1.61	17.1 ± 1.21
Chapaev	*Chapaev3-1_AC*	15	1910	TWA	12	CACTGRWAAACA	18.1 ± 0.99	12.3 ± 1.10
	Chapaev3-2_AC	4	1767	TWA	18	CACTAGGAAACACAATTT	19.7 ± 1.46	10.9 ± 0.90
	Chapaev3-1N1_AC	46	390	TWA	11	CACWGSCCAAC	28.0 ± 2.39	15.9 ± 0.92
	Chapaev3-1N2_AC	988	320	TWA	19	CACTATGTAACAAAATTT	12.5 ± 0.77	6.00 ± 0.42

[a]Only the first 20bp of the TIR are presented here.

The Helitron superfamily

The third superfamily in anole is the *Helitron* superfamily. It is represented by two autonomous families, *Helitron-1_AC* and *Helitron-2_AC*, both in very low copy number (less than 5 copies) and 6 abundant non-autonomous families (Table 4). All these elements are typical members of the *Helitron* superfamily, lacking TIRs and with TA TSDs (Kapitinov and Jurka 2001). Unlike *hAT* and *Mariner* non-autonomous elements, all non-autonomous *Helitron* families are internally deleted versions of autonomous families; *Helitron-1N1_AC* is derived from *Helitron-1_AC* while the remaining 5 non-autonomous families (*Helitron-2N1-5_AC*) resulted from large deletions of *Helitron-2_AC*. Helitrons appear to have been recently active as suggested by their relatively low level of divergence and, as expected, the amplification of most non-autonomous families is concomitant to the activity of their autonomous progenitors. The

only exception is the oldest family, *Helitron-2N1_AC*, which divergence lies clearly outside the divergence distribution of all other families. Yet, this family is unambiguously related to the autonomous *Helitron-2_AC* family. This suggests that the *Helitron-2_AC* family has been active in the anole genome much longer than its current divergence suggests. It is possible that *Helitron-2_AC* is in fact an ancient resident of the anole genome but remained in such low copy numbers that older copies are no longer identifiable. However, it is surprising that *Helitron-2_AC* failed to produce additional non-autonomous families after *Helitron-2N1_AC*. It is thus possible that, once a resident of the anole genome, *Helitron-2_AC* became extinct and only recently re-colonized the anole genome possibly through lateral transfer.

The Chapaev superfamily

The recently described *Chapaev* superfamily (Bao et al. 2009) is also present in the anole genome and consists of 4 families, 2 of which are autonomous. Elements in this superfamily display a 3bp TSD of consensus TWA and TIRs beginning with the tri-nucleotide CAC. They belong to the subsets of elements classified as *Chapaev3* in Repbase. These families are the oldest one we found and are no longer active: the oldest family contains elements that are over 35% divergent while the youngest contains elements that are no less than 12% divergent. Although these families are old, ORFs are still detectable in the two autonomous families, *Chapaev3-1_AC* (1910bp in total length) and *Chapaev3-2_AC* (1767bp in total length) of 485 and 561aa's respectively but no putative conserved domains have been identified by domain finder. The two remaining families of Chapaev elements are non-autonomous families (*Chapaev3-3N1_AC* and *Chapaev3-3N2_AC*) and are also ancient, yet they are not directly derived from the autonomous families.

Discussion

The genome of *Anolis carolinensis* harbors an extraordinary diversity of class II transposons. A total of 68 families, including 10 autonomous ones, representing four recognized superfamilies (*hAT*, *Mariner*, *Helitron* and *Chapaev*) were identified in the anole genome. Only one superfamily, *Chapaev*, is extinct, whereas the other three superfamilies have recently been active increasing the genetic diversity of the green anole. Each superfamily is also represented by a large diversity of non-autonomous families that are either internally deleted versions of autonomous copies (most *hAT* and all *Helitron* families) or do not show similarity with autonomous copies (some *hAT* and all *Mariner* families). As in plants and other animals, non-autonomous families greatly out-number their autonomous counterparts (Le Rouzic and Capy 2006; Hartl, Lozovskaya, and Lawrence 1992). The most active and diverse superfamily, *hAT*, exemplifies a number of mechanisms that can increase the diversity of class II transposons. First, the frequent lateral transfer of active hAT families has a dramatic impact on transposons diversity. 13 out of 38 *hAT* families (*hAT-1*, *2*, *3* and their non-autonomous relatives) are the direct or indirect result of lateral transfer (Pace et al. 2009; Novick et al. 2009b). Although we failed to find elements similar to *hAT-4_AC* in other genomes, we cannot exclude that this family was also laterally transferred. In fact, the recent burst of activity of *hAT-4* elements and the absence of old (>5% divergence) autonomous or non-autonomous *hAT-4* related families suggests that *hAT-4* is a new inhabitant of the anole genome; hence, we cannot exclude that all *hATs* in anole result from lateral transfer. Second, non-autonomous and autonomous *hAT-4* elements show considerable structural variation resulting from their ability to capture and incorporate extraneous DNA sequences. This mechanism produced autonomous elements that are the longest reported in the *hAT* superfamily and some non-autonomous *hAT-4* elements are longer than autonomous copies of other families. As the filler DNA often contains TEs of other classes and superfamilies, the amplification of *hAT-4* families significantly contributes to

increasing the copy number of the elements they transport. The unusual ability of *hAT-4* to incorporate very large extraneous sequences suggests that *hAT-4* would be an excellent candidate for the development of DNA delivery vectors. Finally, inter-element recombination seems to occur readily and is responsible for the diversification of both autonomous and non-autonomous families. In particular, recombination between non-autonomous copies resulted in the formation of 6 novel families of elements.

The age distribution of class II families differs among superfamilies and reveals different dynamics of amplification in the anole genome. The *hAT* superfamily shows the highest level of recent activity as most *hAT* families have very low level of divergence from their consensus (<5%). In fact, some of these families have such low divergence (<0.5%) that they are certainly active in extant anoles. The lack of older *hAT*-related families is likely to reflect the fact that this genome was recently colonized by *hAT* elements through lateral transfer. It is however surprising that 4 (or maybe 5) hAT families invaded the anole genome recently but that lateral transfer did not occur in the more distant past. The presence of a couple of ancient orphan families (>10% from consensus) suggests that episodes of lateral transfer indeed occurred but had limited success at invading the genome of ancestral anoles. The Mariner superfamily shows a more evenly distributed range of ages, from families that are extremely young and possibly active, to much older families (up to 15% divergent from their consensus) that have long been extinct. Yet, young families seem to be predominant and, to a lesser extent than for hATs, the age distribution of Mariner families is skewed towards young families. Similarly, Helitrons families tend to be very young, although the presence of an ancient non-autonomous family suggests they are not newcomers to the anole genome. The only exception to this pattern is the long-extinct Chapaev superfamily that contains only very divergent elements. The overall pattern of divergence of class II families indicate that DNA transposons readily reach fixation and accumulate in the anole genome as revealed by the continuous distribution of families

diverging by 0 to 10%, yet old (>10%) families are comparatively very rare and under-represented in this genome. A possible explanation is that DNA transposons had a very low level of activity in the ancestral anole genome. However, it is surprising that the same pattern of low activity, followed by a more recent burst of activity, is shared by three unrelated superfamilies (*hAT*, *Mariner* and *Helitron*). Another possibility is that transposons in anole are decaying rapidly by accumulation of indels and that they are no longer recognizable past a certain age. This rapid decay of mobile elements in the anole genome was previously reported for the *RTE Bov-B-1_AC* family of retrotransposons (Novick et al. 2009a). This family's average divergence is only 4%, yet more than 70% of the full-length elements *RTE Bov-B-1_AC* generated have lost considerable amount of sequences through deletions. It would therefore not be surprising that elements from families older than 10% could be so fragmented that our approach failed to recognize them.

Class II transposons coexist in anole with a plethora of retrotransposon families. At least 42 families of non-LTR retrotransposons belonging to 5 clades are concurrently active in the anole genome (Novick et al. 2009a). However, the dynamics in the genome of class I and class II elements differ drastically. The vast majority of non-LTR retrotransposons families are much younger than class II families as they all have divergence lower than 2% (except 4 families out of 46). In addition, retrotransposon families are on average far less numerous than class II families. It seems that, unlike class II elements, the vast majority of retrotransposon insertions does not reach fixation. This pattern is best explained by a high turnover of insertions resulting from a balance between the transposition of new elements and the selective loss of deleterious alleles. Interestingly, class II transposons do not seem to be subjected to the same level of purifying selection as class I elements. It was proposed that selection against non-LTR retrotransposons is length-dependent and was caused, for the most part, by their ability to mediate ectopic recombination (Novick et al. 2009a; Song and Boissinot 2007). Therefore, short

elements should not be eliminated by purifying selection to the same extent as long ones are and, consequently, should accumulate in the genome of their host. Indeed, most DNA transposons, in particular non-autonomous ones, are less than 2 Kb long and much shorter than non-LTR retrotransposons. It is therefore likely that these elements are not as deleterious as retrotransposons and readily reach fixation. Interestingly, the longest class II transposons, the *Helitrons* and *hAT-4* elements (autonomous and non-autonomous), have a very low level of divergence and relatively low copy numbers, reminiscent of retrotransposons. It is therefore possible that those elements are subject to the same selective forces as non-LTR retrotransposons and are not reaching fixation, which could explain the lack of older *Helitron* and *hAT-4* related families.

The very large diversity of class II transposons in anole is similar to the one reported in teleost fish (Krasnov et al. 2005; repbase reports). The analysis of teleostean genomes including salmonids, fugu and zebrafish revealed that these genomes are also littered with DNA transposons, in particular from the TC1/Mariner superfamily. In addition, Repbase contains a large number of DNA transposon sequences from fugu and zebrafish. In contrast, mammalian genomes tend to lack active DNA transposons, although class II elements have been diverse and prolific in early mammalian evolution. The human genome for instance contains a large number of class II families that account for 3% of its size. However, these elements have become extinct around 40MY ago (Pace and Feschotte 2007; Lander et al. 2001). The examination of other mammalian genomes such as the dog, cow, rat and mouse reveal a similar trend with an extinction of all DNA transposons activity about 40MY ago (Lander et al. 2001; Lindblad-Toh et al. 2005; Waterston 2002). However, mammalian genomes are not inhospitable to DNA transposons as revealed by the analysis of the little brown bat genome that harbor a large diversity of active transposons (Ray et al. 2007; Ray et al. 2008). It seems clear that the diversity of DNA transposons in the anole genome is more similar to the one observed in fish

than in mammals. The same pattern was reported for non-LTR retrotransposons: the low divergence and low copy number of retrotransposon families resulting from a high turnover of elements in anole parallels the one reported in teleost fish (Novick et al. 2009a; Duvernell, Pryor, and Adams 2004; Furano, Duvernell, and Boissinot 2004; Volff, Korting, and Schartl 2000). This pattern suggests that the regulation of transposable elements, from class I and II, differ radically between mammals and non-mammalian vertebrates and constitute one of the most important evolutionary transitions in evolutionary genomics.

