Using transposable elements as tools to better understand evolution at the genomic level

By

Roy N. Platt II

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Molecular Biology
in the Department of Biochemistry, Molecular Biology, Entomology, & Plant Pathology

Mississippi State, Mississippi

May 2014
Using transposable elements as tools to better understand evolution at the genomic level

By

Roy N. Platt II

Approved:

__________________________
David A. Ray
(Co-Major Professor)

__________________________
Federico G. Hoffmann
(Co-Major Professor)

__________________________
Daniel Peterson
(Committee Member)

__________________________
Andy D. Perkins
(Committee Member)

__________________________
Richard D. Stevens
(Committee Member)

__________________________
Din-Pow Ma
(Graduate Coordinator)

__________________________
George M. Hopper
Dean
College of Agriculture and Life Sciences
Transposable elements (TEs), also known as jumping genes, are DNA sequences capable of mobilizing and replicating within the genome. In mammals, it is not uncommon for 50% of the genome to be derived from TEs, yet they remain an underutilized tool for tracking evolutionary change. With the increasing number of publicly funded genome projects and affordable access to next-generation sequencing platforms, it is important to demonstrate the role TEs may play in helping us understand evolutionary patterns. The research presented herein utilizes TEs to investigate such patterns at the genomic, specific, and generic levels in three distinct ways. First at the genomic level, an analysis of the historical TE activity within the thirteen-lined ground squirrel (*Spermophilus tridecemlineatus*) shows that non-LTR retrotransposon activity has been declining for the past ~26 million years and appears to have ceased ~5 million years ago. Since most mammals, and all other rodents studied to date, have active TEs the extinction event in *S. tridecemlineatus* makes it a valuable model for understanding the factors driving TE activity and extinction. Second, we examined TEs as factors impacting genomic and species diversity. We found that DNA transposon insertions in
*Eptesicus fuscus*, appear to have been exapted as miRNAs. When placed within a phylogenetic context a burst of transposon-driven, miRNA origination and the vespertilionid species radiation occurred simultaneously ~30 million years ago. This observation implies that lineage specific TEs could generate lineage specific regulatory pathways, and consequently lineage specific phenotypic differences. Finally, we utilized TEs to investigate their phylogenetic potential at the level of genus. In particular a method was developed that identified, over 670 thousand Ves SINE insertions in seven species of *Myotis* for use in future phylogenetic studies. Our method was able to accurately identify insertions in taxa for which no reference genome was available and was confirmed using traditional PCR and Sanger sequencing methods. By identifying polymorphic Ves insertions, it may be possible to resolve the phylogeny of one of the largest species radiations in mammals.
DEDICATION

This dissertation is dedicated to my family, each of whom has sacrificed time, money, vacations, and/or anniversaries so that I can achieve my goals.
ACKNOWLEDGEMENTS

I would like to thank each of my lab mates for their support over the past several years, in particular Mike Vandewege, who has done an outstanding job keeping the lab compute cluster running. I would also like to thank my current committee members, Drs. David Ray, Federico Hoffmann, Richard Stevens, Andy Perkins, and Daniel Peterson for their patience and flexibility. Also, I would like to thank my past advisers, Robert Bradley, Llewellyn Densmore, and Robert Baker. Finally I would like to specifically thank my undergraduate adviser Tom Lee for saving me from the “Pre-Med doldrums,” and placing me along my current career path. Finally I would like to thank my wife, Jenn, for supporting me unconditionally throughout this entire process.
# TABLE OF CONTENTS

DEDICATION ................................................................................................................................. ii  
ACKNOWLEDGEMENTS .................................................................................................................. iii  
LIST OF TABLES ............................................................................................................................. vi  
LIST OF FIGURES ........................................................................................................................... vii  

CHAPTER

I. INTRODUCTION .......................................................................................................................... 1
References ........................................................................................................................................ 6

II. A NON-LTR RETROELEMENT EXTINCTION IN *SPERMOPHILUS TRIDECEMLINATUS* .......... 11

Introduction .................................................................................................................................. 11
Materials and Methods .................................................................................................................. 14
   Transposable Element Identification and Quantification ......................................................... 14
   Neutral Mutation Rate and TE Age Estimation ....................................................................... 16
   Identifying Recently Active Non-LTR Retrotransposons ..................................................... 17
Results ........................................................................................................................................... 18
   Repeat Identification and the TE Landscape .......................................................................... 18
   Mutation Rate and Age Estimation ......................................................................................... 19
   Are there Active L1s in the Squirrel genome? ....................................................................... 20
Discussion ..................................................................................................................................... 21
   An Independent Reduction in L1 Activity in *Spermophilus tridecemlineatus* ....................... 21
   Potential impacts of L1 loss in *Spermophilus* ................................................................. 22
   Possible factors for L1 loss in *Spermophilus* ................................................................. 23
Acknowledgements ..................................................................................................................... 26
References ...................................................................................................................................... 32

III. LARGE NUMBERS OF NOVEL MIRNAS ORIGINATE FROM DNA TRANSPOSONS AND ARE COINCIDENT WITH A LARGE SPECIES RADIATION IN BATS ........................................... 43

Introduction .................................................................................................................................. 43
Results and Discussion ..............................................................................................45
Genus *Eptesicus* experienced bursts of DNA transposon activity in the recent past .................................................................45
A large number of miRNAs emerged after the divergence of *Eptesicus*, dog, and horse ...............................................................46
Most recent *Eptesicus* miRNAs are derived from vespertilionid-specific DNA transposons .................................................................47
DNA transposon-derived miRNAs target genes expressed in *Eptesicus* testes ..............................................................................48
The introduction of DNA transposon-derived miRNAs coincides with the rapid diversification of vesper bats ..................................49
How might the burst of miRNAs have impacted vespertilionid bat diversity? ........................................................................................50
Conclusions ..................................................................................................................53
Materials and Methods ..............................................................................................54
TE landscape characterization in *Eptesicus* and *Myotis* .................................................54
Small RNA isolation and sequencing ........................................................................56
miRNA prediction and evolution .........................................................................57
De novo gene prediction in *Eptesicus* ........................................................................58
miRNA target prediction .........................................................................................59
Acknowledgements .................................................................................................59
References ...................................................................................................................64

IV. TARGETED CAPTURE OF VES SINE INSERTIONS FOR PHYLOGENETICS IN *MYOTIS* ...............................................................70
Introduction ................................................................................................................70
Materials and Methods ............................................................................................74
Specimens examined ...............................................................................................74
Library preparation and Ves enrichment ..................................................................75
Data processing .........................................................................................................77
Validation of Ves insertions via PCR and Sanger sequencing .................................78
Results and Discussion ............................................................................................79
Precision and accuracy of Ves identification ..........................................................80
Average sequence coverage of Ves .........................................................................82
Validation with PCR and Sanger Sequencing .........................................................83
Conclusions ...............................................................................................................83
Acknowledgments ....................................................................................................84
References ..................................................................................................................96

