Genomic Diversity and Abundance of LINE Retrotransposons in 4 Anole Lizards

by

Catherine May

A Thesis Presented in Partial Fulfillment of the Requirements for the Degree Master of Science

Approved November 2013 by the Graduate Supervisory Committee:

Kenro Kusumi, Chair
J. Alan Rawls
Jürgen Gadau

ARIZONA STATE UNIVERSITY

December 2013
ABSTRACT

Vertebrate genomes demonstrate a remarkable range of sizes from 0.3 to 133 gigabase pairs. The proliferation of repeat elements are a major genomic expansion. In particular, long interspersed nuclear elements (LINES) are autonomous retrotransposons that have the ability to “cut and paste” themselves into a host genome through a mechanism called target-primed reverse transcription. LINES have been called “junk DNA,” “viral DNA,” and “selfish” DNA, and were once thought to be parasitic elements. However, LINES, which diversified before the emergence of many early vertebrates, has strongly shaped the evolution of eukaryotic genomes.

This thesis will evaluate LINE abundance, diversity and activity in four anole lizards. An intrageneric analysis will be conducted using comparative phylogenetics and bioinformatics. Comparisons within the Anolis genus, which derives from a single lineage of an adaptive radiation, will be conducted to explore the relationship between LINE retrotransposon activity and causal changes in genomic size and composition.
I’d like to acknowledge the staff of the Kusumi Lab for their efforts in creating next-generation whole genome assemblies for three anole lizards, which allowed me to conduct a intrageneric comparison of repeat elements. I’d like to give a special thanks to Dr. Marc Tollis for his help in learning phylogenetics and for all his patience during our collaboration on our genomic projects.

I was supported by the Dean’s Fellowship of the Graduate College at Arizona State University during my graduate research. I also received support from the School of Life Sciences Undergraduate Research Program and Graduate Assistantships during my research on retrotransposons.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>---------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Vertebrate Studies</td>
<td>6</td>
</tr>
<tr>
<td>Anolis Clade</td>
<td>8</td>
</tr>
<tr>
<td>Genome Assembly</td>
<td>13</td>
</tr>
<tr>
<td>2 LINES IN FOUR ANOLE LIZARDS</td>
<td>16</td>
</tr>
<tr>
<td>Introduction</td>
<td>16</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>16</td>
</tr>
<tr>
<td>Methods</td>
<td>17</td>
</tr>
<tr>
<td>Results</td>
<td>25</td>
</tr>
<tr>
<td>Anolis apletohallus</td>
<td>25</td>
</tr>
<tr>
<td>Anolis auratus</td>
<td>34</td>
</tr>
<tr>
<td>Anolis carolinensis</td>
<td>44</td>
</tr>
<tr>
<td>Anolis frenatus</td>
<td>52</td>
</tr>
<tr>
<td>Comparison of LINES in four Anole lizards</td>
<td>61</td>
</tr>
<tr>
<td>Discussion</td>
<td>73</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>83</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Figure 1. LINE structure ..................................................................</td>
<td>6</td>
</tr>
<tr>
<td>2.</td>
<td>Figure 2. Reptile lineages in vertebrate phylogeny.............................</td>
<td>6</td>
</tr>
<tr>
<td>3.</td>
<td>Figure 3. Estimated Anolis lineages based on Nicholson, 2012..............</td>
<td>6</td>
</tr>
<tr>
<td>4.</td>
<td>Figure 4. Anolis species nomenclature.............................................</td>
<td>10</td>
</tr>
<tr>
<td>5.</td>
<td>Figure 5. Geographic Range of Anolis carolinensis...............................</td>
<td>11</td>
</tr>
<tr>
<td>6.</td>
<td>Figure 6. Geographic range of Anolis apletophallus.............................</td>
<td>12</td>
</tr>
<tr>
<td>7.</td>
<td>Figure 7. Geographic range of Anolis auratus.....................................</td>
<td>12</td>
</tr>
<tr>
<td>8.</td>
<td>Figure 8. Geographic range of Anolis frenatus...................................</td>
<td>13</td>
</tr>
<tr>
<td>9.</td>
<td>Figure 9. Anolis genome assembly statistics......................................</td>
<td>15</td>
</tr>
<tr>
<td>10.</td>
<td>Figure 10. NJ tree of Acar L1 DNA ..................................................</td>
<td>21</td>
</tr>
<tr>
<td>11.</td>
<td>Figure 11. NJ tree of Acar L1 protein.............................................</td>
<td>22</td>
</tr>
<tr>
<td>12.</td>
<td>Figure 12. NJ tree of Aapl L1 protein.............................................</td>
<td>22</td>
</tr>
<tr>
<td>13.</td>
<td>Figure 13. ML tree of Aapl CR1 protein...........................................</td>
<td>29</td>
</tr>
<tr>
<td>14.</td>
<td>Figure 14. ML tree of Aapl RTE-1 protein.......................................</td>
<td>31</td>
</tr>
<tr>
<td>15.</td>
<td>Figure 15. ML tree of Aapl RTE-BovB protein....................................</td>
<td>32</td>
</tr>
<tr>
<td>16.</td>
<td>Figure 16. ML tree of Aaur L1 protein...........................................</td>
<td>35</td>
</tr>
<tr>
<td>17.</td>
<td>Figure 17. ML tree of Aaur CR1 protein...........................................</td>
<td>38</td>
</tr>
<tr>
<td>18.</td>
<td>Figure 18. ML tree of Aaur RTE-1 protein.......................................</td>
<td>40</td>
</tr>
<tr>
<td>19.</td>
<td>Figure 19. ML tree of Aaur RTE-BovB protein....................................</td>
<td>42</td>
</tr>
<tr>
<td>20.</td>
<td>Figure 20. ML tree of Acar L1 protein...........................................</td>
<td>46</td>
</tr>
<tr>
<td>21.</td>
<td>Figure 21. ML tree of Acar CR1 protein...........................................</td>
<td>47</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>Figure 22. ML tree of Acar RTE-1 protein</td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>Figure 23. ML tree of Acar RTE-BovB protein</td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td>Figure 24. ML tree of Afre L1 protein</td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>Figure 25. ML tree of Afre CR1 protein</td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td>Figure 26. NJ tree of Afre RTE-1 protein</td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>Figure 27. ML tree of Afre RTE-BovB protein</td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td>Figure 28. ML tree of LINE/L1 consensuses for 4 species</td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td>Figure 29. ML tree of LINE/L1 consensuses for 4 species and outgroups</td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>Figure 30. LINE elements in raw numbers for 4 anole species</td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td>Figure 31. Percent Genome occupied by repeats in 4 species</td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>Figure 32. LINE Content Percentages among four anole species</td>
<td></td>
</tr>
<tr>
<td>33.</td>
<td>Figure 33. Anolis Repeat Content vs. Total Genome Length</td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td>Figure 34. Average Family Divergences for LINE/L1</td>
<td></td>
</tr>
<tr>
<td>35.</td>
<td>Figure 35. Numbers of Families by Average Percent Divergences</td>
<td></td>
</tr>
<tr>
<td>36.</td>
<td>Figure 36. KS-Test with Aapl and Afre</td>
<td></td>
</tr>
<tr>
<td>37.</td>
<td>Figure 37. Total Repeat Content by Vertebrate Species</td>
<td></td>
</tr>
<tr>
<td>38.</td>
<td>Figure 38. Percent of Interspersed Repeat Content in Vertebrate Genomes</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

Repeat elements are viral-like nucleotide sequences present in prokaryote and eukaryote genomes. Repeats have their own history of replicating in the host species’ genomes and have propagated through both vertical and horizontal transfer (Zupunski 2001). Transposons are a specific type of repeat element that encodes enzymes that facilitate the transcription, transposition and integration of their sequences into a host genome. The oldest transposons, non-long terminal repeat (non-LTR) retrotransposons, are sorted into 12 clades, with vertebrates having the L1, CR1 and RTE clades (Zupunski 2001). In this thesis, LINE/L1, LINE/CR1, LINE/RTE-1, and LINE/BovB will be investigated in four anole lizards.

Genome size in vertebrates varies drastically. In mammals, the average genome size is 3.37 pg (1 pg is roughly equal to 1 gigabase pairs), while marsupials average around 4.05 pg and placental mammals at 3.3 pg (Gregory, 2005). In contrast to mammals, reptiles have a mean genome size of 2.24 pg (Gregory, 2005). In particular, lepidosaurs (lizards, snakes and lisamphibasenians) have average genome sizes around 2.10 pg (Gregory, 2005). Fish demonstrate very similar genome sizes, with the average around 1.93 pg (Gregory, 2005). Aves have an average genome size of 1.38 pg (Gregory, 2005). Chelonians and crocodilians display larger genomes at 2.87 pg and 3.11 pg, respectively (Gregory, 2005). In comparison to reptiles and mammals, amphibians have enormous genome sizes, with an average of 18.14 pg (Gregory, 2005). Prior studies on the salamander genome have revealed that the massive size is primarily due to retrotransposons, which averages around 35.9 pg (Mueller 2012; Gregory, 2005). The
largest vertebrate genome is the lungfish, which averages at 132.8 pg (Gregory, 2005). This thesis will explore the notion that LINES are the major driver of genomic expansion. The patterns of genomic composition created by LINE retrotransposons will be compared between species.