CHAPTER III: HORIZONTAL TRANSFER

Independent and Parallel Lateral Transfer of DNA Transposons in Tetrapod Genomes

Introduction

The lateral (horizontal) transfer of genetic information has had a profound impact on the evolution of unicellular organisms, in particular prokaryotes, yet the evolutionary significance of lateral gene transfer in multicellular eukaryotes remains unclear (Andersson 2005). In eukaryotes, beside the transfer of genetic information from organelles to the nucleus, which has occurred repeatedly (Timmis et al., 2004), most cases of lateral transfer have been documented in phagotrophic protists (Andersson 2005). In multicellular eukaryotes, lateral transfer has occurred relatively frequently in plants (Bergthorsson et al. 2003; Won and Renner 2003) but seems exceedingly rare in fungi and animals. Most documented cases of lateral transfer in animals involve the transfer of transposable elements (TEs), usually between closely related species. The lateral transfer of TEs seems to have occurred most frequently in insects but recent studies have demonstrated its occurrence in fish (de Boer et al., 2007) and tetrapods (Kordis and Gubensek 1998; Pace et al. 2008).

TEs are ubiquitous in eukaryotes and have profoundly affected the size, structure and function of eukaryotic genomes. TEs have also been an important source of evolutionary novelties: exaptation of TEs within coding sequences (i.e. exonization) or as regulatory elements seems to have been relatively common (Makalowski 2000; Nekrutenko and Li 2001; Bejerano et al. 2006) and could have driven the evolution of some fundamental biological functions (Feschotte and Pritham 2007; Feschotte 2008). However, the activity of TEs can also negatively affect the fitness of their hosts (Petrov et al. 2003; Boissinot et al. 2006), either by disrupting gene function or by causing chromosomal rearrangements (Kazazian 2004). Two main classes of TEs inhabit animal genomes: class I elements that use an RNA intermediate for their replication and class II elements, also called DNA transposons, that replicate using a cut-and-paste mechanism (for a review on transposons and their impact, see (Feschotte and

Pritham 2007). DNA transposons are mobilized by a self-encoded enzyme called transposase that recognizes specific sites at the end of the element, excises the DNA and inserts it elsewhere in the genome. Although the transposition mechanism is non-replicative, DNA transposons can reach large copy number in some species due to donor site repair. As the only recognition requirements of the transposase are the terminal sequences of the elements, transposons with internal deletions can also be mobilized by the transposase encoded by complete elements. Consequently, families of autonomous DNA transposons are parasitized by a plethora of non-autonomous elements that often out-number their autonomous relatives. DNA transposons are very diverse and are represented by 15 super-families that diverged before the diversification of eukaryotic lineages (Bao et al. 2009).

The general mode of transmission of TEs in animals is vertical. Phylogenetic analyses of class I TEs, such as the LINE-1 retrotransposons, demonstrate that the evolution of these elements can be explained by a strict vertical mode of transmission (Furano 2000; Kordis et al. 2006). The transmission of the RTE Bov-B elements from squamate reptiles to bovids constitutes at current time the only convincing exception to this model (Kordis and Gubensek 1998). The evolution of DNA transposons is also typified by the vertical transmission of elements. However, lateral transfer of DNA transposons can occur, though rarely, and most often between closely related species. It has been unequivocally demonstrated in insects (Daniels et al. 1990; Robertson and Lampe 1995) and fish (de Boer et al. 2007). Recently, Pace et al. (2008) described the first case of lateral transfer in mammals. The elements involved, named *SPIN*, were found in several distantly related lineages (tenrec, bush baby, murine rodents, bat, opossum, lizard and frog), yet they were absent from 27 other vertebrate genomes. The patchy phylogenetic distribution of these elements, coupled with their high sequence similarity, convincingly demonstrated that *SPIN* elements have been transferred laterally to these lineages, from a still unknown source.

During an analysis of repetitive sequences in the genome of the lizard *Anolis carolinensis*, we identified four novel families of DNA transposons that show the hallmark of lateral transfer. Surprisingly, these lateral transfers did not occur randomly in the tree of life as the same species that were hospitable to *SPIN* elements have been colonized by other distantly related families of elements. The parallel and independent invasion of the same genomes by different DNA transposons suggests that some genomes might be intrinsically more hospitable to DNA transposons (or horizontal transfer in general) than others, possibly because they lack defense against these elements or because they developed mechanisms to tolerate their impact.

Materials and Methods

The genome of *A. carolinensis* was searched for the presence of DNA transposons using the RepeatMasker (www.repeatmasker.org) and PILER programs (Edgar and Myers, 2005). Additionally, we performed a BLASTX search of the Anole genome using amino acid sequences derived from the transposon library available from Repbase (v13.01). The resulting hits of at least 100 bp, with a maximum score of $2e^{-150}$, were extracted along with 5,000 bp of flanking sequence to identify precisely the termini of each element. Elements that were recovered were aligned to each other and classified into families, and consensus sequences were constructed for each family of autonomous and non-autonomous transposons. Each consensus sequence was then screened for target site duplications, terminal inverted repeats and the presence of a transposase domain for proper classification. Sequences were manipulated and aligned using the BioEdit program (Hall 1999).

To identify related elements in other species, nucleotide consensus sequences and their translated proteins were used as queries in a BLAST and BLASTX search of GenBank and in a BLAT search of the genomes curated on the UCSC webpage (http://genome.ucsc.edu). Significant hits (>80%) were collected and aligned, and consensus sequences were created and

compared among species. Silent (*ds*) and replacement (*dn*) divergences between consensus sequences, as well as their ratio, were calculated using the maximum likelihood approach developed by Yang and Nielsen (Yang and Nielsen 2000), implemented in PAML (Yang 2000). The mean divergence between elements and its standard deviation was calculated for each family as an estimator of its age, using Kimura 2-parameters method (Kimura 1980). Phylogenetic analyses were performed with the neighbor joining method (NJ) (Saitou and Nei 1987) and maximum likelihood method using PHYML (Guindon et al. 2005). Distances and NJ trees were calculated using the MEGA 4.0 program with 1,000 bootstrap replicates (bootstrap values lower than 75 are not shown) (Kumar et al. 2001). Copy numbers were estimated using BLAST or BLAT and only elements containing >85% of the consensus query were tallied as elements representative of that family. Because some of the genome sequences are not complete, the copy number presented here are only rough estimates.

We confirmed the presence of transposon families by Polymerase Chain Reaction (PCR). PCRs were conducted in organisms that harbor the putatively transferred transposons families and in organisms for which we did not expect to find amplification. Twenty microliters of PCR reactions was conducted with an initial denaturing step of 94° for 5 min, followed by 30 cycles of 94°, 52° and 72° for 30s each and ended with a 7-min extension at 72°. Primers were constructed to amplify elements in all taxa in which they are found by utilizing the most conserved regions in the ORF while minimizing ambiguous nucleotides. The primers used are as follows: hAT-HT1, Forward: 5'CTCTACAAATTTAACATCWG, Reverse: 5'AATCAACYTC-CWGTAAAAGT ; hAT-HT2, Forward: 5'TTCCTATTTTTCCTTGGCAC, Reverse: 5'AAGCTCTT-TAAATCTVADTTG ; and hAT-HT3, Forward: 5'AGTTCCCAAATTTATCAGGAG, Reverse: 5'GCAATTAAACAGAAAGATAT.

Results

The screening of the green anole genome (*A. carolinensis*) revealed the presence of 11 autonomous and 69 non-autonomous families of DNA transposons (not shown). Four of these 80 families yielded highly significant (>90%) hits in other species using the BLAST or the BLAT search engines.

Four shared families of DNA transposons

We found five copies of a *hAT* (*hobo/Activator/Tam3*) element in the green anole, which are virtually identical to each other and have intact Open-Reading Frames (ORF). This element is 2246bp long and encodes a 602 amino acid (a.a.) long transposase (Table 5). Highly similar elements (i.e. less than 5% divergence at the DNA level) were found in eight vertebrate genomes (Figure 15), including three primates (two strepsirrhini, *Otolemur garnetii* and *Microcebus murinus*, and the tarsier, *Tarsius syrichta*), the tenrec (*Echinops telfairi*), the little brown bat (*Myotis lucifugus*), the opossum (*Monodelphis domestica*) and the African clawed frog (*Xenopus tropicalis*). In each of these species, except the bat, we found full-length copies of the element. We named this family *hAT-HT2* to reflect its horizontal mode of transmission. Note that this family includes the previously described opossum *hAT2_MD* family but not the *hAT2* family of bats and frog because *hAT2* families in different organisms do not form a monophyletic group. We also found fragments highly similar to *hAT-HT2* in the planarian *Schmidtea mediterranea* indicating that this transposon is not limited to vertebrate genomes. *hAT-HT2* is apparently absent from all other vertebrate and invertebrate species for which genomic data are available (>50 species). The discontinuous phylogenetic distribution of *hAT-HT2* was confirmed by PCR (Figure 16). As expected, we amplified this family in lemur, opossum, anole and planarian but not in any other species tested. Surprisingly, we failed to amplify *hAT-HT2* in the

Table 5: Characteristics of 3 horizontally transferred autonomous *hAT* families and their non-autonomous realatives