V. CONCLUSION ....................................................................................................103
References ................................................................................................................108
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Transposable element content in <em>Spermophilus tridecemlineatus</em>, <em>Mus musculus</em>, and <em>Rattus norvegicus</em>.</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Summary of miRNA data for <em>Eptesicus</em>, dog and horse.</td>
<td>60</td>
</tr>
<tr>
<td>4.1</td>
<td>Voucher ID and collecting locality for each specimen examined.</td>
<td>86</td>
</tr>
<tr>
<td>4.2</td>
<td>Nucleotide adapters used for Ves enriched library preparation.</td>
<td>87</td>
</tr>
<tr>
<td>4.3</td>
<td>PCR Primers used to verify Ves insertions.</td>
<td>89</td>
</tr>
<tr>
<td>4.4</td>
<td>Basic sequence statistics for each taxon during the data analysis.</td>
<td>90</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

2.1 Transposable element load in Rattus norvegicus, Mus musculus, and Spermophilus tridecemlineatus. .................................................................29

2.2 Phylogenetic analyses of ORF sequences show large numbers of mutation in Spermophilus tridecemlineatus compared to Rattus norvegicus and Mus musculus...............................................................30

2.3 A mammalian phylogeny documenting known non-LTR extinction events.................................................................31

3.1 Recent DNA transposon insertions are more common in Eptesicus and Myotis than in dog and horse.................................................................61

3.2 miRNA evolution along a mammalian phenogram.................................................................62

3.3 Recent transposable element history of Eptesicus fuscus. ..................................................63

4.1 Preparation and sequencing of Ves enriched libraries. ..................................................91

4.2 Precision of estimated Ves insertion sites within each species.................................92

4.3 Accuracy of estimated Ves insertion sites shared among species..........................93

4.4 Sequence statistics per Ves insertion. ........................................................................94

4.5 Results of Ves insertion panel on seven species of Myotis with three different loci. ..................................................................................95
CHAPTER I
INTRODUCTION

Transposable elements (TEs) are small segments of repetitive DNA that are capable of mobilizing and accumulating within a genome. Also known as mobile elements or jumping genes, TEs are classified based on their method of mobilization. Class I TEs, also known as retrotansposons, are reverse transcribed from an RNA-intermediate during the mobilization process (Luan et al., 1993). The major types of retrotansposons are the long interspersed nuclear elements (LINEs), short interspersed nuclear elements, and the long terminal repeat retrotansposons (LTR) or endogenous retroviruses (ERVs). Retrotansposons, in particular the L1 LINE superfamily and various SINE families are the most common mammalian TEs. Class II TEs, also known as DNA transposons, do not go through an RNA-intermediate phase during mobilization. Common mammalian DNA transposons are Tc1-Mariners, hATs, piggyBacs, and Helitrons. DNA transposons in mammals are rare when compared to retrotansposons (Lander et al., 2001; Waterston et al., 2002). For example, the human genome is over 41% (1.2 gigabases) derived from retrotansposon while only 2.8% (77.6 megabases) is derived from DNA transposons (Lander et al., 2001). In fact, with few exceptions (discussed below), DNA transposon activity in commonly studied mammalian genomes (Lander et al., 2001; Waterston et al., 2002; Pace and Feschotte, 2007) indicates a universal DNA transposon extinction ~40 million years ago (mya).
Within each class of TEs, elements are classified into two categories based on their ability to self-mobilize. Autonomous TEs encode all the machinery necessary for mobilization, while non-autonomous TEs utilize the enzymatic machinery from an autonomous partner. In mammals, generally only a very few elements are capable of mobilizing at any one time (Mills et al., 2007). Of the 599,000 LINE elements in the mouse genome (Waterston et al., 2002), it is believed that only 2,400 (Zemojtel et al., 2007) to 3,000 (DeBerardinis et al., 1998) of them are currently active.

Over time mutations, host-defense mechanisms (Yoder et al., 1997; Aravin et al., 2007; Houwing et al., 2007; Slotkin and Martienssen, 2007), and competition between TEs (Le Rouzic et al., 2007) cause turnover in active TEs. In addition some TEs are capable of being horizontally transferred between species (Sanchez-Gracia et al., 2005). These phenomena result in a highly dynamic portion of the genome that can vary greatly between closely related species (Hawkins et al., 2006; Li et al., 2009; McLain et al., 2012; Meyer et al., 2012), populations (Ray et al., 2005; Witherspoon et al., 2006) or even individuals (Batzinger and Deininger, 2002; Bennett et al., 2004). When one considers the amount of the genome derived from TEs and their stochastic nature, it follows that TEs could play a significant evolutionary role (Kidwell and Lisch, 2000; Deininger et al., 2003; Kazazian, 2004). This observation has driven the research described in this dissertation.

In Chapter II we examined the TE content of the thirteen-lined ground squirrel, *Spermophilus tridecemlineatus*. After identifying and annotating the TE portion of the genome, we found that non-LTR retrotransposons are less abundant than would be expected based on comparisons to the *Rattus norvegicus* (rat) and *Mus musculus* (mouse).
genomes. By comparing TE insertions and applying a neutral mutation rate, we were able to identify a decline in non-LTR retrotransposon activity that began 19-26 mya. This decline continued till about 4-5 mya, at which point it appears that non-LTR retrotransposon activity in *S. tridecemlineatus* ceased completely. This non-LTR retrotransposon extinction in *S. tridecemlineatus* is unique in that only four groups of mammals have been identified with TE extinction events; the sigmodontine rodents (Casavant et al., 2000; Grahn et al., 2005; Rinehart et al., 2005), the pteropodid bats (Cantrell et al., 2008), and *Ateles* spider monkeys (Boissinot et al., 2004).

In Chapter III we examined the role of DNA transposons in the generation of novel regulatory elements known as microRNAs (miRNAs). DNA transposon activity was believed to have been extinct for the past 40-50 my in mammals. Additional work within vespertilionid bats identified a recent, and extensive, wave of DNA transposon activity that called this observation into question (Pritham and Feschotte, 2007; Ray et al., 2007; Ray et al., 2008; Thomas et al., 2011; Pagán et al., 2012). With a more complete understanding of the DNA transposon landscape in vespertilionid bats, and the availability of small RNA sequence data from *Eptesicus fuscus* (big brown bat), we examined the relationship between TEs and miRNAs. More specifically, we compared miRNA origination and its relationship to TEs in a subset of laurasiatherian mammals to see if lineage specific TEs, in this case DNA transposons, lead to lineage specific miRNAs. We were able to verify that large numbers of DNA transposon derived miRNAs were acquired during the peak period of the vespertilionid radiation.