In eutherian mammals, LINES have amplified to very high copy numbers with 800,000 copies identified in the human genome (Novick et al., 2009). LINES have played a critical role in shaping genome size and is a major determinant of driving genomic expansion (Kidwell 2002). All non-LTR retrotransposon clades are estimated to have originated over 500 million years ago (mya) during the Pre-Cambrian (Malik et al., 1999; Kazazian and Goodier, 2002). One of the most important repeats in vertebrate evolution is the LINE/L1 clade.

There are dramatic differences in the patterns of evolution of L1 between fish and mammals. In mammals, twenty percent of the genomic mass in species derives from a single L1 lineage, while in fish, there are over 30 distinct L1 lineages (Furano et al., 2004). The pattern of evolution for LINE/L1 in eutherian mammals is surprising because only one family is active at a time (Tollis, 2012). LINES are well-studied repeats in vertebrates, and serve as a good representative sample for investigating repeat evolution due to their pervasiveness. Previous studies have revealed that lizard genomic compositions are more similar in patterns of family diversity to fish genomes than mammalian genomes (Furano et al., 2004). In contrast to mammals, where L1 comprises a single distinct lineage, the L1 clade in anoles split into multiple, active families that vary in both diversity and abundance. In this study, LINE clades in four lizard genomes are analyzed with a focus on abundance, diversity and activity.
LINES are long interspersed nuclear elements about 6 kbp in length (Smit 1996). LINES are autonomous repeats that have the ability to replicate through reverse transcription, and transpose and integrate back into the host genome (Tollis, 2012). There are major classes of repeats present in all eukaryotes and the repeat element content can vary even between species.

This thesis will conduct evolutionary comparisons of repeat element sequences among four anole species of the Anolis adaptive radiation. LINE retrotransposons of A. apletophallus, A. auratus, A. carolinensis and A. frenatus will be analyzed. Abundance and diversity of family members within various LINE clades will be investigated.

Adaptive radiations serve as models for studying interspecific diversification (Calsbeek et al., 2007). This study will focus on the effects of retrotransposon evolution on genomic composition within the Anolis adaptive radiation. Comparative phylogenetic approaches will be used to reveal more about the role of repeat elements in shaping genome evolution.

There are two types of repeat elements; Class 1 requires an RNA intermediate (retrotransposon) and Class 2 elements use a DNA intermediate (Tollis, 2012). Having an open reading frame (ORF) that encodes transposase, Class 2 elements transpose as single or double-strand DNA (Tollis, 2012). Both Class 1 and Class 2 repeat elements coexist in a wide range of Eukaryotes, suggesting their ancient evolutionary origins (Sun et al., 2010). Autonomous repeat elements encode for components that allow them to replicate and mobilize, including reverse transcriptase, endonuclease, RNA Pol II and III, enhancers, polyadenylation and transcription factor binding sites (Tollis, 2012; Rebollo et al., 2012; Singer et al., 2010; Thornburg et al., 2006). Non-LTR retrotransposons are
repeats that do not have long terminal repeats in their sequences (Tollis, 2012). LINE/L1 is a dominant non-LTR, which is the focus of this study.

Some families of repeat elements are able to move and replicate in the genome, by encoding reverse transcriptase and endonucleases. Other repeats, like SINES, rely on other elements (i.e. LINES) in order to encode the reverse transcriptase for them. Over time, the integrity of a TE can degrade such that it loses the ability to replicate autonomously. Decaying transposons are present throughout vertebrate genomes. Interspersed repeats scattered throughout the genome are mostly transposons in varying stages of decay (Hillier et al., 2004). Some transposable element fragments are highly conserved across classes (Bejerano et al., 2004).

The A. carolinensis genome has enormous diversity of Class 2 transposons which are active or were recently active (Boissinot 2010). There are still some active repeat elements that can autonomously replicate. Many repeat element sequences have been domesticated by their hosts and have been functionally co-opted and integrated into gene regulatory networks (Feschotte 2008). Many transposons have lost their ability to self-regulate and mobilize, and now contain with fragmented ORFs. Fragments tend to accumulate during the low-fidelity process of transposition (Pavlicek et al., 2006) Poor processivity can lead to the truncation of open reading frames, and sometimes, as with RTE-1 this can result in the generation of non-autonomous SINE elements (Zupunski et al., 2001).

Besides changing the composition of other genes, via transposition, repeat elements can also change the structure of the repeats themselves. Repeat elements tend to
Archosaur and squamate genomes have five clades of non-LTR-RT including CR1, R2, L2, RTE and L1 (Tollis 2012). In *A. carolinensis*, there is evidence of relatively low divergence in L1, which is indicative of recent insertion events (Novick et al. 2009; Tollis 2013). Evidence that older LINEs that have lost their ability to replicate can be found in highly divergent fragments of LINE open reading frames present in the genomes. To date, LINE profiles of only few squamates have been analyzed. This study will contribute three more LINEs analyses to squamates, from anoles deriving from a monophyletic lineage.

Next-generation genome sequencing has resulted in a vast bottleneck of genomic data. There are several databases dedicated to studying repeat elements such as RepeatMasker and Repbase. With new next-gen genomes for three anole species, it is now possible to evaluate genomic changes during very minimal evolutionary time periods to observe patterns of genomic change within a genus.

LINEs are long interspersed nuclear elements about 6 kb in length. SINEs are shorter than LINEs and contain a tRNA-like region, as well as a 3' sequence that is identical to a corresponding LINE partner. LINEs contain an RNA polymerase II promoter, which recruits RNA replication machinery (Rebollo et al., 2012). Almost all lines are truncated at the 5' end, resulting in a high proportion of LINEs that are comprised mainly of 3' terminal fragments (Terai et al., 1998; Fantaccione et al.,) These truncated forms are evidence that LINE encoded RTases recognize the 3' end of an RNA strand to begin translational synthesis (Eickbush, 1992; Fantaccione et al.,). For this
study, the reverse transcriptase sequences in open reading frame 2 (ORF2) were collected, analyzed and used for phylogenetic analysis.

Figure 1 LINE structure

VERTEBRATE STUDIES

To understand reptiles in the broader scope of amniote evolution (with the genomes publically available), species from 3 key groups, the clade Archosauromorpha, this study begins by analyzing the order Squamata and the class Mammalia.
The modern reptilian lineages include archosaurs, squamates and tuataras and chelonians. The extant Archosaurian clade includes crocodiles and birds, which is estimated to have diverged about 100 mya (St. John et al., 2013). Avian and non-avian reptiles all fall within the clade Sauropsida. Squamates include about 9,400 extant lizards and snakes and is one of the most diverse terrestrial radiation of vertebrates (Pyron et al., 2013). Archosaurs and Squamates are estimated to have diverged about 275 million years ago, while the common ancestor of mammals and reptiles is estimated over 300 mya (Hedges et al., 2006).

Mammals are in the clade Synapsida and are endothermic amniotes. There are roughly 5,000 species (Wilson, 2005). Wide-scale studies of the genomic landscapes of mammals have been conducted on roughly 29 genomes (Lindblad-Toh et al., 2011). Reptiles, a paraphyletic clade, is comprised of roughly 6,500 extant species and includes turtles, archosaurs (crocodiles and birds), squamates (lizards and snakes) and tuataras (Godwin and Crews, 1997). In contrast to mammals, relatively few reptilian genomes are available for large-scale comparisons.

Three crocodilian species are being sequenced; however, only draft assemblies have been used for publication. Initial reports indicate that the alligator and salt water crocodile genomes are comprised of 23% to 28% repeat content, in contrast to 50% in
humans (St. John et al., 2013). Alligator appears to have more repeats than the salt-water crocodile (St. John et al., 2013).

The snake lineage is a radiation that has a particularly diverse phenotypic range, comprised of 3,100 species (Castoe et al., 2011). Snakes and other squamate reptiles diverged around 170 mya (Castoe et al., 2009a). Castoe et al., conducted an intraspecific repetitive genomic study on two snakes that diverged around 100 mya. Likewise, this study investigates lizards that initially diverged ~80 mya. This parallel study provides comparative data to help demonstrate how repeat elements changed the structure and function of genomic regions may be better determined.

In python, repeat element activity not only affects gene regulation (which is intimately tied to metabolic pressures), but can also ontogenetically affect rates of recombination and gene duplication (Castoe et al., 2011). Castoe et al. experimentally determined that some repeat element transcripts are present in relatively high levels. Since other in vivo experiments are lacking, the ontogenetic magnitude of repeat element activity within one organism is unknown. The python genome is small at 1.4 Gbp and has a low repeat element content around 21%, yet the copperhead has more like 45% repeats (Castoe et al., 2011). There are highly divergent repeat element landscapes, even between snakes. The copperhead had 23-fold increase in repeat element transcripts than the python (Castoe et al., 2011).