Family	Organism	TE name	Length, bp	Copy no.[a]	Avg. divergence ± S.E.	TIR
hAT-1	*Myotis lucifugus*	*hAT-HT1_ML*	2,920	38	2.68 ± 0.17	CAGTGATGGCGAACCT
		hAT-HT1N1_ML	235	>5,000	1.44 ± 0.51	CAGTGATGGCGAACCT
	Monodelphis domestica	*hAT-HT1_MD*	3,002	100	8.68 ± 0.50	CYNTGATGGNNAANC
		hAT-HT1N1_MD	1,010	>500	7.29 ± 0.42[c]	CMRTGATGGSNAACCT
	Anolis carolinensis	*hAT-HT1_AC*	2,968	6	5.82 ± 0.40	CAGTGATGGSSAACCT
		hAT-HT1N1_AC	585	868	4.75 ± 0.64	CARTGATGGSCAACCT
hAT-2	*Tarsius syrichta*	*hAT-HT2_TS*	3,381	5	4.56 ± 0.32	CAGGGGTCCTCAAAC
		hAT-HT2N1_TS	674	50	4.50 ± 0.55	CAGGGGTCCTCAAACT
		hAT-HT2N2_TS	506	65	4.39 ± 0.40	CAGGGGTCCTCAAACT
		hAT-HT2N3_TS	626	65	4.33 ± 0.36	CAGGGGTCCTCAAACT
	Otolemur garnetii	*hAT-HT2_OG*	3,277	23	7.81 ± 0.43	CAGGGGTCCTCAAAC
	Microcebus murinus	*hAT-HT2_MM*	2,169	<10	7.23 ± 0.58	CAGGGGTCCTCAAAC
		hAT-HT2N1_MM	205	>1,000	5.71 ± 1.27	CAGGSGTCCTCAAAC
	Myotis lucifugus	*hAT-HT2_ML*	NA	<5	3.51 ± 0.72	CAGGGGTCCCCAAAC[b]
		hAT-HT2N1_ML	746	27	2.16 ± 0.33	CAGGGGTCCTCAAACT
		hAT-HT2N2_ML	205	>1,000	1.48 ± 0.57	CAGGGGTCCTCAAACT
	Echinops telfairi	*hAT-HT2_ET*	3167	16	7.55 ± 0.40	CAGGGGTCCTCAAAC
		hAT-HT2N1_ET	225	>1,000	4.38 ± 0.81	CAGGGGTCCTCAAACT
	Monodelphis domestica	*hAT-HT2_MD*	3,128	267	9.11 ± 0.37	CAGGAGTCCCCAAAC
	Anolis carolinensis	*hAT-HT2_AC*	2246	5	0.14 ± 0.07	CAGGGGTCCCCAAAC
		hAT-HT2N1_AC	1485	28	1.34 ± 0.31	CAGGGGTCCCCAAACT
	Xenopus tropicalis	*hAT-HT2_XT*	NA	<5	NA	YAGGARTCCTCAAAC[b]
	Schmidtea mediterranea	*hAT-HT2_SM*	NA	<5	11.74 ± 1.41	NA
hAT-3	*Myotis lucifugus*	*hAT-HT3N1_ML*	326	537	0.24 ± 0.16	CAGTGRTTCCCAAA
	Anolis carolinensis	*hAT-HT3_AC*	2,754	<25	5.62 ± 0.29	CAGTGRTTCCCAAA
		hAT-HT3N1_AC	326	344	3.98 ± 0.37	CAGTGRTTCCCAAA
	Schmidtea mediterranea	*hAT-HT3_SM*	NA	>20	12.23 ± 1.55	CAGTGGTTCCCAAA

[a]Due to the fact that many of these genomes are not yet fully sequenced, copy number depicts only elements already sequenced and collected from NCBI with >90% sequence identity to the consensus.

[b]TIR is categorized by only the 3' end of the element

[c]5' end removed from divergence calculations due to variation and subfamily structure.

frog, possibly because the *hAT-HT2* copies in this species are too fragmented to yield an amplification.Full-length or partial consensus sequences were constructed for each species (except for *S. mediterranea* because most elements were highly fragmented) and compared to each other. The synonymous divergence (*ds*) between transposase genes, that solely reflects the neutral history of the sequence, is relatively low, < 7% for all inter-specific comparisons (table 6), considering how distantly related these species are. For instance, the *ds* between *Anolis* and the little brown bat is about 5%, although reptiles and mammals diverged about 300

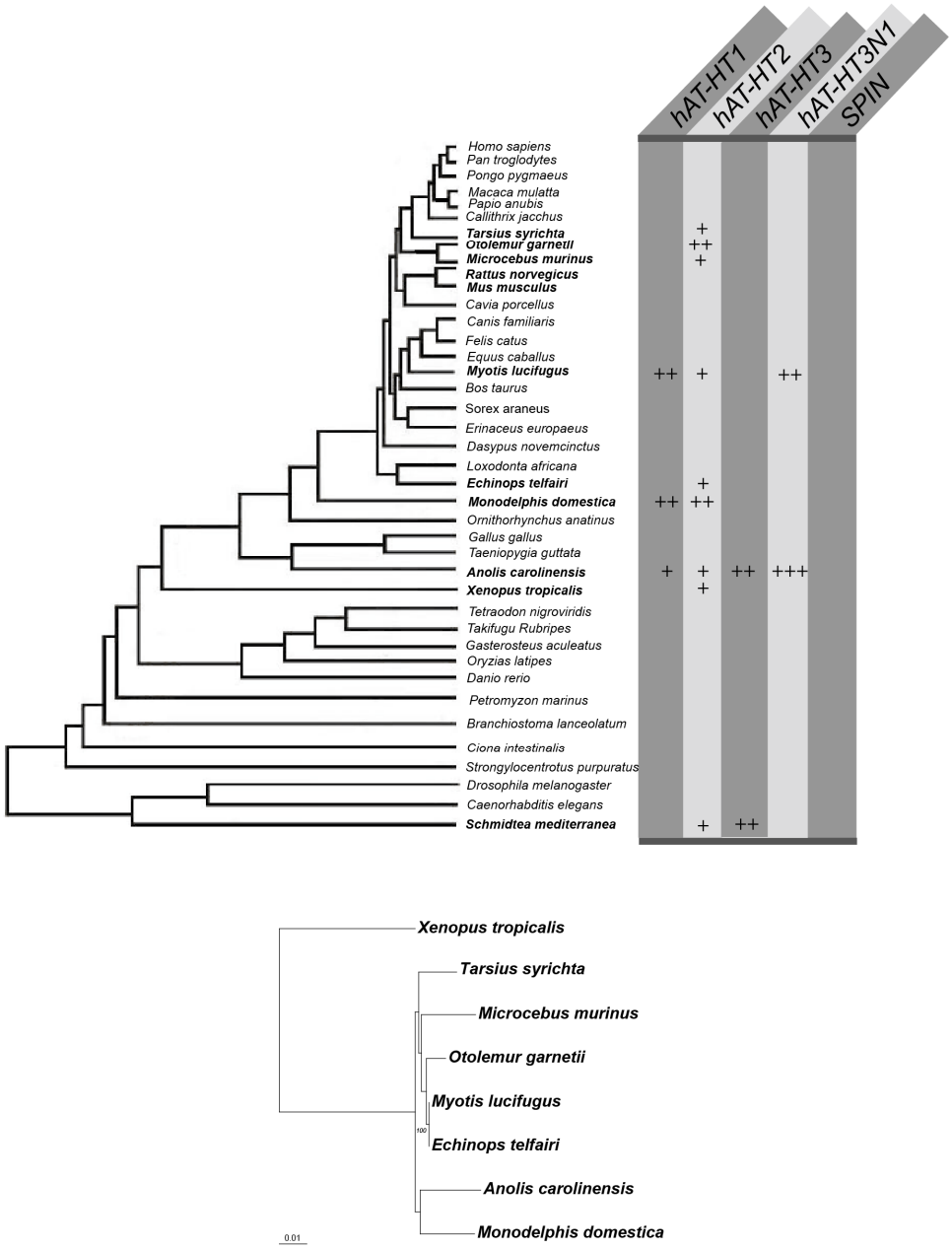

Figure 15: (A) Distribution of five laterally transferred *hAT* DNA transposon families in animal species for which sufficient genome sequence is available. Species in bold contain at least one of the 5 families of horizontally transferred elements. The abundance of each family is symbolized as follows: + less than 10 elements, ++ more than 10 and +++ more than 1,000 elements. Branch lengths are not indicative of time or genomic divergence. (B) Maximum likelihood tree based on the consensus sequence of the transposase domain of *hAT-2* elements in eight of the nine species that genomic copies were found (elements in *S. mediterranea* are too fragmented for this analysis). The model of substitution (HKY+G) was determined using Modeltest (Posada and Crandall, 1998) and the tree was built under this model using PHYML with 1000 bootstrap replicates (Guindon et al., 2005). Bootstrap values less than 75 have been removed.

MY ago (Donoghue and Benton 2007). Similarly among mammals, the *ds* between the opossum and the other mammals is remarkably low (<5%) considering that marsupials and eutherians split 120 MY ago (Donoghue and Benton 2007). Such a low level of silent divergence between distantly related lineages is not consistent with a vertical mode of transmission of the *hAT-HT2* transposons and, instead, strongly suggests it has been laterally transferred. A phylogenetic analysis based on *hAT-HT2* consensus sequences shows a star-like phylogeny, with extremely short internal branches (Figure 15-B), also supporting the lateral transfer of *hAT-HT2.* The tree also suggests that the lateral transfer likely originated from a single common source. The length of the terminal branches indicates that *hAT-HT2* elements have resided in each genome long enough to accumulate nucleotide substitutions since they were laterally transferred. Interestingly, transposases differ at many non-synonymous positions which is surprising considering that the integrity of the transposase is critical for the mobilization of DNA transposons. We calculated the ratio *dn/ds* between transposase consensus sequences and we found that the values of the ratio are remarkably high for a protein coding sequence, although only a few were higher than 1 (Table 6). These high *dn/ds* values suggest that the transposase gene evolved neutrally, which is consistent with the very discrete, and relatively short lifespan of *hAT-HT2* in the anole and in the opossum (figure 17). However, the distribution of pairwise divergence in *O. garnetii* (the bush-baby; figure 17) suggests that *hAT-HT*2 has persisted in this species for about 18MY (assuming a mutation rate in prosimians of ~0.2% per MY (Liu et al. 2003)). It is very unlikely that *hAT-HT2* would have persisted so long in this genome if the transposase was evolving neutrally. Instead, the high *dn/ds* values could result from the action of positive selection, i.e. selection in favor of amino acid changes. Under this scenario, the colonization of the naïve bush-baby genome by *hAT-HT2* could have been followed by a phase of adaptation of the transposon to its new genomic environment, possibly in response to host defense.

Table 6: Summary of non-synonomous (dn) and synonomous (ds) substitutions of the four families of autonomously replicating horizontally transferred *hAT* families.