Finally, in Chapter IV we used recently developed sequencing technologies and protocols to identify large numbers of phylogenetically informative SINE insertions in
seven species of *Myotis*. Retrotransposon insertions are useful as phylogenetic markers and can resolve phylogenies that limited amounts of sequence data cannot (Okada, 1991; Hillis, 1999; Ray et al., 2006; Ray, 2007). The phylogeny of *Myotis* has been studied extensively, yet some relationships remain poorly resolved due to conflicting signals from morphology, geography, and DNA sequence data (Findley, 1972; Ruedi and Mayer, 2001; Stadelmann et al., 2007). This makes *Myotis* an excellent test case for the development of phylogenomic tools in a non-model organism. Unfortunately, identifying SINEs have seen limited use as a phylogenetic tool due to the difficult in identifying large numbers of insertions in taxa lacking a reference genome. By modifying a sequence-capture method used to identify Alu insertions in humans (Witherspoon et al., 2010; Witherspoon et al., 2013) we were able to identify over 796 thousand Ves SINE insertions in six new world and one old world *Myotis*. We tested the accuracy of our method for identifying insertion positions within an individual, and the ability to detect shared positions between closely related species. A handful of these loci were validated using wet-lab techniques showing that our method, could serve as the basis for large-scale phylogenomic data sets in the future.

TEs are major contributors of genomic variation; the necessary precursor for selection (Böhne et al., 2008). Yet, often times, repetitive sequences are excluded from genomic analyses. For example, a recent study of the minke whale genome made only three mentions of repetitive sequences or transposable elements (Yim et al., 2014). Meanwhile, more than 37% of the minke whale genome is comprised of repetitive elements. The work in this dissertation demonstrates that a better understanding of the
TEs within a genome often times leads to a more complete understanding of the organism itself.

Each chapter represents a standalone unit of research at various stages in the publication process. For this reason each chapter follows the formatting requirements for the publishing journal. Chapter II was published in 2011 in *Gene* and titled “A non-LTR retroelement extinction in *Spermophilus tridecemlineatus*”. Chapter III has been accepted for publication with revisions by *Molecular Biology and Evolution*. Revisions for this paper have been completed and we are awaiting a final decision on acceptance. The publication title for Chapter III is “Large numbers of putative miRNAs originate from DNA transposons and are coincident with a large species radiation in bats”.

Research on Chapter IV is ongoing. Our goal is to use the protocol developed in Chapter IV, and the resulting SINE insertion data, to build a robust phylogeny of the new world *Myotis*. Currently, we are planning to submit Chapter IV to *Genome Biology and Evolution*.
References


CHAPTER II
A NON-LTR RETROELEMENT EXTINCTION IN SPERMOPHILUS TRIDECEMLINATUS

Introduction

Transposable elements (TEs) are repetitive DNA sequences that accumulate in genomes via multiple mechanisms and are particularly powerful mutagens. For example, in the course of their mobilization, they may influence gene expression via the introduction/disruption of regulatory elements, exons, and splice junctions (Jurka, 1995; Speek, 2001; Nigumann et al., 2002; Kazazian, 2004; Peaston et al., 2004; Cordaux et al., 2006; Matlik et al., 2006; Babushok et al., 2007; Hasler et al., 2007). However, TEs need not be actively mobilizing to have an effect on genome structure. TE-mediated genome rearrangements through non-homologous recombination and chromosomal rearrangements are well-documented (Lim and Simmons, 1994; Gray, 2000; Batzer and Deininger, 2002; Lonnig and Saedler, 2002; Eichler and Sankoff, 2003). Deletions, duplications, inversions, translocations and chromosome breaks have all been linked to the presence of TEs in a variety of genomes (Weil and Wessler, 1993; Lim and Simmons, 1994; Mathiopoulos et al., 1998; Caceres et al., 1999; Gray, 2000; Zhang and Peterson, 2004).

TEs are classified into two major classes. DNA transposons, or Class II elements, mobilize via a DNA intermediate and are often described as using a “cut and paste”
mechanism in which they excise and relocate themselves within the genome.
Conversely, Class I elements, the retrotransposons, utilize a “copy and paste” mechanism
known as retrotransposition. Retrotransposition involves transcription of a
retrotransposon by an RNA polymerase and reintegration of the element into a novel
genomic location via reverse transcriptase (Kajikawa and Okada, 2002). Two types of
retrotransposons are common; Long Terminal Repeat (LTR) elements such as
endogenous retroviruses and non-LTR elements.

The non-LTR elements are subdivided into two groups, the Long INterspersed
Elements (LINEs) and Short INterspersed Elements (SINEs). LINEs are considered
autonomous elements, in that they can mobilize themselves using self-encoded enzymatic
machinery, while SINEs are non-autonomous elements that require the protein machinery
form a LINE counterpart for their own mobilization (Kajikawa and Okada, 2002). With
few exceptions (Pritham and Feschotte, 2007; Ray et al., 2008; Pagan et al., 2010;
Thomas et al., 2011) non-LTR retrotransposons (LINEs and SINEs) are the predominant
TEs in mammalian genomes. In nearly all mammals examined, a single lineage of LINE,
the LINE-1 (L1) superfamily, and/or a SINE counterpart(s) dominate the active
retrotransposon repertoire. As a result, the L1 superfamily has played a major role
shaping the mammalian genome including but not limited to X-chromosome inactivation
(Lyon, 1998; Bailey et al., 2000), double stranded DNA break repair (Morrish et al.,
2002), and coding exaptation (reviewed in Burns and Boeke, 2008).

Structurally, LINEs contain a 5” untranslated region (UTR), two open reading
frames (ORF1 and ORF2), a 3” UTR, and a poly(A) tail (Malik et al., 1999). Several of
these components are required for successful mobilization of any active LINE. For
example, an internal promoter in the 5’ UTR recruits RNA Pol II (Swergold, 1990) while ORF1 and ORF2 encode enzymes (RNA chaperones and endonuclease/reverse transcriptase, respectively) that are required for nuclear import and reincorporation into the genome (Mathias et al., 1991; Feng et al., 1996; Martin and Bushman, 2001).

Given the potential impact of TEs in general and LINE-1 elements in particular on mammalian genomes as well as the increasing availability of genomic sequence data and the development of powerful computational resources, we began a survey of available mammalian genomes for unique L1 activity. As part of our investigation into TEs in rodents we detected that L1 activity has been substantially reduced and possibly eliminated in the 13-lined ground squirrel, *Spermophilus tridecemlineatus*. At least three other independent examples of reduced or eliminated L1 activity have been previously identified. Convincing evidence exists for an L1 extinction event dating back 22 MY in pteropodid bats (Cantrell et al., 2008). In sigmodontine rodents, the L1 lineage is believed to have gone extinct within the past 9 MY (Casavant et al., 2000; Grahn et al., 2005; Rinehart et al., 2005). Finally, the spider monkey (*Ateles paniscus*) genome exhibits such an extreme reduction of L1 activity that an extinction event beginning 25 MYA is likely though it could not be confidently stated (Boissinot et al., 2004). As a result, an extinction/quiescence event in *S. tridecemlineatus* represents only the second independent L1 loss in rodents and only the fourth example in mammals as a whole. The *S. tridecemlineatus* genome therefore represents an important addition to our understanding of mammalian retrotransposon dynamics. Herein, we present evidence for this decrease in L1 activity within the context of other well-studied rodents, the model
organisms *Mus musculus* (mouse) and *Rattus norvegicus* (rat) as well as the human genome.