**ANOLIS CLADE**

The *Anolis* clade is a group of Iguanian lizards (Losos 2009). Comprised of roughly 400 species, *Anolis* is the largest genus of land vertebrates (Nicholson, 2012). There is much controversy over the phylogenetics of the *Anolis* clade, in particular
whether or not it should be split into multiple genera. The newest study offering molecular evidence for phylogenetic inference is from Nicholson (2012), which places Dactyloa at 87 million years old. Older studies indicate that the entire *Anolis* genus is roughly 66 mya based on relaxed molecular clock estimates using ectotherm mtDNA evidence (Glor, Losos 2010; Nicholson, 2012). However, in the absence of fossil evidence, even with molecular estimates, the controversy will continue. According to the most recent publication, which utilizes Bayesian methods, the lineage including *Anolis carolinensis* appeared roughly 74.7 million years ago (Nicholson, 2012). The Norops radiation, which includes *Anolis auratus* and *Anolis apletophallus* is dated at 51.6 million years (Nicholson, 2012). While some of these radiations may be old, the species may be quite young.

![Figure 3. Estimated Anolis lineages based on Nicholson, 2012.](image)

*Anoles are a model system for studying adaptive radiation, as they derive from a single lineage but have come to inhabit a wide-range of ecological niches. Anoles underwent a rapid speciation that resulted in a high level of morphological differences*
within the genus (Losos, 2009). The term ecomorphs is used to describe the characterization of anoles based on body shape and presumably the ecological constraints that helped to shape it. Ecomorphs are anoles with distinct habitats and likewise different morphology adapted to its niche (Calsbeek et al., 2007)

In this study, three different ecomorphs are investigated. In this study, two trunk-ground anoles, a grass-bush anole and a crown-giant anole are used for comparative phylogenetics. Anoles’ abilities to inhabit new niches and adapt physically to the environment raises interesting questions at the intersection of ecology, development, genetics and evolution.

The Dactyloa clade is an Iguanian monophyletic family comprised of 499 species and subspecies (Nicholson et al., 2012). Dactyloan anoles split from more recent anole radiations around 87 mya, based on Bayesian phylogenetic analyses, while Norops is estimated to have appeared between 50 and 60 mya (Nicholson et al., 2012). The “crown-giant” *A. frenatus* is a member of the Dactyloa clade and occupies the deepest node out of the four anoles investigated here on the phylogenetic tree of *Anolis* evolution.

Mainland anoles of Central and Northern South America are in the Norops and Dactyloa clade (Losos, 2009). *A. auratus* and *A. apletophallus* are part of the Norops clade, which is a monophyletic group comprised of roughly 190 species and subspecies (Nicholson et al., 2012). Norops radiated from the mainland to the neighboring islands, and some species radiated back to the mainland once again (Losos, 2009).

The names of anoles are abbreviated according to the naming conventions proposed by the *Anolis* Gene Nomenclature Committee in “Developing a community-
based genetic nomenclature for anole lizards” (Kusumi et al., 2011). The four-letter abbreviation will be used to refer to the four species in Chapters 2 and 3.

<table>
<thead>
<tr>
<th>Scientific Name</th>
<th>Common Name</th>
<th>Abbreviation</th>
<th>Ecomorph</th>
<th>Repeat Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. carolinensis</em></td>
<td>green anole</td>
<td>Acar</td>
<td>trunk-crown</td>
<td>L1_AC_1</td>
</tr>
<tr>
<td><em>A. apletophallas</em></td>
<td>slender anole</td>
<td>Aapl</td>
<td>trunk-ground</td>
<td>L1_AP_1</td>
</tr>
<tr>
<td><em>A. auratus</em></td>
<td>grass anole</td>
<td>Aaur</td>
<td>grass-bush</td>
<td>L1_AA_1</td>
</tr>
<tr>
<td><em>A. frenatus</em></td>
<td>bridled anole</td>
<td>Afre</td>
<td>crown-giant</td>
<td>L1_AF_1</td>
</tr>
</tbody>
</table>

*Figure 4. Anolis species nomenclature*

*A. carolinensis*, or the green anole, is native to the southern United States, ranging from Texas to the East Coast and as far north as Tennessee (Tollis, 2012). The green anole occupies a trunk-crown niche and is the only member of the *Anolis* genus native to the United States. All other anoles live in the sub-tropics or tropics, with the exception of the invasive *A. sagrei* in southern Florida. While no consensus exists for estimating the origination of *Anolis*, it is estimated to have appeared 74.7 million years ago (Nicholson, 2012). While there is a lot of controversy in the *Anolis* community as to when the clade originated, the estimates range from 66 to 75 mya. There is strong evidence to suggest that *A. carolinensis* dispersed from Cuba via an across-water dispersal event during the late Miocene to Pliocene. (Glor and Losos, 2005; Losos 2010).
A. apletophallas, the slender anole, formerly classified as A. limifrons, is part of a species complex together with A. cryptolimifrons and the remaining A. limifrons. It is native to Central America. It belongs to the Norops clade and primarily is found near the ground (Losos, 2010) and is most closely classified as a trunk-ground anole. The range of A. apletophallas is limited to Central America and is limited by oceans on the East and West coasts of Panama and Costa Rica.
Figure 6. Geographic range of *A. apletophallus*, reproduced from Reptile-database.org

*A. auratus*, the grass anole, has an enormous range from Central America to northern South America, as far south as Brazil. It belongs to the Norops clade and is characterized as a grass-bush ecomorph (Losos, 2009).
Figure 7. Geographic range of *A. auratus*, reproduced from Reptile-database.org

*A. frenatus*, or the bridled anole or Central American giant anole, has a range from Costa Rica to Colombia. It belongs to the Dactyloa clade and is characterized as a crown-giant ecomorph (Losos, 2009).
Figure 8. Geographic range of *A. frenatus*, reproduced from Reptile-database.org

**GENOME ASSEMBLY**

This study began with the sequencing of the genomes generated by the Kusumi Lab, and applies bioinformatic analysis to assess evolutionary questions at the intersection of development and genetics. It includes the use of four genomes, which were produced using very different methods. Because of these base differences, it is important to first discuss how the genomes were respectively constructed, so that they may be accurately compared. The *A. carolinensis* genome assembly was Sanger sequenced and was built by the Broad Institute at Harvard. The three other genomes were
constructed using Illumina High-seq next-generation genome assemblies from the Kusumi Lab. In addition, Repbase and RepeatMasker were used to collect other repeat elements datasets in other vertebrate species.

*A. carolinensis* was the first non-avian reptile and lizard sequenced. The Broad Institute used Sanger sequencing technology, which generates reads of roughly 500-700 base pairs, on primarily one female individual. The *A. carolinensis* genome has about 7.1x coverage and the N50 median contig length is 83 kilo-base pairs (kb). The size of the genome assembly is 1.8 gigabase pairs (Gbp), although the total estimated genome is predicted at 2.2 Gbp.

The sequencing of the *A. apletophallus, A. auratus* and *A. frenatus* genomes was carried out by the Kusumi Lab, in collaboration with the Translational Genomics Research Institute staff. The group constructed a mate-pair library was constructed using an Illumina mate-pair prep kit. 200 bp, 300 bp, 1000 bp, and 3000 bp insert size libraries were constructed. The Illumina’s HiSeq 2000 next-generation sequencing platform was used to generate 104 bp paired-end reads. A whole shotgun genome sequencing technique was used to generate millions of raw reads. Finally, contigs and scaffolds were assembled and the current versions of the genome assemblies were given to me for analysis.

While the genome sequencing efforts are ongoing, the current genome builds include: *A. auratus* had a 87x coverage of breadth, with 93.5% of the genome covered. The N50 contig length was 11.2 kb and the scaffold N50 was 16 kb. *A. apletophallus* had a 43x coverage of breadth, with 71.2% of the genome covered. The N50 contig length was 1.7 kb and the scaffold N50 was 3. 1kb. *A. frenatus* had a 55x coverage of breadth,
with 68.5% of the genome covered. The N50 contig length was 3.1 kb and the N50 scaffold length was 8.9 kb.

<table>
<thead>
<tr>
<th></th>
<th>Percent Complete</th>
<th>Coverage</th>
<th>contig N50</th>
<th>scaffold N50</th>
<th>quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. auratus</em></td>
<td>93.5%</td>
<td>87x</td>
<td>11.2 kb</td>
<td>16 kb</td>
<td>Q20</td>
</tr>
<tr>
<td><em>A. apletophallus</em></td>
<td>71.20%</td>
<td>43x</td>
<td>1.7 kb</td>
<td>3.1 kb</td>
<td>Q20</td>
</tr>
<tr>
<td><em>A. frenatus</em></td>
<td>68.50%</td>
<td>55x</td>
<td>3.1 kb</td>
<td>8.9 kb</td>
<td>Q20</td>
</tr>
</tbody>
</table>

*Figure 9. Anolis* genome assembly statistics
INTRODUCTION

Comparisons of repeat elements between organisms have been conducted at the class, order and family level. While other genomic studies have shown the differences in repeat element abundance, diversity and structure among classes, little research has been done on intrageneric repeat element divergence. This study investigates changes to genomic composition between four anole species in the Anolis adaptive radiation. In order to statistically determine significance, average pairwise divergences from family consensuses were collected for each L1 repeat family, for each respective species. Mann-Whitney U tests and Kolmogorov-Smirnov tests were performed on the average divergences from family consensuses to determine if species-level differences are due to LINE activity. This thesis conducts preliminary comparisons in order to lay a foundation for exploring how LINE retrotransposons might alter genomic composition and what potential effects these changes, if any, may have on the phenotype.