dn

hAT-HT1

	Md	*Ml*	*Ac*
Monodelphis domestica (Md)			
Myotis lucifugus (Ml)	0.0983		
Anolis carolinensis (Ac)	0.0157	0.0932	

hAT-HT2

	Ts	*Ml*	*Mm*	*Md*	*Og*	*Et*	*Ac*
Tarsius syrichta (Ts)							
Myotis lucifugus (Ml)	0.0444						
Microcebus murinus (Mm)	0.0298	0.0527					
Monodelphis domestica (Md)	0.0337	0.0432	0.0407				
Otolemur garnetii (Og)	0.0191	0.0364	0.0270	0.0230			
Echinops telfairi (Et)	0.0147	0.0342	0.0225	0.0208	0.0065		
Anolis carolinensis (Ac)	0.0229	0.0363	0.0308	0.0257	0.0179	0.0135	

hAT-HT3

	Ac
Schmidtea mediterranea (Sm)	0.0098

SPIN

	Xt	*Ac*	*Og*	*Et*	*Ml*	*Md*
Xenopus tropicalis (Xt)						
Anolis carolinensis (Ac)	0.1002					
Otolemur garnetii (Og)	0.0580	0.0529				
Echinops telfairi (Et)	0.0541	0.0513	0.0101			
Myotis lucifugus (Ml)	0.0578	0.0514	0.0118	0.0067		
Monodelphis domestica (Md)	0.0997	0.0952	0.0606	0.0581	0.0558	

ds

hAT-HT1

	Md	*Ml*	*Ac*
Monodelphis domestica (Md)			
Myotis lucifugus (Ml)	0.3002		
Anolis carolinensis (Ac)	0.0315	0.3201	

hAT-HT2

	Ts	*Ml*	*Mm*	*Md*	*Og*	*Et*	*Ac*
Tarsius syrichta (Ts)							
Myotis lucifugus (Ml)	0.0299						
Microcebus murinus (Mm)	0.0205	0.0371					
Monodelphis domestica (Md)	0.0407	0.0493	0.0428				
Otolemur garnetii (Og)	0.0171	0.0206	0.0194	0.0420			
Echinops telfairi (Et)	0.0102	0.0229	0.0124	0.0347	0.0068		
Anolis carolinensis (Ac)	0.0450	0.0525	0.0470	0.0608	0.0388	0.0412	

hAT-HT3

	Ac
Schmidtea mediterranea (Sm)	0.0027

SPIN

	Xt	*Ac*	*Og*	*Et*	*Ml*	*Md*
Xenopus tropicalis (Xt)						
Anolis carolinensis (Ac)	0.1198					
Otolemur garnetii (Og)	0.0366	0.0862				
Echinops telfairi (Et)	0.0304	0.0987	0.0111			
Myotis lucifugus (Ml)	0.0304	0.0931	0.0067	0.0044		
Monodelphis domestica (Md)	0.2537	0.2688	0.2323	0.2425	0.2370	

dn/ds

hAT-HT1

	Md	*Ml*	*Ac*
Monodelphis domestica (Md)			
Myotis lucifugus (Ml)	0.3276		
Anolis carolinensis (Ac)	0.4986	0.2912	

hAT-HT2

	Ts	*Ml*	*Mm*	*Md*	*Og*	*Et*	*Ac*
Tarsius syrichta (Ts)							
Myotis lucifugus (Ml)	**1.4849**						
Microcebus murinus (Mm)	**1.4537**	**1.4205**					
Monodelphis domestica (Md)	0.8280	0.8763	0.9509				
Otolemur garnetii (Og)	**1.1170**	**1.7670**	**1.3918**	0.5476			
Echinops telfairi (Et)	**1.4412**	**1.4934**	**1.8145**	0.5994	0.9559		
Anolis carolinensis (Ac)	0.5089	0.6914	0.6553	0.4227	0.4613	0.3277	

hAT-HT3

	Ac
Schmidtea mediterranea (Sm)	**3.6360**

SPIN

	Xt	*Ac*	*Og*	*Et*	*Ml*	*Md*
Xenopus tropicalis (Xt)						
Anolis carolinensis (Ac)	0.8369					
Otolemur garnetii (Og)	**1.5848**	0.6138				
Echinops telfairi (Et)	**1.7789**	0.5201	0.9017			
Myotis lucifugus (Ml)	**1.9037**	0.5519	**1.7662**	**1.5286**		
Monodelphis domestica (Md)	0.3931	0.3542	0.2608	0.2395	0.2356	

The average divergence between *hAT-HT2* copies is 0.14% in the anole indicating that these copies have been inserted very recently in this genome. In other species, *hAT-HT2* insertions are a little older, the divergence between elements ranging from 11.74% in the planarian to 3.51% in the bat (table 5). Assuming that variations in mutation rate are small among species, differences in pairwise divergence suggest that the lateral transfer of *hAT-HT2* occurred at different times in different lineages or that the elements remained dormant in some species until they began amplifying in their host genome. *hAT-HT2* is found in low copy

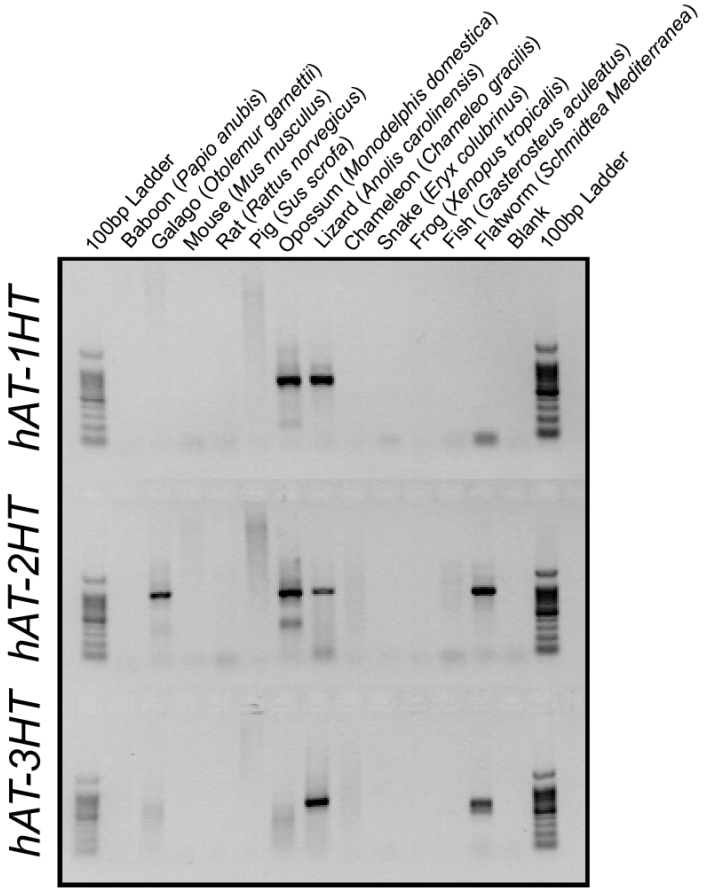

Figure 16: PCR amplification of three horizontally transferred *hAT* families. Using the same primers, *hAT-1HT* amplifies in the lizard and the opossum, *hAT-2HT* in the lemur, opossum, lizard and flatworm and *hAT-3HT* in only the lizard and the flatworm. As *hAT-2HT* elements in frog are extremely fragmented, we were unable to amplify this family. In all other organisms screened, we repeatedly failed to amplify any of the three *hAT* families.

numbers in most species (<50 copies), except in the tenrec and in the opossum where this element is found in more than 200 copies. This family is also responsible for the mobility of non-autonomous elements in most lineages (Table 5). We identified one non-autonomous family related to *hAT-HT2* in *Microcebus,* anole and tenrec, at least two non-autonomous families in the bat and at least three in the tarsier. These non-autonomous families have amplified extremely successfully and largely outnumber their autonomous progenitors.

A second autonomous *hAT* family, *hAT-HT1*, is shared only among the anole, the brown bat and the opossum (Figure 15 and 16). The *hAT-HT1* family includes the opossum *hAT1_MD* and bat *Myotis_hAT1* families described in RepBase. In each of these species we were able to recover full-length elements and to reconstruct complete consensus sequences. The divergence between the anole and opossum *hAT-HT1* is extremely low, in particular at synonymous sites (3.15%), and is suggestive of lateral transfer. The *ds* between the anole and the bat is much higher (about 30%) suggesting a distinct origin for the *hAT-HT1* family in the bat. The distributions of pairwise divergence among elements (figure 17) barely overlap indicating that the *hAT-HT1* family amplified at different time in these three species. A phylogenetic analysis of genomic copies of *hAT-HT1* elements (Figure 18) shows that the anole and opossum elements form an unresolved polytomy, suggesting a single source for the lateral transfer of *hAT-HT1* into these two species. In contrast, bat elements form a clearly distinct monophyletic group indicating that bat *hAT-HT1* might be coming from a different source or that the elements were transferred at different times. This independent origin is also supported by the observation that bat *hAT-HT1* elements have very different (non-homologous) termini than the anole and opossum elements (Figure 19). *hAT-HT1* transposases also differ remarkably at non-synonymous sites, although the values of the *dn/ds* ratio were not as high as the values for *hAT-HT2* (table 6). Autonomous *hAT-HT1* elements are found in very low copy number in the

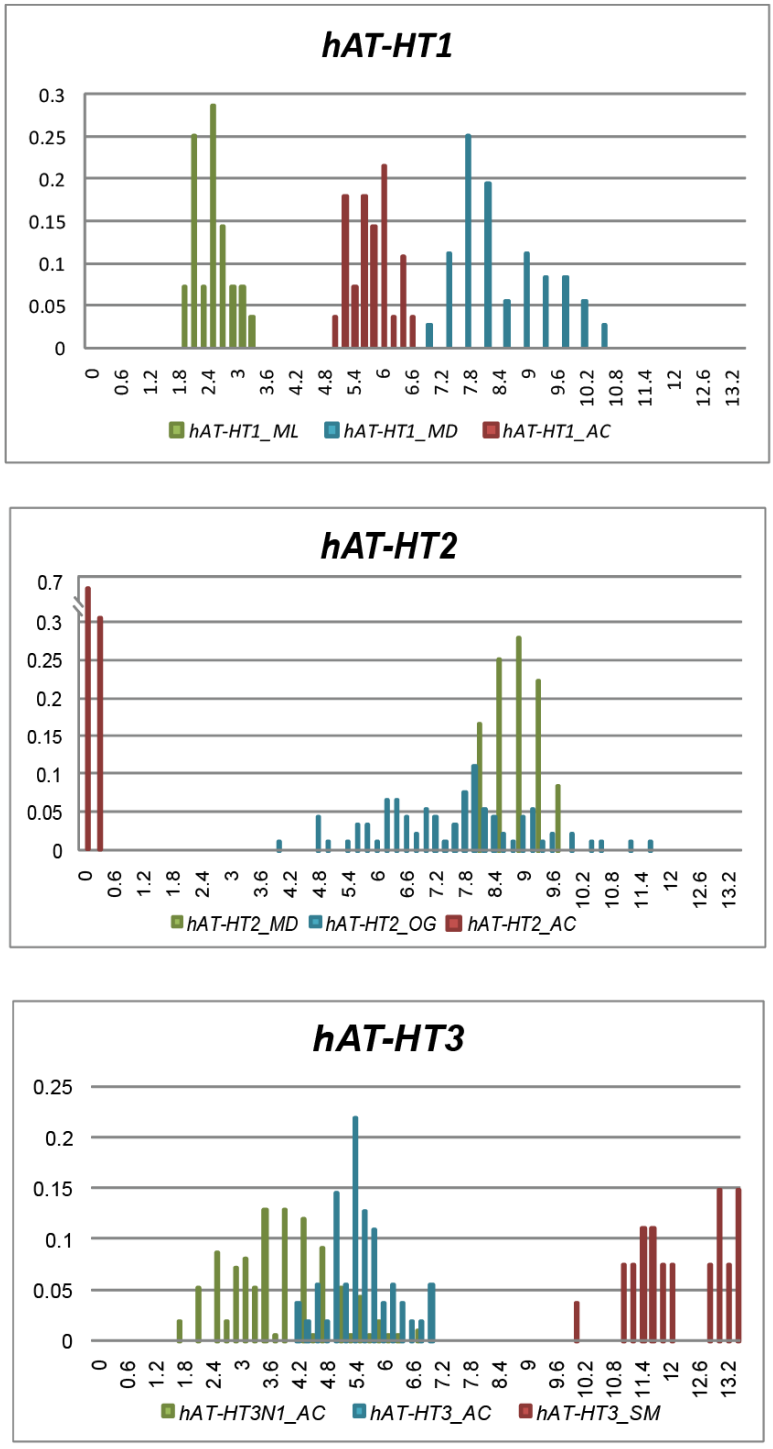

Figure 17: Pairwise divergence distribution of *hAT-HT1*, *hAT-HT2* and *hAT-HT3* families. Abbreviations: *Anolis carolinensis* (AC), *Myotis lucifugus* (ML), *Monodelphis domestica* (MD), *Otolemur garnetii* (OG) and *Schmidtea mediterranea* (SM). Genomic elements were collected and at least 1,000bp across the element were used to calculate pairwise divergence.

anole (about 6) but have been successful at colonizing the bat and opossum genomes, with more than 30 copies in the bat and more than 100 copies in the opossum.