We note that Helgen et al. (2009) recently renamed *Spermophilus* to *Ictidomys*. However, for ease of comparison with other work recently published on the genome, in particular the analyses of Lindblad-Toh et al. (2011) we have chosen to retain the name *Spermophilus*.

**Materials and Methods**

**Transposable Element Identification and Quantification**

Many previously identified repetitive elements for *S. tridecemlineatus* are cataloged in the common transposable element repository, RepBase (Jurka et al., 2005). However, variations in program usage, annotation, and data sources presented the possibility that the TE landscape of *S. tridecemlineatus* might be incompletely represented. With this in mind, we conducted a *de novo* TE identification to complement the available Repbase library.

Repetitive sequences in one quarter (~450 Mb) of the *S. tridecemlineatus* early stage (2x) WGS assembly (Genbank accession number: AAQQ01000000) were identified de novo using RepeatScout (Price et al., 2005). RepeatMasker (Smit et al., 2004) searches were used to quantify copy number for each repeat, and those present in less than ten copies were removed from further analysis. Repeats consisting of low sequence complexity (satellite sequences) were also removed. Retrotransposons mobilize under a master gene model, in which only a few elements are capable of mobilizing at any one time (Deininger et al., 1992). The presumed master gene sequence can be inferred by comparing multiple progeny and creating a consensus sequence which
ignores mutations that occurred in the progeny sequences after insertion. To infer the master gene (consensus) sequences for each repeat, the filtered RepeatScout (Price et al., 2005) output was used to query the entire *S. tridecemlineatus* WGS using BLAST v2.2.23 (Altschul et al., 1997). Hits of at least 75 base pairs (bp) were extracted along with a minimum of 500 bp of flanking sequence using custom PERL scripts. Extracted sequences were aligned using MUSCLE v3.8.31 (Edgar, 2004) and from these alignments consensus sequences were reconstructed using a using a 50% majority rule. Full length elements were presumed only when single copy DNA was identifiable on the 5’ and 3’ end of the alignments. In cases where the full length of the consensus sequence had not been captured the process was repeated until single copy DNA sequence was identifiable at both ends. The resulting library of elements was then submitted to CENSOR (Kohany et al., 2006) to ascertain their identity with regard to previously classified elements in RepBase (Jurka et al., 2005). Each element identified in our analysis fell within the classification parameters of Wicker et al. (2007) therefore subsequent analyses used the existing RepBase classification and naming system.

Recently, a higher coverage assembly the genome was released (AGTP00000000) and all subsequent analyses utilized this version.

The TE complements of rat and mouse have been studied extensively (Cabot et al., 1997; Saxton and Martin, 1998; Hardies et al., 2000; Rebuzzini et al., 2009) and multiple TEs have been characterized for each. RepeatMasker tables were accessed via the UCSC genome browser (www.genome.ucsc.edu) for each species and used to compare their TE landscapes to that of *S. tridecemlineatus*. In addition, BLAST searches were used to identify and extract elements used in the distance based analyses below.
Neutral Mutation Rate and TE Age Estimation

The presence of multiple TE insertions with little nucleotide divergence from each other and from the master (consensus) element is considered evidence of recent activity (Ivics et al., 1997). To estimate the active periods for each TE family however, a neutral mutation rate is needed. Many studies have examined phylogenetic relationships within Sciuridae (Steppan et al., 1999; Herron et al., 2004; Steppan et al., 2004), but have not produced a robust neutral mutation rate specific to the *Spermophilus* lineage. To resolve this problem, we extracted and concatenated nine exons (9,069 bp) from the squirrel, rat, and mouse genomes using the ENSEMBL:55 database (Birney et al., 2004). Kimura 2-parameter (K2P; Kimura, 1980) values were calculated at third base synonymous sites using MEGA5 (Tamura et al., 2011). Fossil calibration dates of 16 and 75 MYA were used to date the *Mus – Rattus* (Horner et al., 2007; Huchon et al., 2007; Murphy et al., 2007), and sciurid-murid (Bininda-Emonds et al., 2007; Huchon et al., 2007; Murphy et al., 2007) divergences respectively. K2P divergence values between each TE insertion and its consensus were calculated based on sequence alignments generated using RepeatMasker (Smit et al., 2004; Pagan et al., 2010). By applying the neutral mutation rate to each K2P distance value we were able to estimate activity periods for selected families.

In mammals increased methylation of cytosine at CpG sites has been demonstrated to act as a regulatory mechanism to suppress gene expression and TE activity (Yoder et al., 1997; Xing et al., 2004). To determine the rate of CpG mutations in the *S. tridecemlineatus* genome, the rate of cytosine to thymine conversion at CpG sites was calculated compared to all other non-CpG mutations. Five hundred random
insertions from each squirrel-specific SINE and LINE subfamily were queried using PERL scripts developed by Xing et al. (2004).

**Identifying Recently Active Non-LTR Retrotransposons**

L1 transposition relies on transcription of intact ORF1 and 2 regions. To determine the number of potentially active L1s elements in the genome, ORFs 1 and 2 of the *S. tridecemlineatus* L1 subfamilies, L1-1_Str and L1-2_Str, were used as BLAST queries against the genome. Each hit was extracted and translated. Potentially functional sequences were defined *a priori* as those harboring appropriate start codons and containing a single stop codon within a window 10% of the expected range.

Most L1 copies incompletely insert resulting in 5” truncations that are non-functional at the moment of insertion. We could therefore infer the history of L1 activity by identifying intact L1 loci by assaying the more common truncated L1 insertions. We estimated the ages of truncated L1 elements by querying the *S. tridecemlineatus* WGS with 500 bp from the 3” end of L1-1_Str and L1-2_Str (excluding the poly-A tail). Potential dates for the most recent L1 transposition in *S. tridecemlineatus* were determined based on the average pairwise K2P distance values among the twenty-five most similar truncated elements after having removed hypermutable CpG dinucleotides as per Pagan et al. (2010).

Mobilization of SINEs relies on functional L1 elements (Kajikawa and Okada, 2002; Dewannieux and Heidmann, 2005). Thus, SINEs can serve as indirect markers for L1 activity because identification of recently active SINE families would refute the hypothesis that L1 expression has been subjected to extinction or reduction in activity and suggest other possibilities. For example, lowered L1 transposition rates might be due
to increased competition for the ORF protein machinery from SINEs. To test this hypothesis, recent SINE activity was also dated as described.

Finally, TE dynamics can be diagnosed using phylogenetic tools. As insertions age and accumulate mutations, branch lengths on a phenogram of representative insertions will increase. Younger elements, on the other hand, are expected to form polytomies with very short branch lengths (Grahn et al., 2005; Scott et al., 2006; Cantrell et al., 2008). We therefore compared L1 phenograms from taxa with confirmed recent activity (human, mouse, and rat (Cabot et al., 1997; Hardies et al., 2000; Rebuzzini et al., 2009)) to *S. tridecemlineatus*. Using BLAST, the 250 insertions most similar to the ORF2 consensus sequence were identified and extracted from the *S. tridecemlineatus* WGS. Two hundred and fifty insertions with the highest similarity to the respective ORF2 consensus sequences from human, rat and mouse were also identified. Finally, a single neighbor joining tree encompassing each of the four data sets was inferred with MEGA5 (Tamura et al., 2011) based on K2P distance values.