HYPOTHESIS

The null hypothesis is that LINEs have not created significant changes to the genomic compositions of species within the Anolis genus. If this is the case, similar copy number, diversity and activity of LINE retrotransposons should be found in all four anoles. The alternative hypothesis is that LINES have created significant genomic changes within the Anolis genus. LINE/L1 was the primary LINE family used to determine divergences between the four species. In particular, the average divergences within families and between families were calculated for each respective species, for
Statistical tests were used to determine if significant differences exist between the L1 family divergences. If the p-values of the statistical tests are less than 0.05, then the alternative hypothesis may be accepted. If the p-value is greater than 0.05 than there is no statistical significance for the evolutionary divergences between species. Prior studies have identified that in Acar, the average pairwise divergence for the L1 family from consensus is between 1 and 3% (Novick et al., 2011). This study determines whether the low rate of divergence within LINE families is also observed in the other three species sequenced. Analyses were conducted using bioinformatic tools to extract LINEs from large genomic datasets. Phylogenetic reconstructions and analyses were conducted to identify statistically significant changes in genomic composition among four species in the Anolis clade.

METHODS

RepeatModeler. RepeatModeler was used to survey the repeat landscape of four anole genome assemblies. RepeatModeler’s function is the de novo identification of repeat elements, which utilizes a repeat database called Repbase. The RepeatModeler Suite was run on genome assemblies for Aapl, Aaur and Afre. An Acar repeat library was obtained from Dr. Stéphane Boissinot and used as an Anolis custom repeat library with the RepeatModeler Suite. RepeatMasker was also used to generate data files for each of the three species, using Anolis as the assumed species. A master file with repeat element data was generated and then used for comparison to other vertebrate classes. The RepeatModeler file contained overall statistics and percentages of elements in the
genome and was used to determine differences between the four species in abundance, as a percentage of genome occupied by repeats.

Several clades of LINES were analyzed, including the L1, CR1, RTE-1 and RTE-BovB repeat element clades. Since the full-length elements range from 5 to 10 kb, only the reverse transcriptase (RT) domains were used to differentially search for and select LINE sequences. The RT is roughly 300 amino acids, but in some cases, as with AapL sequences were truncated due to contig length and alignments of proteins 150 base pairs or more were used.

**collection of line elements.**

Four LINE clades were chosen that display particularly interesting patterns of diversification: L1, CR1, RTE-1 and RTE-BovB. Using Repbase, consensus sequences for the Acar LINE/L1 elements were obtained. The DNA sequences were uploaded into NCBI’s ORF Finder program. Next blastp was used to find matching protein sequences for each L1 element. The protein sequence for L1-encoded reverse transcriptase was extracted. Elements were also checked for correct clade identity through Repbase’s CENSOR program. In particular, the similarity of CR1 and L2 elements was problematic. In order to prevent errors in the pairwise divergences and in phylogenetic reconstruction, I ensured that each aligned sequence belonged to the correct LINE clade.

LINES are on average around 6 kb in length and are often highly divergent, making it very difficult to conduct a sequence analysis on them. The ability to locate LINES is also highly dependent on the quality of the genome assembly; in particular the presence of long scaffolds over 6 kb are required to find full length LINE elements. Only Afre and Aaur contained average scaffold lengths larger than 6 kb. Subsequently, protein
sequences of the highly conserved reverse transcriptase domains were used for a four way comparisons of LINES in *Anolis*. Using tblastn, I used the *Anolis carolinensis* consensus sequence as a query and each respective anole genome assembly as the database. Matching sequences were saved in a protein fasta format.

Prior to generating phylogenetic reconstructions, I first made sure that the L1_AC proteins within one species aligned to each other using the MUSCLE algorithm in Geneious (Biomatters Ltd.), with a minimum of eight iterations. Once the proteins were aligned, a consensus sequence was generated and a tree was then constructed. Initially, neighbor-joining (NJ) trees were created using the Jukes-Cantor genetic distance model, and a NJ tree building method. Between 500 and 1,000 bootstraps were used for each phylogenetic reconstruction. In order to prevent the construction of biased trees, each alignment was also screened for sequence duplicates. After creating an initial survey of NJ protein trees for all four species, for all four clades of LINE elements, maximum likelihood (ML) trees were then created.

**phylogenetic reconstruction.**

The phylogenetic analysis of protein sequences is based on comparisons of amino acid substitution rates and statistical models that imply relatedness during the pairwise comparison of sequences (Pevsner 2009). In constructing phylogenetic trees, several methods were used. Geneious produced two types of trees were constructed that rely on a distance, or similarity matrix. While phylogenetic reconstructions were conducted primarily on protein sequences, some DNA trees were also generated. The fragmented nature of 6 kb DNA sequences and rates of divergence between species proved slightly problematic, prompting the use of proteins for phylogenetic reconstruction. The DNA
trees provided additional support for the phylogenetic relationships I identified using proteins, in particular the identification of new repeat element families in Aaur, Aapl and Afre.

Several criteria were used for designating a family status within the LINE/L1, the most important of which was evidence of phylogenetic relatedness. Family designations were assigned based on the bootstrap values of multiple trees (NJ, UPGMA, ML), tree topology and distance matrices. An iterative search and refine method was used to generate trees, identify perceived families, extract those families, and to locate more family members via tblastn. Every tree was generated at least five times in an effort to create the most accurate phylogenetic reconstruction, with the most family members available. Redundancy in phylogenetic trees allowed for a greater degree of certainty when designating families that consistently appeared across tree methods and with the addition of more family members. Bootstrap values over 80 were considered strong evidence of relatedness, while values under 60 were considered weak and only used circumstantially to infer relatedness based on general topology, for very distal branches with high bootstrap values, and in conjunction with other modes of evidence such as distance matrix data or maximum likelihood trees. Pairwise deletion was used to evaluate all of the gaps.

Each phylogenetic tree model was tested in MEGA for reliability via the maximum likelihood statistical method. For all three of our species, it was determined that the best model to use is the JTT+G, or the Jones-Taylor-Thornton model with a gamma distribution rate among sites. Since these options followed recommendations from MEGA on the evolutionary analysis of amino acids, little revision needed to be
done. To model continuous variables with a skewed distribution, the gamma distribution
is often used, which is particularly useful for dealing with unequal substitution rates
across sites (Pevsner, 2011). I used the MEGA defaults for the gamma parameter, with
one as a gamma parameter for the distance estimation, and five for constructing
maximum likelihood trees. Gaps were treated using partial deletion with a 95% cutoff.

Because prior publications on Acar LINE/L1s were based on nucleotide
sequences, a phylogenetic DNA tree was reproduced using only the protein ORF for the
reverse transcriptase. Indeed, the protein tree is nearly identical to the DNA tree
published in Novick et al., 2011. In order to compare Acar with the three new species,
protein trees were constructed for all four species.
Figure 10. Neighbor-joining phylogenetic tree for the LINE/L1 clade in Acar, based on consensus DNA sequences. 1,000 bootstraps were executed.
Figure 11. A Neighbor-Joining tree of Acar LINE/L1 clade based on consensus protein sequences. 1,000 bootstraps were executed.

Well-supported branching structures were observed on both the DNA and protein tree, demonstrating similar family relationships. Due to the close similarity of the DNA and protein NJ trees, the sole use of protein sequences in the molecular phylogenetic analysis was used. DNA trees were also produced, but are not pictured here.
Each family alignment was imported into MEGA, a family consensus sequence was generated and pairwise divergences were conducted in MEGA using a Jones-Taylor-Thornton (JTT) model with 500 – 1,000 bootstraps and a pairwise deletion treatment of gaps.

**comparative genomic analysis.**

A maximum likelihood tree for all four anole lizards was constructed for the protein consensus sequences of the L1 families. The L1 consensuses for human, zebrafish and *Xenopus* were also included as outgroups, when possible.

In addition to the LINE/L1 clade, other LINE members were evaluated, as well. Proteins from the LINE/CR1 clade were collected via tblastn. CR1 was originally found to be one of the few repeats that dominate the generally repeat-sparse chicken genome, but it is present in all vertebrate genomes (Fantaccionne et al., 2004). While multiple lineages of CR1 were active in chicken, they now appear to be inactive (Tollis 2012), as indicated by a lack of full-length elements.

Protein sequences for the LINE/RTE-1 clade were collected via tblastn for each species. RTE is an old lineage of non-LTR repeats, first found in *C. elegans* (Youngman et al., 1996), which has two divergent lineages, RTE-1 and RTE-BovB that appeared before the origin of vertebrate (Novick 2009). RTEs are found in vertebrates and arthropods (Malik and Eickbush, 1998). In a divergence versus time study, the RTE lineage was found to demonstrate increasing divergence over time, with the split of tetrapods and teleosts having maximal saturation of the elements around 400 mya (Zupunski et al., 2001). RTE-BovB arose in vertebrate squamates through horizontal transfer [CITE].