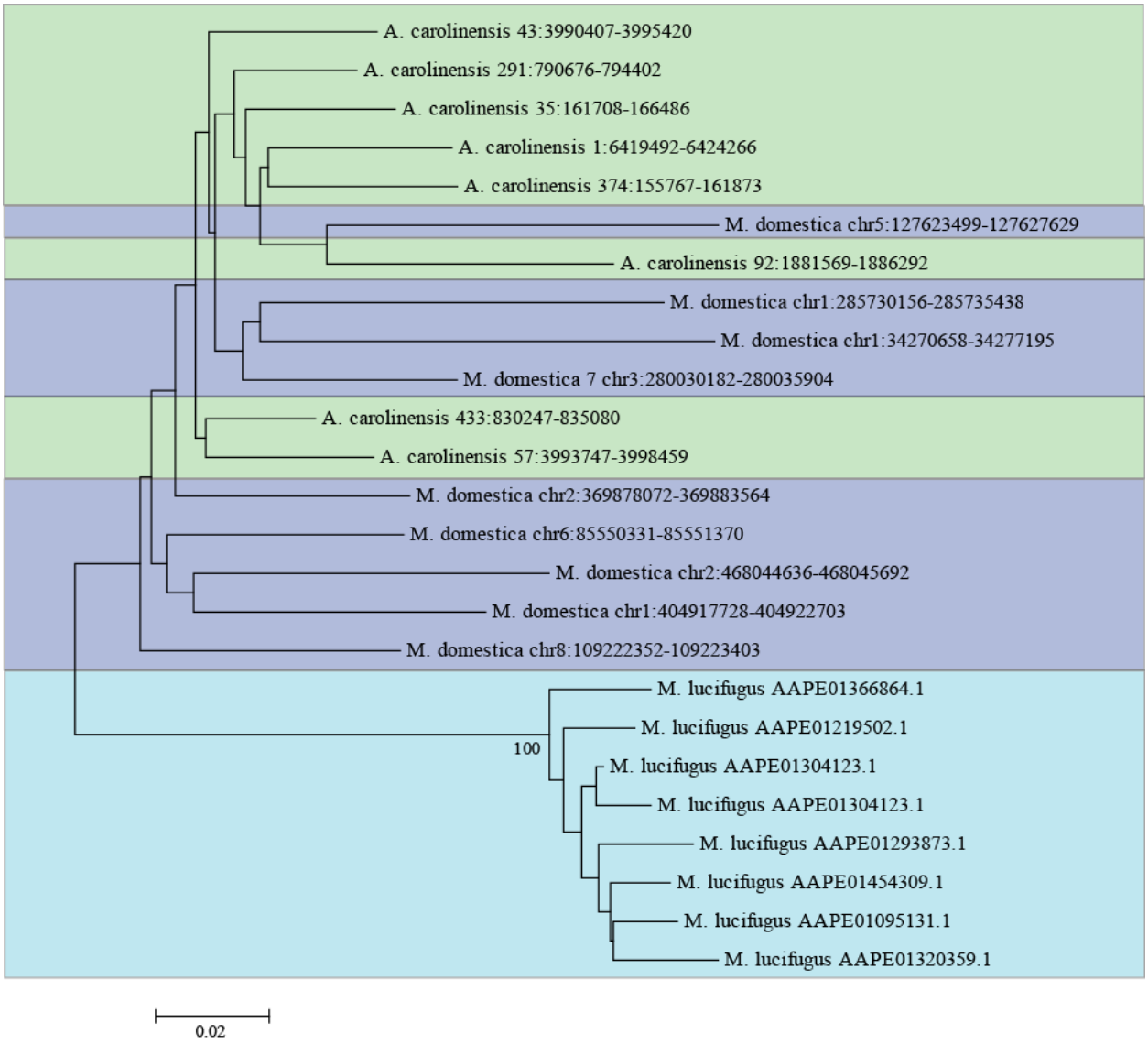

Figure 18: Neighbor joining phylogeny of genomic copies of *hAT-HT1* elements from the three species known to foster this family, *Monodelphis domestica* (blue), *Anolis carolinensis* (green) and *Myotis lucifugus* (teal). The tree is based on a 1000bp alignment of the transposase domain. The robustness of the tree was assessed with 1000 bootstrap replicates. Only bootstrap values higher than 75 are shown. Accession numbers for each element are delineated on each branch (chromosome or scaffold locus from the UCSC website for the opossum and lizard respectively, or its NCBI accession number for the bat).

A third autonomous *hAT* family, *hAT-HT3*, is shared by the anole and the planarian *S. mediterranea* (Figure 15 and 16). We recovered a number of full-length elements in the anole but only fragments in the planarian, yet we were able to construct full-length *hAT-HT3* consensus for both species. The *ds* value between the anole and the planarian consensus is the lowest reported in this study, with a value of 0.27%, suggesting a lateral transfer from the very same source or directly from a planarian species to the anole (because the *hAT-HT3* has apparently resided longer in the planarian than in the lizard).

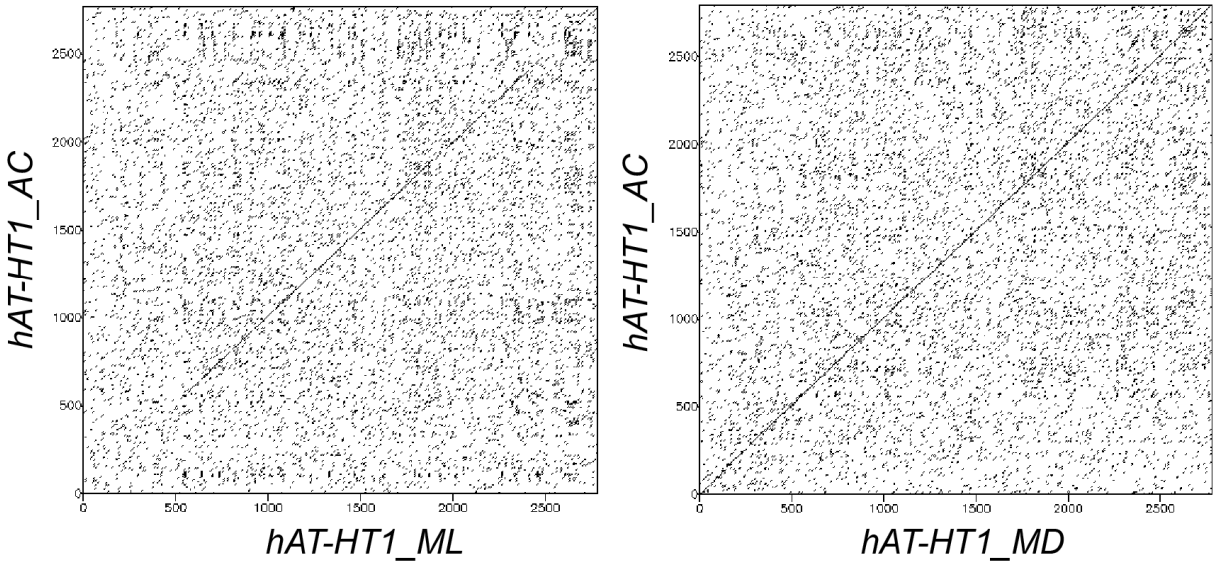

Figure 19: Dot plot comparison of the complete *hAT-HT1* consensus sequences of the anole to the little brown bat, *Myotis lucifugus* (Left), and to the gray short-tailed opossum, *Monodelphis domestica* (Right)

Finally, we found in anole several families of non-autonomous elements related to *hAT-HT3*. One of these non-autonomous elements *hAT-HT3N1*, is also found in the bat. The anole and the bat *hAT-HT3N1* elements are very similar and diverge by only 0.65%, a value best explained by a lateral transfer of this non-autonomous element. The transfer and successful amplification of the non-autonomous element is surprising because we failed to find in the bat genome an autonomous element similar to the *hAT-HT3* family. It is plausible that, following its

transfer, the *hAT-HT3N1* recruited another autonomous element for its mobilization. We cannot exclude that the number of autonomous *hAT-HT3* elements in the bat genome is very small and that they reside in genomic regions that have yet to be fully sequenced. It is also possible that the autonomous element was so recently transferred that it is still polymorphic in little brown bat populations and is missing from the individual sequenced.

Evidence for Lateral Transfer

The evidence we presented above unequivocally demonstrates that the presence in distantly related genomes of three autonomous and one non-autonomous *hAT* transposons results from lateral transfer. The first line of evidence comes from the phylogenetically discontinuous distribution of the families. If these elements had been vertically transmitted from a common ancestor, they would have had to been independently lost from other lineages. In fact, we would need to assume as many as 14, 8, 9 and 9 independent losses to account for the phylogenetic distribution of the *hAT-HT2*, *hAT-HT1*, *hAT-HT3* and *hAT-HT3N1* families, respectively. We would also expect to find the remnants of ancient elements in the primates, tenrec, bat and opossum genomes. TEs readily accumulate in mammalian genomes where they remain as DNA fossils (Gilbert and Labuda, 1999; Bejerano et. al., 2006). Given the relatively slow mutation rate of mammals, ancient elements remain detectable long after their insertion. In the case of these families, we failed to find ancient copies in any of these genomes suggesting these families are recent colonizers of mammalian genomes. This reasoning applies only to mammals because TEs, especially long ones, do not accumulate in the anole genome, which is characterized by a high turnover of inserts (Novick et al. 2009a). Thus, we do not expect to find ancient elements in this genome. The second line of evidence in favor of lateral transfer comes from the very low divergence between consensus sequences, especially at silent sites. Silent divergences faithfully reflect the evolutionary history of sequences because synonymous mutations are virtually invisible to natural selection. For all autonomous families, we report

levels of silent divergence considerably lower than expected for such ancient lineages. For instance, lophotrochozoans and vertebrates separated more than 800 MYA, yet the *ds* between the anole and the planarian *hAT-HT3* consensus is 0.27%. Similarly among mammals, the *ds* values between *hAT-HT2* consensus sequences were all lower than 5%, although chiropters diverged from primates 65 MY ago and eutherians split from marsupials more than 120 MY ago (Donoghue and Benton 2007). The third line of evidence comes from phylogenetic analyses. Phylogenetic trees showed a striking lack of structure, for *hAT-HT2* and for part of the *hAT-HT1* tree. This lack of structure is not consistent with a vertical mode of transmission which yields bifurcating phylogenies that mimic the phylogeny of the host species. The phylogenies we obtained are best explained by the invasion of different genomes from a single or a small number of sources.