**Results**

**Repeat Identification and the TE Landscape**

Our *de novo* analysis recovered 98 of the 104 *S. tridecemlineatus* TEs present in Repbase. Each element was subsumed within subfamilies listed in Repbase based on the parameters proposed by Wicker et al. (2007) and we have adopted the names recognized by Repbase throughout the rest of this work. The genome of *S. tridecemlineatus* is dominated by L1 LINEs, proto-B1 and ID SINEs as well as spuma-like (ERV3) retroviruses (Table 2.1). This pattern is similar to *Mus* and *Rattus* with a few general exceptions. First, TEs are much less abundant in the squirrel genome than in the murid
rodents. In *S. tridecemlineatus* TEs comprise ~26.3% (608.9 Mb) of the genome compared to ~39.2 (1,072 Mb) and ~41.5% (1,030.5 Mb) in *Mus* and *Rattus*, respectively per our Repeatmasker runs and repeat tables from the genome browser. The calculation for *Rattus* and *Mus* are comparable to those presented previously (Gibbs et al., 2004). Next, beta-like retroviruses (ERV2) present in large numbers in the murid genomes are much less abundant in the squirrel genome. Finally, our data confirm the post-divergence expansion of proto-B1 SINEs in *S. tridecemlineatus* compared to the B1 SINE dominance in the murid genomes (Veniaminova et al., 2007a; Veniaminova et al., 2007b; Churakov et al., 2010).

**Mutation Rate and Age Estimation**

Mutation rate analysis revealed an average of 39.6% divergence at third base synonymous sites between murids and *S. tridecemlineatus*. This yielded a neutral mutation rate of ~2.64 X 10^-9 (0.264% per MY), only slightly higher than the average mammalian mutation rate of 2.22 X 10^-9 (0.222% per MY) (Kumar and Subramanian, 2002). Using this rate, we dated peak activity periods for twelve *S. tridecemlineatus* specific LINE and SINE families. CpG mutation rates were slightly elevated within SINE and LINE insertions 1.15 - 3.80X (mean = 2.1X). TE expansion profiles (Fig. 2.1) suggest an overall period of declining L1 activity in *S. tridecemlineatus* based on insertion divergence from the consensus element. Combined with the neutral mutation rate calculated above, this decline began around 26 MYA has continued to the present. This decline appears to have affected all TE classes.
Are there Active L1s in the Squirrel genome?

We estimated the number of potentially active L1 elements by identifying intact ORFs. Using tBLASTn searches, over 400,000 hits to ORF1 and ORF2 were identified and extracted. Of these only three insertions contained ORF sequences that exhibited a methionine start codon and whose stop codon fell within a 10% window of the expected position. Each of these insertions matched L1-1 ORF1, not ORF2. To confirm that intact ORF1 sequences were indeed contained within unviable elements, we identified multiple premature stop codons in the ORF2 portion of the corresponding L1 insertions.

To estimate the likely most recent mobilization periods for L1 and *Spermophilus* SINEs, five hundred bp from the 3” region of L1, excluding the poly-A tail, were used as queries for BLAST searches. The twenty-five best hits for L1-1 and L1-2 exhibited 2.8 ± 0.3% and 2.7 ± 0.3% sequence divergence at their 3” ends corresponding to ~5.3 ± 1.13 MY since their insertion. Compared to murid rodents, this suggests minimal activity at best. The average genetic distance within the 3” portion of L1 among the twenty-five most recent insertions in the murid genomes drastically lower average divergences are observed (K2P: *Rattus* 0.3 ± 0.1%; *Mus* 0%) indicating recent mobilization in those genomes. Recent activity in *Mus* is also supported by the presence of polymorphic L1 insertions in laboratory strains (Akagi et al., 2008). To secondarily query L1 activity we searched for evidence of recently active SINEs. The squirrel SINE STRID3 is the most homogeneous family and therefore likely to be the most recently active SINE in the genome. Analysis suggested an average genetic distance of 2.1% when comparing the twenty-five most similar insertions. This suggests that the most recent SINE activity was
centered ~4.0 ± 0.8 MYA, corresponding well with the period of most recent activity seen for L1-1_Str and L1-2_Str.

Phenograms of the 250 most similar L1 ORF2 sequences from S. tridecemlineatus, Rattus, Mus and Homo, of which the latter three are known to harbor active L1s, provide striking graphical support for our contention that L1 activity in ground squirrels has been dramatically curtailed (Fig. 2.2; Cabot et al., 1997; Hardies et al., 2000; Rebuzzini et al., 2009). The L1 ORF2 sequences for Mus and Rattus form one large polytomy for each species with very short terminal branches, indicating a large number of recent insertions from a single L1 subfamily. Human ORF2 sequences exhibit slightly more variation (K2P 1.7%), than Rattus and Mus but exhibit a similar pattern overall. ORF2 insertions in S. tridecemlineatus recover long terminal branches (14.9% divergence within S. tridecemlineatus ORF2) and multiple clades, indicating either a very high mutation rate or long periods of L1 quiescence. Given our estimate of the mutation rate calculated above, the former is unlikely.

Discussion

An Independent Reduction in L1 Activity in Spermophilus tridecemlineatus

The TE landscape of the S. tridecemlineatus genome is distinct from Mus and Rattus (Fig. 2.1 and Table 2.1). The oldest elements in all three taxa exhibit similar profiles, as would be expected given their common ancestry. However, after the split with the murid lineage, S. tridecemlineatus exhibits a decline in activity for all TE classes. Our analyses suggest that there is very little, if any, TE activity in the squirrel genome. A lack of DNA transposon activity is not unusual in mammals. With a few notable exceptions, DNA transposon activity in mammals ceased 40-50 MYA (Lander et
al., 2001; Waterston et al., 2002; Giordano et al., 2007; Mikkelsen et al., 2007; Pace et al., 2008; Ray et al., 2008; Zhao et al., 2009; Thomas et al., 2011). However, a lack of recent non-LTR retrotransposon activity is a rare observation. Two L1 extinction events had been identified previously in mammals (Fig. 2.3), the pteropodid extinction 22 MYA (Cantrell et al., 2008) and the sigmodontine extinction 9 MYA (Grahn et al., 2005). A reduction of L1 activity was identified in the spider monkey beginning within the last 25 MY (Boissinot et al., 2004). Our data suggest that between 19 and 26 MYA L1 activity begin declining in the proto-Spermophilus genome, a decline that has continued in S. tridecemlineatus and likely ceased altogether ~4-5 MYA.