26
The RTE-BovB family has propagated very successfully in *Anolis*, with about 3,000 copies present in the genome (Novick 2009). Protein sequences for the LINE/RTE-BovB clade were collected for each species. For all 4 anoles, the LINE/RTE-BovB had just one distinct family. The pairwise divergences of the RTE-BovB family for three anoles were calculated, using 500 bootstraps.

**Statistical Tests**

In order to test statistical significance, the Mann-Whitney U test, or the Wilcoxon rank-sum test, was used to determine if the average divergences for L1 family members from the family consensus significantly differ. The Kolmogorov-Smirnov test was also conducted on each pair of species.

**RESULTS**

The results section is divided up into sub-sections. Individual summaries of the results for each species will appear first. Following species’ sections will be an overall section comparing LINES between the four species.

**Anolis apletophallus**

*repeatmodeler.*

RepeatMasker was used to scan the Aapl genome assembly for repeats. The query species was assumed to be *Anolis*. In Aapl, nearly 14.4% of the genome is comprised of repeats. 9.8% of the Aapl genome is comprised of retroelements, including 7.3% occupied by LINES and 1.8% occupied by SINES. Of that 7.3% occupied by LINES,
4.6% is comprised of CR1 and L2, 0.8% is occupied by RTE and 1% is occupied by L1. LTR elements occupy 0.6% of that 9.8% of retroelements that comprise the repeat landscape. DNA transposons occupy 1.9% of the repeat landscape. In addition, total interspersed repeats comprise 12.28% of the repeat landscape and simple repeats comprise 1.2%.

**collection of LINE elements.**

Acar LINE consensuses were collected from the Repbase repeat database. The DNA sequence was entered into NCBI’s ORF finder. I identified the reverse transcriptase (RT) domain of the ORF2 for the LINE sequences and used blastp to retrieve just the RT protein sequence. This RT consensus (for each respective LINE clade) was used to retrieve protein sequences in the Aapl genome using the tblastn program. The protein sequences were then uploaded into Geneious and aligned using eight iterations of the MUSCLE program. A total of 295 LINE elements from Aapl were retrieved from protein blasts and used to construct phylogenetic trees for LINE/L1, LINE/CR1, LINE/RTE-1, and LINE/RTE-BovB.

**phylogenetic reconstruction.**

The alignments were exported and loaded into MEGA. Maximum likelihood trees were generated using 500 or more bootstraps. The maximum likelihood statistical method model was tested in MEGA for reliability, specifically for Aapl. Gaps were treated using partial deletion with a 95% cutoff. The best model for Aapl is the JTT+G+F, or the Jones-Taylor-Thornton model with a gamma distribution rate among sites.

LINE/L1.
The LINE/L1 alignment included 139 protein sequences found through tblastn. NJ and ML trees were generated for LINE/L1 elements found in the tblastn. Input from both the NJ and ML trees were used to designate LINE families. The naming convention for LINE/L1 repeat families (i.e. L1_AP_1) in Aapl followed Acar repeat family names in the Repbase database. In Aapl, at least 16 distinct families were identified in the LINE/L1 clade. The L1 tree for Aapl was highly structured with well-supported nodes.

Once families were designated, a consensus sequence was generated for each family and re-aligned to the family. The alignment with consensus was then imported into MEGA for pairwise divergences. The pairwise divergences within the families ranges from 0 to 20%. The average divergence from the family consensus was 10%.
Figure 12. Maximum likelihood phylogenetic tree for the LINE/L1 clade in *Anolis apletophallus*, based on protein sequences.

LINE/CR1.

A total of 56 protein sequences for CR1 were collected from the Aapl genome assembly. The Aapl CR1 clade has two distinct families. The average divergence from the family consensus was 13.7%. However, between those two families, the average divergence from consensus was 46.50%.
Figure 13. A Maximum likelihood tree was constructed for LINE/CRI proteins for *Anolis apletophallus*. 500 bootstraps were executed.

LINE/RTE-1.

A total of 12 protein sequences for RTE-1 were collected from the Aapl genome assembly. Aapl has one distinct RTE-1 family. The average divergence from the RTE-1 family consensus was 24.3%.
Figure 14. A Maximum likelihood tree was constructed for LINE/RTE-1 proteins for *Anolis apletophallus*. 500 bootstraps were executed.

LINE/RTE-BovB.
A total of 88 protein sequences for RTE-BovB were collected from the Aapl genome assembly. Aapl has just one RTE-BovB family. The average divergence from family consensus is 13.9%.
Figure 15. A Maximum likelihood tree was constructed for LINE/RTE-BovB proteins for *Anolis apletophallus* with 1,000 bootstraps were executed.

*Anolis auratus.*

*repeatmodeler.*

RepeatMasker was used to scan the Aaur genome assembly for repeats. The query species was assumed to be *Anolis*. In Aaur, nearly 25% of the genome is comprised of repeats. 16% of the Aaur genome is comprised of retroelements, including 13% occupied by LINES and 2% occupied by SINES. Of that 13% occupied by LINES, 7.7% is comprised of CR1 and L2, 2% is occupied by RTE and 1.3% is occupied by L1. LTR elements occupy 1.1% of that 16% of retroelements that comprise the repeat landscape. DNA transposons occupy 2.6% of the repeat landscape. In addition, total interspersed repeats comprise 20% of the repeat landscape and simple repeats comprise 4%.

*collection of LINE elements.*

Acar LINE consensuses were collected from the Repbase repeat database. The DNA sequence was entered into NCBI’s ORF finder. I identified the reverse transcriptase (RT) domain of the ORF2 for the LINE sequences and used blastp to retrieve just the RT protein sequence. This RT consensus (for each respective LINE clade) was used to retrieve protein sequences in the Aaur genome using the tblastn program. The protein sequences were then uploaded into Geneious and aligned using 8 iterations of the MUSCLE program. A total of 289 LINE elements for Aaur were retrieved from protein blasts and used to construct phylogenetic trees for LINE/L1, LINE/CR1, LINE/RTE-1, and LINE/RTE-BovB.
*phylogenetic reconstruction.*

The alignments were exported and loaded into MEGA. Maximum likelihood trees were generated using 500 or more bootstraps. The maximum likelihood statistical method model was tested in MEGA for reliability, specifically for Aaur. Gaps were treated using partial deletion with a 95% cutoff. The best model for Aaur is the JTT+G+F, or the Jones-Taylor-Thornton model with a gamma distribution rate among sites.

LINE/L1.

The LINE/L1 alignment included 135 protein sequences found through tblastn. NJ and ML trees were generated for LINE/L1 elements found in the tblastn. Input from both the NJ and ML trees were used to designate LINE families. The naming convention for LINE/L1 repeat families (i.e. L1_AA_1) in Aaur followed Acar repeat family names in the Repbase database. In Aaur, at least 29 distinct families were identified in the LINE/L1 clade. The L1 tree for Aaur was highly structured with well-supported nodes.

Once families were designated, a consensus sequence was generated for each family and re-aligned to the family. The alignment with consensus was then imported into MEGA for pairwise divergences. The pairwise divergences from family consensus ranges from 0 to 24%. The average divergence from the family consensus was 6.2%.
Figure 16. Maximum likelihood phylogenetic tree for the LINE/L1 clade in *Anolis auratus*, based on protein sequences. 500 bootstraps were executed.
A total of 33 protein sequences for CR1 were collected from the Aaur genome assembly. The Aaur CR1 clade is comprised of two families. The first has a high abundance of elements that are closely related, or have low divergence between them. The average pairwise divergence from family protein consensus for CR1 is 0.45%. Yet the divergence between the two families is 30.7%. As seen in Acar, the families for CR1 are highly divergent from each other, but within the family there is very low divergence. The low divergences in family (0.2 and 0.7%) may indicate recent or current activity of CR1 in Aaur.
Figure 17. A Maximum likelihood tree was constructed for LINE/CR1 elements for *Anolis auratus* with 500 bootstraps were executed.
LINE/RTE-1.

A total of 67 protein sequences for RTE-1 were collected from the Aaur genome assembly. Aaur has just one RTE-1 family. The average divergence from family consensus is 20.7%. 
Figure 18. A Maximum likelihood tree was constructed for LINE/RTE-1 proteins for *Anolis auratus* with 500 bootstraps were executed.
LINE/RTE-BovB.

A total of 54 protein sequences for RTE-BovB were collected from the Aaur genome assembly. Aaur has just one RTE-BovB family. The average divergence from family consensus is 13.8%.
Figure 19. A Maximum likelihood tree was constructed for LINE/RTE-BovB proteins for *Anolis auratus* with 500 bootstraps were executed.
**Anolis carolinensis**

*repeatmodeler.*

RepeatMasker was used to scan the Acar genome assembly for repeats. The query species was assumed to be *Anolis*. In Acar, nearly 26.8% of the genome is comprised of repeats. 18.9% of the Acar genome is comprised of retroelements, including 11.5% occupied by LINES and 3.1% occupied by SINES. Of that 11.5% occupied by LINES, 6.1% is comprised of CR1 and L2, 1.5% is occupied by RTE and 1.1% is occupied by L1. LTR elements occupy 4.3% of that 18.9% of retroelements that comprise the repeat landscape. DNA transposons occupy 5.2% of the repeat landscape. In addition, total interspersed repeats comprise 25% of the repeat landscape and simple repeats comprise 1.3%.