Discussion

We identified four new families of DNA transposons that have been horizontally transferred in tetrapods. Because these families represent distinct monophyletic groups (Figure 20), they signify independent events of lateral transfer. The discovery of these four cases results from searching other genomes for the presence of transposons discovered in *A. carolinensis*. It is therefore biased towards elements found in this species and certainly underestimates the frequency of transposon lateral transfer in tetrapods. The only case previously reported was the *SPIN* family, which was shared between bushbaby, bats, murine rodents, marsupials, anole and frog (Pace et al. 2008). Thus, at least five events of lateral transfer have occurred during the evolution of tetrapods and we suspect many more will be discovered. Lateral transfer of transposons had previously been described in animals but, in most cases, was between closely related species and was mostly limited to insects. Our analysis suggests that the transfer of DNA transposons might be much more common in vertebrates than previously thought. In fact,

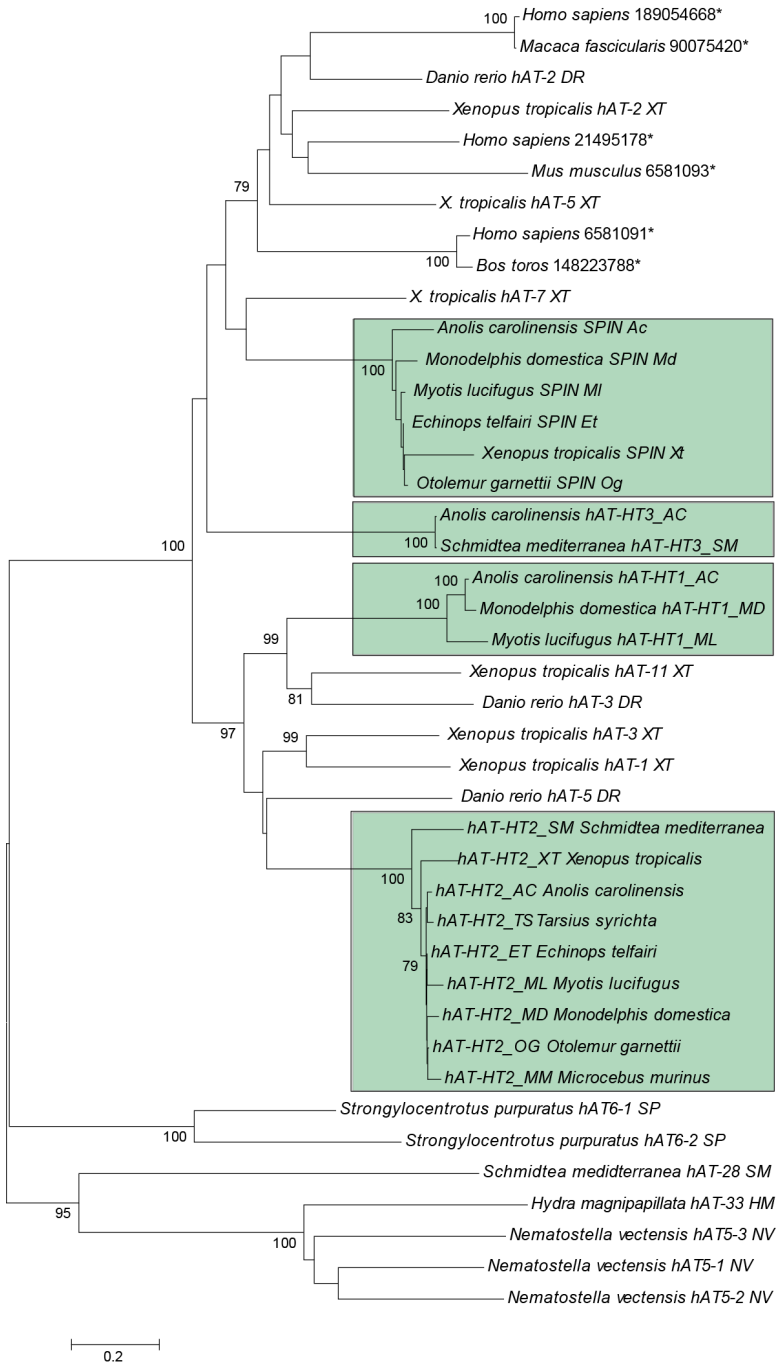

Figure 20: Neighbor joining phylogeny of 35 sequences consisting of 600 amino-acids of the 3' end of the *hAT* transposase domain in vertebrates. In addition to the 14 sequences included in this study (shaded in green), six mammalian sequences were included from the NCBI database (protein accession numbers are available next to the scientific name and indicated by an asterisk) as well as 15 sequences retrieved from repbase (SPIN elements are also shaded in green). As an outgroup we have included seven sequences from four species of invertebrates including *S. mediterranea* (also retrieved from repbase). Bootstrap values less than 75 have been removed from the tree.

it probably happens more often than our data suggests because we detected only the successful events of lateral transfer.

This has important implications for the genomic evolution of mammals. Until the discovery of recent transposon activity in the bat *M. lucifugus* (Ray et al. 2007; Ray et al. 2008), it was believed that DNA transposons were extinct in mammals and had no impact on the evolution of this group. However, lateral transfer provides a means for the re-colonization of genomes and we cannot exclude that DNA transposons are active in many species for which genomic data are not currently available. The successful amplification of laterally transferred DNA transposons can have a significant impact on genomic diversity. Altogether the five events of lateral transfer discussed here are responsible for the amplification of approximately 1300, 57,000 and 100,000 elements in the anole, the bat and the tenrec genomes respectively. As the amplification of DNA transposons can trigger the evolution of new genes or even new biological functions by providing coding and regulatory motifs on which natural selection can act (reviewed in Feschotte and Pritham, 2007), it is plausible that some insertions have been recruited by their host to perform a new function.

The evidence in favor of lateral transfer is overwhelming, yet our data provide no information about the source of the transfer and its mechanism(s). The main problem is that in amniotes (reptiles and mammals) the germ line and the early stages of embryogenesis are sequestered within the gonads and the genital tract of the female. In species that have free living unprotected eggs and/or external fertilization, such as the frog and the planarian, it is easier to imagine a mechanism of transfer of genetic material directly in the germ line. In amniotes, the lateral transfer of transposons implies the existence of an agent that is mediating the transfer to the different species. It was recently proposed that poxviruses could mediate lateral transfer of TEs because a snake retroposon was found in the genome of a rodent

poxvirus (Piskurek and Okada 2007), raising the possibility that this virus could transfer the reptilian element to its rodent host. Poxviruses are good candidates for mediating lateral transfer due to their low host and cell specificity. However, we are not aware of any poxvirus (or any other viruses) with a range of hosts broad enough to explain the phylogenetic distribution of DNA transposons, from planarian to primates. The geographic distributions and ecology of the host species differ drastically: tarsiers are found in the Philippines, the bush baby and *Xenopus* frog in Africa, lemurs and tenrecs in Madagascar, the little brown bat and the green anole in North America, the opossum *M. domestica* in South America and the planarian *S. mediterranea* around the Mediterranean. Not only do these species inhabit different geographic areas, but their center of origin and their evolution took place on different continents. Therefore, the source of the transposons must have (or had if it is now extinct) a very wide, possibly global, geographic distribution. A majority of the species that experienced lateral transfer have an insectivorous diet (anole, bats). It is therefore tempting to speculate that a transfer of transposons occurred from prey to predators. However, none of the laterally transferred elements described here were found in the insect genomes currently available, although we cannot exclude that these transposons are present in an insect species that has not yet been sequenced.

Although the phylogenetic distribution of the five transposon families is discontinuous, it is not random. For instance, the anole shares four families with the bat, three with the opossum and three with the planarian *S. mediterranea* but none with the other 24 vertebrates and more than 25 invertebrates for which genomic data are available. This pattern suggests that some genomes have a higher chance of lateral transfer or are more hospitable to transposons than others or both. It is plausible that genomes that are vulnerable to lateral transfer have a weakened resistance to transposon invasion and, at first, lose control of the amplification whereas genomes with a strong response never let the transposons amplify in their genome.

Our data suggest that a mechanism of resistance might exist in some species. When consensus transposase sequences are compared, we found high values for the ratio *dn/ds*, with a number of values higher than 1. Values of *dn/ds* higher than 1 indicate that selection in favor of amino-acid changes is acting on a gene. One can wonder why a gene as indispensable to the fitness of the transposon would evolve rapidly at the amino-acid level. A possibility is that the amplification of the transposon could trigger the evolution of a response by the host, leading to an arms race between the transposon and its host. This evolutionary race could cause a rapid evolution of the transposase, and of the host factor responsible for its control. The role of a genomic conflict has previously been proposed to explain the rapid evolution of the first open-reading frame of the LINE-1 retrotransposon in primates (Boissinot and Furano 2001; Khan et al. 2006) and of transposase genes in bacterial DNA transposons (Petersen et al. 2007).

CONCLUSIONS

Summary and Future Research

The anole genome is characterized by an extraordinary diversity of Class I and Class II transposable elements. This diversity results from the persistence of vertically inherited families of non-LTR retrotransposons and DNA transposons since the origin of amniotes and from the frequent lateral transfer of DNA transposons. It is notable that the vast majority of TEs in the anole show very recent sign of activity or are still active. The young age of most TEs insertions and the rarity of older elements suggest that most new insertions do not reach fixation, possibly because TE insertions are strongly deleterious in reptiles. However, it seems that some categories of elements, in particular shorter elements, do accumulate. These fixed elements decay very rapidly suggesting a high rate of DNA loss in reptiles. The constant generation of novel insertions, their selective loss as well as the high rate of DNA deletion suggests that the anole genome is very flexible and has probably evolved mechanisms to tolerate this high level of activity.