**Potential impacts of L1 loss in Spermophilus**

The addition of S. tridecemlineatus to the list of mammalian genomes with little or no L1 activity represents an important fourth instance in furthering our understanding of the functional impact of these events. Three potential examples of the significance of this observation spring to mind. First, it is well known that LINEs and SINEs are substrates for homologous and non-homologous recombination (Deininger et al., 2003; Kazazian, 2004). As time passes after the cessation of activity, individual insertions diverge thereby reducing the risk of non-homologous recombination. Thus, we might now ask whether rates of LINE- and SINE-mediated recombination events are correspondingly lowered in the affected taxa. Second, in humans and other mammals, transposable elements play a substantial role in introducing regulatory sequences and pathways are derived from transposable elements (reviewed in Feschotte, 2008). Given that significant periods of time have passed since the LINE extinctions it may be possible to detect differences in these taxa with regard to the evolution of cis regulatory elements.
Unfortunately, the non-model status of these animals suggests a long road to a basic understanding of their regulatory pathways before this can be confidently addressed.

Third, the reduction and/or extinction of non-LTR activity raises interesting questions regarding the evolution of genome size in these taxa. For example, without the constant introduction of new sequences from the mobilization of LINEs and SINEs, has there been a corresponding decrease in genome size when compared with other mammals?

Taxonomically, the scope of the L1 extinctions described herein calls into question the general presumption of active non-LTR retrotransposon activity in mammalian genomes. While it is clear that most mammals examined to date exhibit L1 activity, it may be that L1 extinction/quiescence is more widespread than previously thought. Assuming the L1 reduction observed in *S. tridecemlineatus* is shared by other taxa in the lineage (Sciuridae: Xerini), these four independent extinction/quiescence events affect up to 675 species (Fig. 2.3), or around 12% of mammals (Wilson and Reeder, 2005). With increasing sequencing capacity the taxonomic sampling of mammalian genomes will improve and allow us to examine the non-LTR landscape of additional non-model taxa.

**Possible factors for L1 loss in Spermophilus**

An important question to ask is what drives the loss of retrotransposon activity in certain mammalian genomes? Our total knowledge of therian L1 extinction or quiescence events totals an N of four (Fig. 2.3). Therefore, an attempt to definitively answer this question based on current data would be premature. With such a small dataset, the evolutionary history of each species (or group) will be of utmost importance. However, certain scenarios can be considered.
Several studies (Charlesworth et al., 1994; Sanchez-Gracia et al., 2005) support an equilibrium model of transposon accumulation versus removal via selection; although this scenario has also been criticized (Le Rouzic et al., 2007). Many factors alter the equilibrium between TE accumulation and removal. These include but are not limited to: 1) genetic drift; 2) competition between TEs; and 3) evolution of host defense mechanisms. Genetic drift does not appear to be playing a role in the ground squirrel. For example, one could hypothesize an ancestral population of proto-squirrels that encountered a bottleneck, allowing genetic drift to drastically impact the number of active L1 elements in the genome. However, our observation that TE activity in the squirrel experienced a slow reduction over a prolonged period of time contradicts this scenario.

Much as their counterparts in macroscopic ecosystems, TE families contest for limited genomic resources, and this could lead to more efficient elements outcompeting their less efficient rivals (Casavant et al., 1998; Brookfield, 2005; Dewannieux and Heidmann, 2005; Veniaminova et al., 2007b; Venner et al., 2009). For example, in the rice rat, *Oryzomys palustris*, loss of L1 activity is correlated with an increase activity of the ERV MysTR (Cantrell et al., 2005) a pattern that is congruent in all the sigmodontines (Erickson et al., 2011). Such competition fails to explain the recent suppression of L1 activity in *S. tridecemlineatus*. Instead, our data suggest that there has been a reduction in TE activity in general and we detect no evidence of increased activity from a competing family of elements.

Finally, host defense mechanisms are powerful factors in limiting or decreasing TE activity in the genome. These mechanisms range from increased CpG methylation
(Yoder et al., 1997; Xing et al., 2004), epigenetic suppression (Slotkin and Martienssen, 2007), and small RNA silencing (Aravin et al., 2007). It has been suggested that increased CpG methylation evolved as a defense against genomic invaders (Yoder et al., 1997; Xing et al., 2004). Indeed, some taxa with high levels of TE activity exhibit high CpG mutation rates (~6x in humans, Xing et al., 2004; ~8x in bats, Ray et al., 2008). CpG mutations were only slightly elevated in *S. tridecemlineatus* TEs (2.1x) and this could be interpreted two ways. First, CpG mutation rates may have been much higher in the past, when TEs were more active. At that time, CpG mutations reduced TE activity, and the mutation rate is now returning to a lower equilibrium level. Alternatively, it is possible that CpG mutation rates never deviated greatly from their current levels. If increased CpG mutation rates played a role in the reduction of TE activity in *S. tridecemlineatus*, then it follows that CpG mutation rate in the past must have been much higher than current levels, a hypothesis that remains to be tested.

Another interesting hypothesis involves the presence of Piwi RNA processing genes (PIWIL1, PIWIL2, PIWIL3, PIWIL4) and may be worthy of further study. Piwi proteins and piRNAs (Piwi-interacting RNAs) are known to influence TE activity in mammals (O'Donnell and Boeke, 2007; Seto et al., 2007; Aravin and Bourc'his, 2008). Pteropodid bats and *S. tridecemlineatus* contain four Piwi genes and have both experienced drastic reductions in TE activity. By contrast, *Rattus, Mus,* and several other mammals lack PIWIL3 while harboring active retrotransposons (F. Hoffmann, unpublished data). Even more striking, *Myotis lucifugus* lacks two Piwi homologs, PIWIL1 and PIWIL3 (F. Hoffmann, pers. comm.), and exhibits DNA transposon activity unprecedented in any mammalian genome investigated to date (Pritham and Feschotte,
We might therefore ask if a more efficient piRNA system could have driven the reduction of TE activity in *S. tridecemlineatus*.

In conclusion, the reduction/extinction of L1 activity in *S. tridecemlineatus* cannot confidently attributed to any single mechanism described above. Indeed, it would be overly simplistic to attribute this phenomenon to a single factor. The complex interactions between TEs and host genomes undoubtedly have led to the diverse array of TE activity and distributions observed. TE diversity and quantity varies not only among species (Volff et al., 2000; Rebuzzini et al., 2009) but among individuals (Witherspoon et al., 2010), and temporally (Khan et al., 2006; Filatov et al., 2008) yet few mammals have experienced as drastic a reduction in L1 activity as *S. tridecemlineatus*. Our analysis revealed only three L1 insertions with an intact ORF1 and zero intact ORF2 sequences. By comparison, mouse has confirmed L1 activity and is estimated to contain 2,400 (Zemojtel et al., 2007) to 3,000 (DeBerardinis et al., 1998) active L1 copies. Regardless of the assumptions made, it is apparent that *S. tridecemlineatus* genome contains at most a fraction of the L1 activity found in other rodents. With this addition to the list of mammals having experienced such a significant reduction in L1 activity, we are now better equipped to identify factors contributing to TE suppression in mammals.