**collection of LINE elements.**

Acar LINE consensuses were collected from the Repbase repeat database. The DNA sequence was entered into NCBI’s ORF finder. I identified the reverse transcriptase (RT) domain of the ORF2 for the LINE sequences and used blastp to retrieve just the RT protein sequence. This RT consensus (for each respective LINE clade) was used to retrieve protein sequences in the Acar genome using the tblastn program. The protein sequences were then uploaded into Geneious and aligned using 8 iterations of the MUSCLE program. A total of 330 LINE elements were retrieved from protein blasts and used to construct phylogenetic trees for LINE/L1, LINE/CR1, LINE/RTE-1, and LINE/RTE-BovB.

**phylogenetic reconstruction.**
The alignments were exported and loaded into MEGA. Maximum likelihood trees were generated using 500 or more bootstraps. The maximum likelihood statistical method model was tested in MEGA for reliability, specifically for Acar. Gaps were treated using partial deletion with a 95% cutoff. The best model for Acar is the JTT+G+F, or the Jones-Taylor-Thornton model with a gamma distribution rate among sites.

**LINE/L1.**

A total of 104 protein sequences for LINE/L1 were collected from the Acar genome assembly. NJ and ML trees were generated for LINE/L1 elements found in the tblastn. In Acar, 20 distinct families were identified in the LINE/L1 clade, as found in previous studies. A consensus sequence was generated for each family and re-aligned to the family. The alignment with consensus was then imported into MEGA for pairwise divergences. The pairwise divergences within the families range from 0 to 17%. The average divergence from the family consensus was 3.4%.
Figure 20. A ML tree of Acar LINE/L1 clade. 500 bootstraps were executed.

LINE/CR1.

A total of 72 protein sequences for CR1 were collected from the Acar genome assembly. I designated five Acar CR1 families, although Novick et al., 2009 indicate there are only four families. The average pairwise divergence from family consensus is 3.4%. The divergences for each family range from 0 to 21.3%.
Figure 21. A Maximum likelihood tree was constructed for LINE/CR1 proteins for Acar. 500 bootstraps were executed.

LINE/RTE-1.
A total of 104 protein sequences for RTE-1 were collected from the Acar genome assembly. Acar had only one distinct LINE/RTE-1 family. RTE-1 had very low divergence from the family consensus at 6.3%.
**Figure 22.** A Maximum likelihood tree was constructed for LINE/RTE1 proteins for Acar. 500 bootstraps were executed.

LINE/RTE-BovB.

A total of 50 protein sequences for RTE-BovB were collected from the Acar genome assembly. Acar had just one distinct family of LINE/RTE-BovB. RTE-BovB had an average divergence from family consensus of 4.9%.
Figure 23. A Maximum likelihood tree was constructed for LINE/RTE-BovB proteins for Acar. 500 bootstraps were executed.

Anolis frenatus

repeatmodeler.
RepeatMasker was used to scan the Afre genome assembly for repeats. The query species was assumed to be *Anolis*. In Afre, nearly 9.7% of the genome is comprised of repeats. 7.3% of the Afre genome is comprised of retroelements, including 6.2% occupied by LINES and 0.2% occupied by SINES. Of that 6.2% occupied by LINES, 4.7% is comprised of CR1 and L2, 0.4% is occupied by RTE and 0.96% is occupied by L1. LTR elements occupy 0.85% of that 7.3% of retroelements that comprise the repeat landscape. DNA transposons occupy 0.6% of the repeat landscape. In addition, total interspersed repeats comprise 8.1% of the repeat landscape and simple repeats comprise 0.6%.

**collection of LINE elements.**

Acar LINE consensuses were collected from the Repbase repeat database. The DNA sequence was entered into NCBI’s ORF finder. I identified the reverse transcriptase (RT) domain of the ORF2 for the LINE sequences and used blastp to retrieve just the RT protein sequence. This RT consensus (for each respective LINE clade) was used to retrieve protein sequences in the Afre genome using the blastn program. The protein sequences were then uploaded into Geneious and aligned using 8 iterations of the MUSCLE program. A total of 333 LINE elements were retrieved from protein blasts and used to construct phylogenetic trees for LINE/L1, LINE/CR1, LINE/RTE-1, and LINE/RTE-BovB.

**phylogenetic reconstruction.**

The alignments were exported and loaded into MEGA. Maximum likelihood trees were generated using 500 or more bootstraps. The maximum likelihood statistical method model was tested in MEGA for reliability, specifically for Afre. Gaps were treated using
partial deletion with a 95% cutoff. The best model for Afre is the JTT+G+F, or the Jones-Taylor-Thornton model with a gamma distribution rate among sites.

LINE/L1.

The LINE/L1 alignment included 169 protein sequences found through tblastn. NJ and ML trees were generated for LINE/L1 elements found in the tblastn. Input from both the NJ and ML trees were used to designate LINE families. The naming convention for LINE/L1 repeat families (i.e. L1_AF_1) in Afre followed Acar repeat family names in the Repbase database. In Afre, at least 26 distinct families were identified in the LINE/L1 clade. The L1 tree for Afre is highly structured with well-supported nodes.

Once families were designated, a consensus sequence was generated for each family and re-aligned to the family. The alignment with consensus was then imported into MEGA for pairwise divergences. The pairwise divergences within the families range from 0 to 50%. The average divergence from the family consensus was 13.7%.
Figure 24. Maximum likelihood phylogenetic tree for the LINE/L1 clade in Afre based on protein sequences. 500 bootstraps were executed.
Afre had one distinct CR1 family. The average divergence from family consensus was 15%.
Figure 25. A Maximum likelihood tree was constructed for LINE/CR1 elements for Afre.
500 bootstraps were executed.
LINE/RTE-1.

Afre RTE-1 had just one distinct family. The average divergence from family consensus was 15.7%.
Figure 26. A Neighbor-joining tree was constructed for LINE/RTE-1 proteins for Afre. 500 bootstraps were executed.
LINE/RTE-BovB.

Afre has just one distinct RTE-BovB families. The average divergence from family consensus is 110%.

*Figure 27.* A Maximum likelihood tree was constructed for LINE/RTE-BovB proteins for Afre. 500 bootstraps were executed.
comparison of LINES in four anoles.

LINE/L1.

Most of the consensuses for the L1 family are similar across species with the exception of Aapl. Aapl was a clear outlier, with its consensus sequences for L1 far more divergent and diversified than the other anole L1 families. As found in the Novick et al., 2009 publication on Acar L1 families, there are some much older L1 families that are far more divergence than other L1 families. Surprisingly, all twelve of the Aapl L1 family consensuses failed to cluster with other anoles, instead creating a highly divergent branch.
Figure 28. ML tree of LINE/L1 family consensus for all four anole species.

The Aapl L1 family consensuses clustered with zebrafish and Xenopus family consensus, indicating that Aapl L1 families may be more ancient than it’s other anole counterparts. A lack of newer, less divergent L1 families in Aapl suggests that there is not recent activity of L1.
Next, an ML tree was created with the L1 consensuses for all four anole species, as well as human, zebrafish and *Xenopus* outgroups. Again, Aapl clusters by itself and with *Xenopus*. Most of Aaur (AA), Acar (AC) and Afre (AF) L1 consensus sequences cluster together.
Figure 29. ML tree of LINE/L1 family consensus for all four anole species and outgroups.
**repeatmasker results.**

The RepeatMasker output was analyzed for all four anole species. Because of the differences in genome assembly and coverages, absolute numbers of repeat copies cannot be used for comparative purposes. However, the total percentage of each LINE family can be quantified relative to the total base pairs present in the genome assembly build. I used these percentages to compare LINE content between the three genomes that were assembled using next-gen sequences.

Repeat content was found to vary significantly between the four anole species, among major classes LINES. The RTE-BovB in Afre comprises 4.6% of the genome, whereas in Acar it is 1.54%. The CR1 element in Aapl comprises 4.6% of the genome, whereas Aaur is 7.71%. In terms of total LINES, Aaur has almost double the percent of LINES that Afre has. Afre has the lowest repeat content at 9.67%, while Acar has the highest at 30%.

Although absolute numbers cannot be quantitatively compared because of differences in the initial coverages of various genome assemblies, the graph below shows raw numbers of LINE elements. Aaur has the most coverage, around 87x, while Acar is just 7.1x.
A more accurate way to measure repeat abundance is by finding the percent of the genome that the repeats occupy. This may remove a potential quantitative bias present when comparing genome assemblies of different coverages. The RepeatModeler output generated this data for different classes and families of repeats.

The most striking difference was in the percent of genome that total repeats comprise. In Acar and Apre and Aapl, the percentage of the genome occupied by repeats is 9.67 and 14.4% respectively. The repeat composition between the four anole species ranges widely from roughly 10 to 25%. This was a surprising result, as variations
in repeat abundance were hypothesized to be very low. Acar and Aaur are again, most similar, as seen in the comparison of the average percent divergence from family consensus. Likewise, Afre and Aapl are most similar to each other as well.