The diversity of young families in reptiles is unparalleled in amniotes and clearly exceeds the diversity of TEs in mammals, birds and possibly amphibians (Lander et al. 2001; Shedlock et al. 2007; Warren et al. 2008; Waterston et al. 2002). In fact, the only genomes that resemble the anole genome are fish and insect genomes, where a large diversity of young class I and II families coexists (Duvernell, Pryor, and Adams 2004; Kaminker et al. 2002;.Krasnov et al. 2005). Similarly to the anole, the vast majority of insertions in fish and insects are very young, suggestive of the high turnover of elements in these species. It is still unclear why vertebrate genomes differ so much in terms of TE diversity and abundance. Eickbush and Furano (2002) proposed that the rate of ectopic recombination in insects and lower vertebrates was higher than in mammals. Assuming that inter-element recombination is the main explanation for TE's deleteriousness, this hypothesis suggests that TE insertions are more deleterious in insects and fish resulting in their selective loss and their failure to accumulate in these groups. Three lines of evidence in anoles seem to confirm the idea that ectopic recombination is more common in

reptiles than in mammals, accounting for the profile of TE diversity in this group. First, we found that the only elements that accumulate were the shortest one, possibly because short elements are less likely to mediate ectopic recombination events (Song and Boissinot 2007). Second, inter-element recombination seems to be common in anoles and has driven the diversification of entire families of class II elements. Finally, the rapid decay of fixed insertions resulting from the loss or translocation of large fractions of elements could very well be caused by inter-element recombination. A similar recombination-mediated decay of TE was reported in plants and accounts for the loss of DNA in monocots (Bennetzen, Ma, and Devos 2005).

Our work raised a number of questions that remain to be answered. First, a higher deleterious effect of TE insertions in reptiles than in mammals is at this point only a hypothesis that remains to be tested. A population genetics similar to the ones Petrov et al. (2003) and Boissinot et al. (2006) performed on *Drosophila* and humans respectively would answer this question. The second hypothesis to be confirmed is the one relative to the difference of DNA loss among vertebrates. Studies in insects have shown that variations in the rate of DNA loss even among closely related species can occur and accounts for differences in genome size (Petrov and Hartl 1998). However, similar comparative studies have not been conducted in vertebrates possibly because the field of vertebrate comparative genomics is relatively new. Finally, a possible relationship between TE activity and adaptation in anoles needs to be examined. Anoles constitute the most diverse group of amniotes. They have undergone multiple adaptive radiations and their ecological and morphological diversity is unparalleled in a vertebrate genus (Losos et al. 1998). It has been proposed that TEs can impact the rate of speciation of organisms (Furano 2000) or generate important adaptive genetic changes by exaptation (Reviewed in Sorek 2007; Feschotte and Pritham 2007). However, this issue has not been rigorously examined in a group of organisms as diverse and speciose as the *Anolis*

genus. Clearly, expanding the analysis of TEs diversity to other anole species could possibly reveal new mechanisms by which TEs can impact the genome of their host.

BIBLIOGRAPHY

Abrusan, G., H. J. Krambeck, T. Junier, J. Giordano, and P. E. Warburton. 2008. Biased distributions and decay of long interspersed nuclear elements in the chicken genome. Genetics **178**:573-581.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.

Andersson, J.O. Lateral gene transfer in eukaryotes. 2005. Cellular and Molecular Life Sciences **62**: 1182-1197.

Bao, W., M.G. Jurka, V. V. Kapitonov, and J. Jurka. 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol Biol Evol **5**: 583-593.

Basta, H. A., A. J. Buzak, and M. A. McClure. 2007. Identification of novel retroid agents in Danio rerio, Oryzias latipes, Gasterosteus aculeatus and Tetraodon nigroviridis. Evol Bioinform:179-195.

Baxter, L.R., R.F. Ackermann, E.C. Clark and J.E. Baxter. 2001. Brain mediation of *Anolis* social dominance displays: I. Differential basal ganglia activation. Brain, Behavior and Evolution **57**:169-184.

Bejerano, G., C.B. Lowe, N. Ahituv, B. King, A. Siepel, S.R. Salama, E.M. Rubin, W.J. Kent, and D. Haussler. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature **441**:87-90.

Ben-Aroya, S., P. A. Mieczkowski, T. D. Petes, and M. Kupiec. 2004. The compact chromatin structure of a Ty repeated sequence suppresses recombination hotspot activity in *Saccharomyces cerevisiae*. Molecular Cell **15**:221-231.

Bennetzen, J.L., J. Ma, and K.M. Devos. 2005. Mechanisms of recent genome size variation in flowering plants. Ann Bot **95**:127–132.

Bergthorsson, U., K.L. Adams, B. Thomason,and J.D. Palmer. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature **424**:197-201.

Boissinot, S., P. Chevret, and A. V. Furano. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol **17**:915-928.

Boissinot, S., J. Davis, A. Entezam, D. Petrov, and A.V. Furano. 2006. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A **103**: 9590-9594.

Boissinot, S., A. Entezam, and A. V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol **18**:926-935.

Boissinot, S. and A.V. Furano. 2001. Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol **18**:2186-94.

Boissinot, S., and M. Song. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene **390**:206-213.

Burke, W. D., H. S. Malik, J. P. Jones, and T. H. Eickbush. 1999. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. Mol Biol Evol **16**:502-511.

Cabot, E.L., B. Angeletti, K. Usdin, and A.V. Furano. 1997. Rapid evolution of a young L1 (LINE-1) clade in recently speciated Rattus taxa. J Mol Evol **45**:412-423.

Callinan, P.A., and M.A. Batzer. 2006. Retrotransposable elements and human disease. Genome Dyn **1**:104-15.

Casola, C., A.M. Lawing, E. Betrán, C. Feschotte. 2007. PIF-like transposons are common in drosophila and have been repeatedly domesticated to generate new host genes. Mol Biol Evol **24**:1872-1888.

Charlesworth, B., and D. Charlesworth. 1983. The population dynamics of transposable elements. Genetical Research **42**:1-27.

Charlesworth, B., and C. H. Langley. 1989. The population genetics of *Drosophila* transposable elements. Annual Review of Genetics **23**:251-287.

Chen, J.M., P.D. Stenson, D.N. Cooper, and C. Ferec. 2005. A systematic analysis of LINE-1 - dependent retrotranspositional events causing human genetic disease. Hum Genet **117:**411-427.

Cooper, D. M., K. J. Schimenti, and J. C. Schimenti. 1998. Factors affecting ectopic gene conversion in mice. Mammalian Genome **9**:355-360.

Cost, G. J., Q. Feng, A. Jacquier, and J. D. Boeke. 2002. Human L1 element target-primed reverse transcription in vitro. EMBO J **21**:5899-5910.

Craig, N.L., R. Craigie, M. Gellert, and L.M. Lambowitz. 2002. Mobile DNA II. Washington, DC: Am. Soc. Microbiol. Press.

Daniels, S.B., K.R. Peterson, L.D. Strausbaugh, M.G. Kidwell, and A. Chovnick. 1990. Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics **124**:339-55.

de Boer, J.G., R. Yazawa, W.S. Davidson, and B.F. Koop. 2007. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics **8**:422.

Dewannieux, M., C. Esnault, and T. Heidmann. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet **35**:41-48.

Dewannieux, M., and T. Heidmann. 2005. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. J Mol Biol **349**:241-247.

Donoghue, P.C.J. and M.J. Benton. 2007. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. Trends in Ecology and Evolution **22**:424-431.

Duvernell, D. D., S. R. Pryor, and S. M. Adams. 2004. 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 **59**:298-308.

Edgar, R.C. and E.W. Myers. 2005. PILER: identification and classification of genomic repeats. Bioinformatics **21**:i152-8.

Eickbush, T. H., and A. V. Furano. 2002. Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev **12**:669-674.

Engels, W.R., D.M. Johnson-Schlitz, W.B. Eggleston, and J. Sved. 1990. High-frequency P element loss in *Drosophila* is homolog dependent. Cell **62**:515–25.

Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. Nature Genetics. **24**:363-367.

Feng, Q., J. V. Moran, H. H. Kazazian, Jr., and J. D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87**:905-916.

Feschotte, C. 2008 Transposable elements and the evolution of regulatory networks. Nat Rev Genet **9**:397-405.

Feschotte, C. and E.J. Pritham, E.J. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331-68.

Feschotte, C., L. Swamy and S.R. Wessler. 2003. Genome-wide analysis of mariner-like transposable elements reveals complex relationships with Stowaway MITEs. Genetics **143**: 747-758.

Feschotte, C., X. Zhang and S.R. Wessler. 2002. Miniature Inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In: Mobile DNA II (Craig N, Craigie R, Gellert M & Lambowitz A, eds.). ASM Press, Washington D.C.

Furano, A. V. 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol **64**:255-294.

Furano, A. V., D. D. Duvernell, and S. Boissinot. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet **20**:9-14.

Gentles, A. J., M. J. Wakefield, O. Kohany, W. Gu, M. A. Batzer, D. D. Pollock, and J. Jurka. 2007. Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. Genome Res **17**:992-1004.

Gibbs, R.A., G.M. Weinstock, M.L. Metzker, D.M. Muzny, and E.J. Sodergren. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Gilbert, N., and D. Labuda. 1999. CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. Proc Natl Acad Sci USA **96**:2869-2874.

Gish, W. 1996-2004. WU-BLAST.

Gregory, T.R. 2009. Animal Genome Size Database. http://www.genomesize.com.

Guindon, S., P. Lethiec, P. Duroux, and O. Gascuel. 2005. PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res **33**:W557-559.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series **41**:95-98.

Hartl, D.L., E.R. Lozovskaya and J.G. Lawrence. 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica **86**:47-53.

Hsia, A.P., and P.S. Schnable. 1996. DNA sequence analyses support the role of interrupted gap repair in the origin of internal deletions of the maize transposon, *MuDR*. Genetics **142**:603–18.

Hull, R. 2001. Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. International Committee on Taxonomy of Viruses. Arch Virol **146**:2255-2261.

Jiang, N., Z. Bao, X. Zhang, S.R. Eddy, and S.R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature **431**:569–73.

Jiao, Y., O.S. Lau, and X.W. Deng. **2007.** Light-regulated transcriptional networks in higher plants. Nat Rev Genet **8**:217-30.

Jones, J.M. and M. Gellert. 2004. The taming of a transposon: V(D)J recombination and the immune system. Immunol Rev **200**:233-48.

Jordan, I.K., I.B. Rogozin, G.V. Glazko, and F.V. Koonin. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet **19**:68-72.

Jurka, J. and E. Zuckerkandl. 1991. Free left arms as precursor molecules in the evolution of Alu sequences. J Mol Evol **33**:49-56.

Kajikawa, M., and N. Okada. 2002. LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell **111**:433-44.

Kaminker J.S., C.M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biology **3**:1-20.

Kapitonov, V.V., and J. Jurka. 2001. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA **98**:8714–8719.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. 2003. The UCSC Genome Browser Database. Nucleic Acids Res **31**:51-54.

Kawasaki, S., and E. Nitasaka. 2004. Characterization of *Tpn1* family in the Japanese morning glory: *En/Spm*-related transposable elements capturing host genes. Plant Cell Physiol **45**:933–44.

Kazazian, H.H. 2004. Mobile elements: Drivers of Genome Evolution. Science **303**:1626-1632.

Khan, H., A. Smit, and S. Boissinot. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res **16**:78-87.

Kidwell, M.G., and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci USA **94**:7704–11.

Kimura, M. A. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16**:111-120.

Kohlstaedt, L. A., J. Wang, J. M. Friedman, P. A. Rice, and T. A. Steitz. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science **256**:1783-1790.