**Acknowledgements**

We thank the Broad Institute Genome Sequencing Platform and Genome Sequencing and Analysis Program, F. Di Palma and Kerstin Lindblad-Toh for making the early assembly of *S. tridecemlineatus* available. This work was supported by the National Science Foundation [MCB-0841821 and DEB-1020865]. Additional support was provided by the Institute for Genomics, Biocomputing and Biotechnology, the
College of Agriculture and Life Sciences (both at Mississippi State University) and the Eberly College of Arts and Sciences at West Virginia University. Constructive criticism of the manuscript was provided by F.G. Hoffmann, A. Chong, M.P. Ramakodi, M.W. Vandewege, J.D. Smith, A. Sharma, C. Thompson, C. Lavoie, and anonymous reviewers. S. DiFazio provided computational support.
Table 2.1  Transposable element content in *Spermophilus tridecemlineatus*, *Mus musculus*, and *Rattus norvegicus*.

<table>
<thead>
<tr>
<th>TE Classification</th>
<th><em>S. tridecemlineatus</em></th>
<th><em>M. musculus</em></th>
<th><em>R. norvegicus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% genome</td>
<td>Coverage (Mb)</td>
<td>% genome</td>
</tr>
<tr>
<td>DNA Transposons</td>
<td>1.2%</td>
<td>28.2</td>
<td>1.1%</td>
</tr>
<tr>
<td>LTR</td>
<td>6.4%</td>
<td>147.1</td>
<td>10.7%</td>
</tr>
<tr>
<td>LINE</td>
<td>10.7%</td>
<td>248.1</td>
<td>19.8%</td>
</tr>
<tr>
<td>CR1</td>
<td>0.4%</td>
<td>9.2</td>
<td>0.1%</td>
</tr>
<tr>
<td>L1</td>
<td>10.3%</td>
<td>238.6</td>
<td>19.4%</td>
</tr>
<tr>
<td>RTE</td>
<td>&lt;0.1%</td>
<td>0.3</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>L2</td>
<td>0%</td>
<td>0</td>
<td>0.4%</td>
</tr>
<tr>
<td>SINE</td>
<td>8%</td>
<td>185.5</td>
<td>7.6%</td>
</tr>
</tbody>
</table>

* Genome proportion and content for the major transposable element classes and LINE subclasses identified in the WGS of *Spermophilus tridecemlineatus* with comparisons to *Rattus* and *Mus* RepeatMasker tracts from the UCSC genome browser.
Figure 2.1  Transposable element load in *Rattus norvegicus*, *Mus musculus*, and *Spermophilus tridecemlineatus*.

Transposable element insertions quantified as percent of genome for different age classes of SINE, LINE, LTR, and DNA transposons. Elements for the *Spermophilus tridecemlineatus* genome were identified using RepeatMasker and a custom TE library. UCSC RepeatMasker tracks were used to quantify repeats in the *Mus musculus* and *Rattus norvegicus* genomes.
Figure 2.2  Phylogenetic analyses of ORF sequences show large numbers of mutation in *Spermophilus tridecemlineatus* compared to *Rattus norvegicus* and *Mus musculus*.

Results of phylogenetic analyses of ORF2 sequences from the 250 most similar (and therefore most likely to be recent) L1 insertions in *Mus musculus*, *Rattus norvegicus*, *Homo sapiens*, and *Spermophilus tridecemlineatus* demonstrating the large inter-element divergences for even the most similar L1 insertions in *S. tridecemlineatus*. 


Figure 2.3  A mammalian phylogeny documenting known non-LTR extinction events.

A general mammalian phylogeny of mammals compiled from several sources (Steppan et al., 2004; Teeling et al., 2005; Perelman et al., 2011; McCormack et al., 2012). Currently, two L1 extinctions (marked with an “X”) and a quiescence (marked with a “/”) event have been convincingly demonstrated in mammals. Our analysis indicates that a similar event (!) has occurred in the *S. tridecemlineatus* genome.
References


Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P.,
Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E.,
Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B.,
Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown,
S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S.,
Chioromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins,
F.S., Cook, L.A., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V.,
Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis,
E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn,
Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S.,
Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N.,
Goodstadt, L., Graffham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R.,
Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs,
A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I.,
E.K., et al., 2002. Initial sequencing and comparative analysis of the mouse

Ds elements is caused by aberrant transposition. Plant Cell 5, 515-522.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A.,
Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman,
Nat Rev Genet 8, 973-982.


Mobile element scanning (ME-Scan) by targeted high-throughput sequencing.
BMC Genomics 11, 410.

element mutation spectra: molecular clocks and the effect of DNA methylation.
Journal of Molecular Biology 344, 675-682.


Exonization of active mouse L1s: a driver of transcriptome evolution? BMC
Genomics 8, 392.

Zhang, J. and Peterson, T., 2004. Transposition of reversed Ac element ends generates
CHAPTER III
LARGE NUMBERS OF NOVEL MIRNAS ORIGINATE FROM DNA TRANSPOSONS AND ARE COINCIDENT WITH A LARGE SPECIES RADIATION IN BATS

Introduction

MicroRNAs (miRNAs) are short, ~ 22 nucleotides long, noncoding RNAs that act as post-transcriptional regulators of gene expression (Bartel, 2004; Chen and Rajewsky, 2007; Carthew and Sontheimer, 2009) and changes in miRNA repertoires have great potential to generate evolutionary novelties. Expansions in miRNA repertoires have been linked to major innovations in the evolution of vertebrates (Heimberg et al., 2008). Many miRNAs arise from transposable elements (TEs) and in mammals several TE-derived miRNAs have been identified (Piriyapongsa and Jordan, 2007; Piriyapongsa et al., 2007; Devor et al., 2009; Ahn et al., 2013). miRNAs originate from transcribed hairpins (He and Hannon, 2004; Chen and Rajewsky, 2007) suggesting that DNA sequences tending to form such structures would be a valuable source of novel miRNA loci. Class II TEs, i.e. DNA transposons, commonly harbor palindromic sequences that form such hairpins (Feschotte et al., 2002; Sijen and Plasterk, 2003; Ahn et al., 2013). Thus, fortuitous transcription initiated by nearby promoters would be all that is needed to generate pri-miRNAs from DNA transposon insertions. Given that these elements are high in copy number, they represent a particularly robust potential source of novel miRNAs, a
hypothesis supported by the observation that miRNAs derived from ancient DNA transpositions are overrepresented in the human genome (Piriyapongsa et al., 2007).

Mammals in general have not experienced DNA transposon activity in the recent past (Pace and Feschotte, 2007; Hellen and Brookfield, 2013) and instead, retrotransposons (Class I TEs) dominate. This has disallowed direct, detailed study of ongoing miRNA origination via DNA transposon deposition. Vesper bats are unique among mammals in that they have experienced a substantial increase in DNA transposon activity, led by the Helitron superfamily. This shift was first observed in the genome of *Myotis lucifugus*, the little brown bat, where multiple DNA transposon families are present in high copy numbers and were deposited within the last 40 million years (Pritham and Feschotte, 2007; Ray et al., 2007; Pace et al., 2008; Ray et al., 2008; Thomas et al., 2011; Pagan et al., 2012).