![RepeatMasker Output](image)

*Figure 31.* Percent of Genome Occupied by Repeat Elements in four anole lizards.

Aapl and Afre were highly similar except for the SINE content, for which Afre had very few SINES. Acar and Aaur were also highly similar in repeat content.
Figure 32. LINE Content Percentages among four anole species

The total genome assembly length was also inspected with regards to the amount of bases repeat-masked. Aapl and Afre were the most similar again for both total length of genome and total bases repeat-masked. Acar and Aaur have greater total genome lengths and a greater percentage of bases repeat-masked. Whether or not the smaller repeat content in Aapl and Afre is due to reduced activity or inactivity of LINE clades needs to be determined in future studies. Novick et al., demonstrated that Acar has newer, active families and interestingly, there is a greater repeat content and total genome size in comparison to Afre and Aapl.
**pairwise divergences**

After making the maximum likelihood trees, then designating families based on several criteria with multiple forms of evidence to support, pairwise distances were calculated for each family. The average divergence from the family consensus sequence was calculated for LINE/L1 families. Aapl (AP) had an average divergence of 10% from the family consensus. Aaur (AA) has an average divergence of 6.2% from the consensus. Afre (AF) has an average divergence of 14% from the consensus. In contrast, to Acar whose average family divergence is 2.7% all three of our anole species demonstrate moderate levels of divergence within families for the LINE/L1 families.
The distributions for the average divergences from consensus also vary among the four species, with Acar and Aaur being similar and Aapl and Afre being similar. Aapl and Afre had more families with higher divergences from family consensuses. Divergences in Aapl and Afre tended to cluster between 10 to 30%, whereas Acar and Aaur had many more families that demonstrated low divergences, between 1 and 10%.

**Figure 34.** Comparison of the Average Family Divergence from Consensus for LINE/L1 elements
Aapl and Afre had a central tendency towards greater family divergences than the other two anoles.

Figure 35. Histogram of Numbers of Families by Average Percent Divergence from Family Consensus

In order to test statistical significance, the Mann-Whitney U test, or the Wilcoxon rank-sum test, was used to determine if the average divergences for L1 family members from the family consensus significantly differ. In each set of two-species comparisons,
the only pair that failed to demonstrate significant difference was Aapl and Afre, which had a p-value of 0.68. All the other comparisons had p-values much less than 0.0001.

Next, the Kolmogorov-Smirnov test was also conducted on each pair of species. Surprisingly, Aapl and Afre demonstrated high similarity. The distributions for the average divergences from consensus also vary among the 4 species, with Acar and Aaur being similar and Aapl and Afre being similar.

![KS-Test Comparison Percentile Plot](image)

*Figure 36. KS-Test with Aapl and Afre*

Across all the classes of repeats depicted above, there are pronounced differences in the repeat abundance. Among the three new anole genomes, 69 LINE/L1 families were identified, using a total of 547 protein sequences to generate the trees. After families
were identified, consensus sequences were generated. LINE/L1 protein consensus trees were built for Afre, Aapl and Aaur. 500-1,000 bootstraps were executed. A comparison tree for L1 was constructed using the consensus sequences for each L1 family, in each respective species. Human, zebrafish and *Xenopus* protein sequences were also included in the tree.

**LINE/CR1.**

In prior studies, Acar showed evidence of recent insertion (Novick, 2009; Janes et al., 2010). Aapl, Afre and Aaur were highly similar to the CR1 patterns found in Acar. The pairwise divergences of the CR1 family for three anoles were also calculated, using 500 bootstraps. The average pairwise divergence from the family protein consensus for was 15% for Afre, 30.7% for Aaur, 46.50% for Aapl and 3.2% for Acar. Aapl and Aaur have two CR1 families, Afre has one CR1 family and Acar has four CR1 families.

**LINE/RTE-1.**

In Acar, the RTE-1 family was found to have very low divergence (0.2%) for RTE-1 DNA sequences, which suggests a young age (Novick 2009). There are about 250 copies of RTE-1 in Acar (Novick 2009). The average pairwise divergence from the RTE-1 family protein consensus for was 15.7% for Afre, 21% for Aaur, 24.3% for Aapl and 6.3% for Acar. All four anoles have just one RTE-1 family. The high divergence rates suggests that RTE-1 has not been active recently in three species, but it is possible RTE-1 may have been recently active or is active in Acar.

**LINE/RTE-BovB.**

The average pairwise divergence from the RTE-BovB family protein consensus for was 110% for Afre, 13.6% for Aaur, 13.9% for Aapl and 4.9% for Acar. The high
rates of divergence for RTE-BovB for Aapl, Aaur, and Afre indicate that RTE-BovB is not likely to be active, whereas in Acar RTE-BovB may be active.

**DISCUSSION**

The two main metrics to compare LINES are abundance and diversity. LINE abundance demonstrates a wide range among the four anoles, both in raw numbers and percentages of the genome. The patterns of family diversification in the four anoles generally conformed to repeat element profiles characteristic of reptiles. As previously published, Acar’s genome is more similar to fish genomes than mammalian genomes. Aapl, afre and aaur also have similar patterns to Acar.

It is clear from several phylogenetic reconstructions that the patterns of diversification of LINES are significantly different among species in the *Anolis* clade. Acar was a clear outlier, with average family divergence from consensus between 0 and 5%, whereas the other three anoles had very high divergences between 15 and 40%. Acar, which is estimated to have appeared 2 mya (Tollis 2012), has the lowest rates of LINE family divergences and has more active families than the other three anoles investigated. Afre, which derives from the Dactyloa radiation (87 mya) has the highest LINE divergences and has lost LINE lineages present in other anoles. Average family divergences from consensuses in the LINE/L1 family vary widely from 0 to 50% among the four anole species. These differences in family divergence between species were confirmed by the Mann-Whitney U test. However, there was one pair of anoles that failed to demonstrate significance between rates of divergence within families. Aapl and Afre, had no significant family divergences for LINE/L1. I expected to see the most similarity
between Aapl and Aaur, which both arose from the Norops radiation, yet they have quite
different patterns of LINE evolution. Acar, from a Cuban radiation, and Aaur, from the
Norops radiation, are more similar.

For LINE/CR1, the average family divergences from consensuses within families
ranged from 0.45% to 15%, with Acar having the lowest divergences and Afre having the
highest divergence. Between families, the divergences ranged from 3.2 to 46.50%. For
LINE/RTE-1, the average family divergences from consensuses ranged from 6.3 to
24.3%, with Acar having the lowest divergences and Aapl having the highest divergence.
For LINE/RTE-BovB, the average family divergences from consensuses ranged from
64.9% to 110%, with Acar having the lowest divergences and Afre having the highest
divergence.

**Anolis LINE abundance**

Repeat content, in both raw numbers and percentages, vary widely between the
four anoles with a difference of 10 to 25% repeat content of the total genome. Total
family numbers of the LINE/L1 clade range from 16 to 29 families. LINE/L1 divergence
rates also vary considerably from 2.7% to 13.6%, with Acar having the lowest divergence
and Afre having the highest divergence. CR1 shows variation in family number amongst
the four anoles with Acar having at least four distinct families, and Aapl, Afre and Aaur
having at least two distinct families. Within the *Anolis* clade, there are several differences
in genomic composition that result from LINE activity. Acar was chosen by the Broad
Institute as the first non-avian reptile to be sequenced. Now that three other anole
genomes are available for comparison, it is apparent that Acar is not the best model to use
for genomic comparisons. With regards to repeat analyses, solely using Acar to make
inferences on retrotransposon dynamics would present a bias because Anolis repeat families are younger, less divergent and more active.

**Vertebrate repeat contents**

Afre and Aapl have a much lower percent of genome occupied by repeats, and are more similar to the turtle or chicken genomes. In contrast, Aaur and Acar have a much higher percent of genome occupied by repeats, and are more similar to crocodile, python and bovine genomes. Within Anolis, there is a wide range of differences in percent of repeats, such that two of the anoles are more similar to species of other vertebrate classes than members of the same genus. Anole LINE repeat content demonstrates interesting correlations to other vertebrate species. For instance, Acar and Aapl have fairly high repeat content, much like alligator, crocodile, python, fish and mammals. However, Aapl and Afre have much lower repeat content, more like avian and turtle genomes. Such a high variation within a genus indicates that genomic composition should not be generalized to represent an entire class of vertebrates based on the closest genome available. The mammalian genomes investigated so far have very similar repeat contents, around 40-50%. However, in reptiles, there is a much larger range of repeat content that is not easily characterized by categories such as snake, turtle, crocodile or lizards. In fact, Aaur and Acar are more similar to crocodile than Aapl and Afre. Afre is more similar to turtle than Acar and Aaur. Similar to the findings of Castoe et al., variation is quite high among related species. Python has a total repeat content of 21% whereas the copperhead has 45% (Castoe et al., 2011). Chicken has nearly double the repeats of zebra finch, at 4 and 8% respectively (Warren et al., 2010). Between Afre and Acar the repeat content varies from 9.7 to 26.8%, respectively.
Figure 37. Total Repeat Content by Vertebrate Species

As found in Acar, the new three anole genomes demonstrate LINE evolution patterns that are more similar to fish genomes than mammalian genomes. Fish have about 30 active L1 families, which have low copy numbers. The four anoles investigated have an average of twenty two L1 families, each with relatively low copy numbers. Whereas humans have very low diversity with one active L1 family, anoles have very high diversity with multiple active families. Humans also have a very high copy number of L1, whereas anoles have low copy numbers.
L1 families in fish tend to disappear from the genome, failing to become fixed (Furano et al., 2004), in contrast to humans where L1 members have amplified to hundreds of thousands of copies. Furano et al., hypothesize that LINE dynamics in fish result from an equilibrium between the decay and generation of new LINES. It is unclear whether copy number is under negative selection in anoles; however, in Drosophila and fish ectopic homologous recombinations are hypothesized to constrain transposon copy numbers (Furano et al., 2004). The Castoe et al., 2011 study revealed that snake genomes (as represented by the copperhead and python) have highly divergent repeat landscapes. Similarly, anoles within the same genus have divergent repeat landscapes, though not as dramatically as the snakes.