Kordis, D., N. Lovsin, and F. Gubensek. 2006. Phylogenomic analysis of the L1 retrotransposons in Deuterostomia. Syst Biol **55**:886-901.

Kordis, D. and F. Gubensek. 1998. Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. Proc Natl Acad Sci U S A **95**:10704-10709.

Krasnov, A., H. Koskinen, S. Afanasyev, and H. Mölsä. 2005. Transcribed Tc1-like transposons in salmonid fish. BMC Genomics **12**:107-117.

Kumar, S., K. Tamura, I.B. Jakobsen, and M. Nei. 2001. MEGA2: Molecular Evolutionary Genetics Analysis Software, Arizona State University, Tempe, Arizona, USA.

Kunze, R.,and C.F. Weil. 2002. The *hAT* and CACTA superfamilies of plant transposons. Mobile DNA II. Washington, DC: Am. Soc. Microbiol. Press 565–610.

Lander, E. S.L. M. Linton, B. Birren, C. et. al. (254 co-authors) 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Langley, C. H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth. 1988. On the role of unequal exchange in the containment of transposable element copy number. Genetical Research **52**:223-235.

Le Rouzic, A. and P. Capy. 2006. Population genetics models of competition between

transposable element subfamilies. Genetics **174**:785–93.

Lindblad-Toh, K., C.M. Wade, T.S. Mikkelsen, E.K. Karlsson, D.B. Jaffe, et. al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438**:803-819.

Lisch, D. 2002. Mutator transposons. Trends Plant Sci*.* **7**:498–504.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res **13**:358-68.

Losos, J. B., T.R. Jackman, A. Larson, K. de Queiroz, and L. Rodriguez-Schettino. 1998. Contingency and determinism in replicated adaptive radiations of island lizards. Science **279:** 2115-2118.

Lovern, M. B., M. M. Holmes, and J. Wade. 2004. The green anole, *Anolis carolinensis*: a reptilian model for laboratory studies of reproductive morphology and behavior. Journal of the Institute for Laboratory Animal Research **45**:54-64.

Lovsin, N., F. Gubensek, and D. Kordis, D. 2001. Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in deuterostomia. Mol Biol Evol **18**:2213-2224.

Lowe, C.B., G. Bejerano, and D. Haussler. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A **104**:8005-8010.

Luan, D. D., and T. H. Eickbush. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol Cell Biol **15**:3882-3891.

Luan, D. D., M. H. Korman, J. L. Jakubczak, and T. H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell **72**:595-605.

Lunyak, V.V., G.G. Prefontaine, E. Núñez, T. Cramer, B.G. Ju, et. al. 2007. Developmentally

regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science **317**:248-251.

Makalowski, W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene **259**:61-67.

Malik, H. S., W. D. Burke, and T. H. Eickbush. 1999. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol **16**:793-805.

Malik, H.S., and T.H. Eickbush. 1998. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol Biol Evol **15**:1123-1134.

Malone, J. H. and W. Davis. 1998. Geographic distribution of Anolis carolinensis. Herpetological Review **29**: 248.

Marchler-Bauer, A., J. B. Anderson, M. K. Derbyshire, et. al. (25 co-authors) 2007. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res **35**:D237-240.

Martin, S.L., and F.D. Bushman. 2001 Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol **21**:467-475.

Martin, S.L., W. Li, A.V. Furano, and S. Boissinot. 2005 The structures of mouse and human L1 elements reflect their insertion mechanism. Cytogenet Genome Res **110**:223-228.

Martin, S.L., J. Li, and J.A. Weisz. 2000. Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. J Mol Biol **304**:11–20.

Martin, F., C. Maranon, M. Olivares, C. Alonso, and M. C. Lopez. 1995. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes. J Mol Biol **247**:49-59.

Mathews, S. 2006. Phytochrome-mediated development in land plants: red light sensing evolves to meet the challenges of changing light environments. Mol Ecol **15**:3483-503.

Mathias, S.L., A.F. Scott, H.H. Kazazian Jr., J.D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. Science **254**:1808-1810.

McClintock, B. 1951. Chromosome organization and genic expression. Cold Spring Harbor Symp Quant Biol **16**:13–47.

McClure, M. A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol **8**:835-856.

McClure, M. A., M. S. Johnson, D. F. Feng, and R. F. Doolittle. 1988. Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. Proc Natl Acad Sci U S A **85**:2469-2473.

McClure, M. A., H. S. Richardson, R. A. Clinton, C. M. Hepp, B. A. Crowther, and E. F. Donaldson. 2005. Automated characterization of potentially active retroid agents in the human genome. Genomics **85**:512-523.

Mikkelsen, T. S., M. J. Wakefield, B. Aken, et. al. (62 co-authors) 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature **447**:167-177.

Miskey, C., Z. Izsvák, K. Kawakami, and Z. Ivics. 2005. DNA transposons in vertebrate functional genomics. Cell Mol Life Sci **62**:629-641.

Monks, S.P. 1881. A partial biography of the green lizard. Am Nat **15**:96-99.

Muehlbauer, G.J., B.S. Bhau, N.H. Syed, S. Heinen, S. Cho, et. al. 2006. A hAT superfamily transposase recruited by the cereal grass genome. Mol Genet Genomics **275**:553-563.

Neafsey, D. E., Blumenstiel, J. P. and Hartl D. L. 2004. Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies. Mol Biol Evol:2310-2318.

Nekrutenko, A., and W. H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet **17**:619-621.

Novick, P.A., H. Basta, M. Floumanhaft, M.A. McClure, and S. Boissinot. 2009a. The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis carolinensis* shows more similarity to fish than mammals. Mol Biol Evol **26**: 1811-1822.

Novick, P.A., J. Smith, D. Ray, and S. Boissinot. 2009b. Independent and parallel lateral transfer of DNA transposons in tetrapod genomes. Gene **449**:85-94.

Ohshima, K., and N. Okada. 2005. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res **110**:475-90.

Ostertag, E.M. and H.H. Kazazian. 2001. Biology of mammalian L1 retrotransposon. Ann Rev Genet **35**:501-538.

Pace, J.K. 2nd, and C. Feschotte. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res **17:**422-432.

Pace, J.K., C. Gilbert, M.S. Clark, and C. Feschotte. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci U S A **105**:17023-17028.

Pascale, E., C. Liu, E. Valle, K. Usdin, and A. V. Furano. 1993. The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. J Mol Evol **36**:9-20.

Petersen, L., J.P. Bollback, M. Dimmic, M. Hubisz, and R. Nielsen. 2007. Genes under positive selection in Escherichia coli. Genome Res **17**:1336-43.

Petrov, D., Y.T. Aminetzach, J.C. Davis, D. Bensasson, and A.E. Hirsh. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol **20**:880-892.

Petrov, D.A. and D.L. Hartl. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol **15**: 293-302.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal, et al. 2006. Doubling genome size without

polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res **16**:1262–1269.

Piskurek, O., C.C. Austin, and N. Okada. 2006. Sauria SINEs: Novel short interspersed retroposable elements that are widespread in reptile genomes. Journal of molecular evolution **62**:630-644.

Piskurek, O. and N. Okada. 2007. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proc Natl Acad Sci U S A **104**:12046-12051.

Posada, D. and K.A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817-818.

Price A.L., Jones N.C. and Pevzner P.A. 2005. De novo identification of repeat families in large genomes. To appear in Proceedings of the 13 Annual International conference on Intelligent Systems for Molecular Biology (ISMB-05). Detroit, Michigan.

Pritham EJ, Feschotte C, Wessler SR. Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. Mol Biol Evol. 2005;22:1751–1763.

Pritham, E.J., T. Putliwala, and C. Feschotte. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390:3-17.

Ray, D.A., C. Feschotte, H.J. Pagan, J.D. Smith, E.J. Pritham, P. Arensburger, P.W. Atkinson, and N.L. Craig. 2008. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res **18**:717-728.

Ray, D.A., H.J. Pagan, M.L. Thompson, and R.D. Stevens. 2007. Bats with hATs: evidence for recent DNA transposon activity in genus Myotis. Mol Biol Evol **24**:632-639.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16**:276-277.

Robertson, H.M. 2002. Evolution of DNA transposons. Mobile DNA II. Washington, DC:

Am. Soc. Microbiol. Press. pp. 1093–110.

Robertson, H.M. and D.J. Lampe. 1995. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Mol Biol Evol* **12:**850-862.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406-425.

Shedlock, A. M., C. W. Botka, S. Zhao, J. Shetty, T. Zhang, J. S. Liu, P. J. Deschavanne, and S. V. Edwards. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. Proc Natl Acad Sci U S A **104**:2767-2772.

Skowronski, J., T. G. Fanning, and M. F. Singer. 1988. Unit-length line-1 transcripts in human teratocarcinoma cells. Molecular and Cellular Biology **8**:1385-1397.

Smit, A. F. A., Hubley R., Green P,. 2004. RepeatMasker Open-3.0. Institute of Systems Biology.

Song, M., and S. Boissinot. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene **390**:206-213.

Sorek R. 2007. The birth of new exons: mechanisms and evolutionary consequences. RNA **13**:1603–1608.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**:1596-1599.

Timmis, J.N., M.A. Ayliffe, C.Y. Huang, and W. Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet **5**:123-135.

Underwood, H., and M. Calaban. 1987. Pineal melatonin rhythms in the lizard *Anolis carolinensis*: II. Photoreceptive inputs. J. Biol. Rhythms **2**:195-206.

Volff, J. N., L. Bouneau, C. Ozouf-Costaz, and C. Fischer. 2003. Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet **19**:674-678.

Volff, J. N., C. Korting, and M. Schartl. 2000. Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol Biol Evol **17**:1673-1684.

Voliva, C. F., C. L. Jahn, M. B. Comer, C. A. Hutchison, 3rd, and M. H. Edgell. 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res **11**:8847-8859.

Warren, W. C.L. W. Hillier, J. A. Marshall et. al. (104 co-authors) 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature **453**:175-183.

Waterston, R. H.K. Lindblad-Toh, E. Birney, et. al. (222 co-authors) 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Won, H. and S.S. Renner. 2003. Horizontal gene transfer from flowering plants to Gnetum. *Proc* Natl Acad Sci U S A **100**:10824-10829.

Xia, X., and Z. Xie. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered **92**:371-373.

Yang, Z. 2000 PAML (phylogenetic analysis by maximum likelihood) version 3.0. University College, London.

Yang, Z. and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**:32-43.

Zabala, G., and L.O. Vodkin. 2005. The wp mutation of *Glycine max* carries a gene-fragmentrich transposon of the CACTA superfamily. Plant Cell **17**:2619–2632.

Zdobnov, E.M., M. Campillos, E.D. Harrington, D. Torrents, and P. Bork. 2005. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res. **16**:946-954.

Zupunski, V., F. Gubensek, and D. Kordis. 2001. Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons. Mol Biol Evol **18**:1849-1863.