**interspersed repeats**

Interspersed repeats, which are likely to be decaying or truncated transposons, are scattered throughout the genome are in varying stages of decay (Hillier et al., 2004). There are differences in the interspersed repeat content between the four anole species. Consistent with the notion that Aapl and Afre have much older, less active LINE families, the repeat content for these two species is quite low, similar to Avians. Interspersed repeat content appears highly constrained in avian genomes. In the chicken genome, less than 9% is occupied by interspersed repeats, in contrast to mammalian genomes, which are roughly 45% interspersed repeats (Hillier et al., 2004). In zebra finch, 4% of the genome is interspersed repeats (Warren et al., 2010). Generally, avian genomes are very compact with a sparse repeat landscape. Randomness and genetic drift is likely to play a large role in shaping genomic composition even among closely related
species. However, future directions should aim to determine if the low interspersed repeat content in Aapl and Afre result from similar constraints in Aves.

![Percent of Interspersed Repeats in Vertebrate Genomes](image)

**Figure 38.** Percent of Interspersed Repeat Content in Vertebrate Genomes

**Anolis LINE diversity**

L1 family diversity in all four anoles is similar, when compared to vertebrates. *Anolis* shows similar patterns of diversification to fish genomes, unlike mammalian genomes. Afre and Aapl have a much lower percent of genome occupied by repeats, and are more similar to the turtle or chicken genomes. In contrast, Aaur and Acar have a
much higher percent of genome occupied by repeats, and are more similar to crocodile, python and bovine genomes. Within Anolis, there is a wide range of differences in percent of repeats, such that two of the anoles are more similar to species of other vertebrate classes than members of the same genus. This was an unexpected finding.

Anolis LINE activity

Differences in retrotansposon activity were found between the four anoles. Acar and Aaur show recent activities of L1, whereas Afre and Aapl L1 families are older, more divergent and demonstrate less recent activity, if any. This study was intended to explore repeat element evolution, by looking at one of the most dominant repeats, members of the LINE clades. L1, CR1, RTE-1, RTE-BovB were investigated. Phylogenetic reconstructions indicate that the patterns of diversification of LINES are significantly different among species in the Anolis clade. However, other classes of repeats can be used in a four-way comparison between the anole lizards. SINES, DNA transposons, LTR elements, helitrons and simple repeats may reveal different patterns in Anolis. Additionally, differences in genomic structure underlie changes in functionality.

adaptive or deleterious roles?

LINES appear to paradoxically be both deleterious and adaptive (Deininger et al., 2003; Gonzalez and Petroc 2009; Oliver and Greene, 2011). LINE insertions are acknowledged to have both negative and positive effects, as they can be large and cumbersome, disruptive and metabolically costly to the host. In Avians, there is a clear constraint on genome size, which is hypothesized to be a result of two factors: metabolic demands involved with flight and the low incidence of repeat elements (Burt, 2002; Hughes and Piontkivska 2005). Constraints on the proliferation of transposons in avian
genomes highlight the negative effects of LINES in vertebrate evolution. Yet, in most eukaryotic genomes, repeat content is twice as high as in birds (Hughes and Piontkivska 2005). The nearly-neutral theory (Ohta, 1998) offers an explanation for how LINES might be able to accumulate even if they are slightly deleterious to the host’s fitness. Studies on eukaryotic transposons have demonstrated that effective population size can affect the strength of purifying selection, and can also allow for the fixation of even slightly deleterious alleles due to genetic drift (Charlesworth, 2009; Tollis 2013).

However, the successful amplification and high copy number of LINES in most eukaryote genomes suggests some long-term benefits might result from transposon activity. The TE-Thrust hypothesis set forth by Oliver and Greene states that transposons are powerful facilitators or genomic evolution (Oliver and Greene, 2011). In primates, transposons have been identified to cause significant coding and regulatory changes, and many cases of the exonization of repeat-affected genes have been identified (Oliver and Greene, 2011). While LINE insertions may deleterious, transposon evolution may contribute to generation of novel regulatory elements and roles in the host species.

LINE insertions have been identified to have several deleterious effects, such as insertions into regions of high homologous recombination that disrupt crossing over, cause ectopic recombinations and chromosome breakage (Kazazian, 2004). In vivo in cultured cells, experiments have shown that 10% of L1 insertions are associated with large deletions of genomic DNA (Kazazian, 2004). Furthermore, while the human genome has 800,000 copies of L1, only 1 in 10,000 are full-length elements capable of transposition (Yang and Kazazian 2006). It has been suggested that regulatory
mechanisms might control transposition, thus limiting autonomous replicative capabilities.

Retrotransposons are subject to cellular regulation (Beauregard et al., 2009). RNAi mechanisms may also play a role in moderating the deleterious effects of LINEs. RNAi mechanisms have evolved to control the replication of genomic parasites (Aravin, Hannon and Brennecke, 2007). For instance, RNA interference (RNAi) mechanisms have been found in Drosophila, where cells produce piRNA homologous to the Penelope repeat element that combats transcription and facilitates degradation (Rozhkov et al., 2010). Experiments involving the suppression of Dicer have demonstrated that RNAi mechanisms greatly limit L1 transposition in humans (Yang and Kazazian 2006). Mechanisms to combat invasive transposition have been found including, piwi-argonaut proteins that silence transposons (Evgen’ev et al., 2013) and the DICER enzyme, which produces siRNA, has been found to actively control transcript abundance (Yang and Kazazian 2006). Mechanisms for controlling transposition might help to temper the deleterious effects of LINE propagation.

However, there are also potential long-term evolutionary benefits from transposon activity, such as the repair of double-stranded DNA breaks (Kazazian, 2004). Their poor processivity is also associated with 3’ transductions that move DNA coding domains and regulatory motifs to novel locations in the genome, giving rise to new genes (Kazazian, 2004; Yang and Kazazian, 2006). L1 transpositions can also create actively expressed “chimeric retrogenes” (Kazazian, 2004).

Repeat elements have more generally been recognized to affect several levels of organismal genomic regulation including insertions that act as developmental enhancers
(Bejerano et al., 2006; Sasaki et al., 2008; Franchini et al., 2011), increasing mRNA expression levels (Smith et al., 2008), modifying transcription factor regulatory networks (Wang et al., 2007; Bourque et al., 2008; Kunarso et al., 2010) and catalyzing the generation of new cell types (Lynch et al., 2011). While other studies have identified regulatory effects from transposon activity, in Anolis these studies have been limited. Future directions should include evaluating the impact of LINE insertions on the evolution of gene regulatory networks within the genus.

**underlying genetic differences and phenotype**

It has been established that LINES have causally created differences in genomic composition between four anole species. However, whether or not transposon-driven genetic shuffling underlies morphological changes has not been determined in this study. However, Oliver and Greene assert that transposon evolution is implicated with the creation of adaptive phenotypes and novel ecological niches (Oliver and Greene, 2011). In the Anolis adaptive radiation, which comprises over 400 species, much research has been focused on understanding the interaction between genetic, ecological and evolutionary factors that contribute to specialized ecomorphs and morphological diversity between species. While it has not been established that LINE-driven changes in genomic composition in Anolis translate to changes in phenotype, it is known that changes in gene regulation are important for phenotypic evolution to occur. Thus, future research should also explore the effect of transposons on phenotypic development.
REFERENCES


Franchini, L. F., López-Leal, R., Nasif, S., Beati, P., Gelman, D. M., Low, M. J., ... &


Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in
eukaryotes. Genetica, 115(1), 49-63.


interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene, 300(1), 189-194.


Quan Fang, Christopher Botka and Andrew Shedlock. Poster Presentation at the 3rd International Conference Genomic Impact of Eukaryotic Transposable Elements.


Tollis, M., & Boissinot, S. (2013). Lizards and LINEs: Selection and Demography Affect the Fate of L1 Retrotransposons in the Genome of the Green Anole (Anolis carolinensis). Genome biology and evolution, evt133.