Why and how DNA transposon activity has increased so substantially in vesper bats is still unknown. Introduction of the DNA transposons was likely via horizontal transfer (Pace et al., 2008; Gilbert et al., 2010; Pagan et al., 2010; Thomas et al., 2011) but a direct source of the transfer has yet to be identified. DNA transposon families have been identified in other vespertilionids (Thomas et al., 2011; Pagan et al., 2012) indicating that the expansions of DNA transposons can be traced to a common ancestor. Our knowledge of TEs in non-vesper bats is limited but informative. In the Pteropodidae (a megabat family), nearly all published work has focused on the retrotransposon extinction seen in several genera (Cantrell et al., 2008; Gogolevsky et al., 2009). Our own studies of *Pteropus vampyrus* (Pteropodidae) and studies from other researchers in the field (Ray et al. unpublished, Feschotte et al. pers. comm.) suggest that there has been
no recent activity from any TE class. Genomic survey sequencing in *Artibeus lituratus* suggests a TE history in Phyllostomidae (a microbat family more closely related to Vespertilionidae) that is more typical of mammals, i.e. extensive retrotransposon activity and little to no DNA transposon activity) (Pagan et al., 2012 and Ray et al. unpublished data). Furthermore, reports by Thomas et al. (2011) and Ray et al. (2008) suggest that no DNA transposon activity had occurred in a close relative of the vespert bats, the family Miniopteraidae.

Given that Class II TEs are disproportionately likely to harbor secondary structures characteristic of miRNA loci (Feschotte et al., 2002; Sijen and Plasterk, 2003; Ahn et al., 2013) and that this single clade of bats has experienced a recent expansion of DNA-based TEs in their genome, the genomes of vespert bats provide a natural system to explore the potential of DNA transposons as sources of novel miRNAs. We hypothesized that the increase in DNA transposon activity could have resulted in an upsurge in miRNA origination. To test this hypothesis, we compared patterns of TE activity and miRNA birth among *Eptesicus fuscus*, a vespertilionid bat whose genome was recently sequenced, horse and dog. More specifically, we compared the TE landscapes and miRNA repertoires in each genome, estimated miRNA origination rates in each lineage, discriminating between DNA transposon, retrotransposon and non-TE derived miRNAs, and compared the observed patterns among species.

**Results and Discussion**

**Genus Eptesicus experienced bursts of DNA transposon activity in the recent past**

Our first step was to compare patterns of TE activity among *Eptesicus*, horse and dog. To this end, we performed a de novo characterization of the TE landscape using the
current genome draft of *Eptesicus* to verify that its genome has experienced similar bursts of Class II transposon activity when compared to *Myotis*. Our results confirm that over the past 40 my, around half of the TE content of the two bat genomes is derived from Class II transposable elements compared to only minor fractions in dog and horse (Fig. 3.1). As expected given that DNA transposons invaded a common ancestor of *Eptesicus* and *Myotis*, many DNA transposon families identified in *Myotis* are also present in *Eptesicus*. Finally, we identified multiple TE families found in *Eptesicus* but were missing or at very low copy number in the genome of *Myotis* (and vice versa), confirming that both lineages have experienced their own unique set of Class II TE expansions after their divergence from a common ancestor (Fig. 3.1).

**A large number of miRNAs emerged after the divergence of *Eptesicus*, dog, and horse**

A previous study of human miRNAs uncovered a burst of miRNA origination shortly before the split between Laurasiatheria and Euarchontoglires (Iwama et al., 2013). Accordingly, the genomes of *Eptesicus*, dog and horse, all members of Laurasiatheria, would be expected to share several miRNAs. Through comparison with 19 other mammal genomes, we identified 224, 345, and 382 miRNAs in *Eptesicus*, dog and horse respectively, that originated prior to their most recent common ancestor (Fig. 3.2). The lower values for the bat are to be expected as we are only including miRNAs identified during our analysis compared to all available sources for the dog and horse. Further analyses reveal that a large proportion of expressed miRNAs arose after the divergence of all three taxa from a common ancestor ~82 mya (Murphy et al., 2007; Meredith et al.,
2011), 60% (396 out of 661), 34% (184 out of 535), and 47% (339 out of 729) of the total miRNAs are unique to the lineages of *Eptesicus*, dog and horse, respectively (Table 3.1).

**Most recent *Eptesicus* miRNAs are derived from vespertilionid-specific DNA transposons**

Even though TEs are an important source of novel miRNAs in all three species analyzed, *Eptesicus* stands out in two respects. Close to two thirds (61.1%) of the 396 post-divergence miRNAs in *Eptesicus* are derived from TEs compared to only 23.9% and 16.2% in the dog and horse, respectively. In dog and horse the sources of these TE-derived miRNAs are also distinct. The majority of TE-derived miRNAs, 90.9% and 65.4% are associated with retrotansposons, which are the dominant TEs in the genome of dog and horse. In *Eptesicus*, however, 58.7% of the TE-derived and 35.9% of all miRNAs arose from DNA transposons (Table 3.1). When one considers that DNA transposons comprise only 10.3% of the *Eptesicus* genome compared to the 16.5% that is derived from retrotansposons, it is clear that DNA transposon-derived miRNAs are overrepresented. Yet, analysis at finer scales reveals even more interesting patterns.

Overall rates of miRNA origination are of similar magnitude when comparing non-TE-derived miRNA in *Eptesicus*, dog and horse (Table 3.1, A – *Eptesicus*, A – Dog, and A - Horse). However, the TE-derived miRNA origination rate in *Eptesicus* is over four times higher than either horse or dog, and the rate of DNA-transposon derived miRNA birth is between 34 and 8.5 times higher in *Eptesicus* than dog or horse, respectively. Even more notable, high rates of miRNA origination map to branches with DNA transposon activity. A conservative estimate based on data derived only from available genome drafts limits DNA transposon activity to after the divergence of
Pteropus from the other bats (Figure 3.2, branch C - Eptesicus). The overall miRNA origination rate along this ~66 my period is ~5.8 miRNAs/my, with many being derived from DNA transposons (~2.2 miRNAs/my).

However, because it is unlikely that DNA transposons were active in either Artibeus or Miniopterus, more closely related bats, a more realistic rate of DNA transposon-derived miRNAs can be calculated. In particular, after the divergence from Miniopterus ~46 mya (Fig. 3.2, node D, we can surmise that DNA transposon-derived miRNAs arose in the lineage leading to Eptesicus (D – Eptesicus) at an increased rate of ~3.1 miRNAs/my. Notably, most miRNAs we identified were present in both Myotis and Eptesicus, indicating that they arose in a common ancestor and narrowing their origin to a window of ~21 my (Fig. 3.2, branch D – E). We estimate that during this period the miRNA origination rate reached a peak and that Class II-derived miRNAs were arising at ~5 miRNAs/my.

A peak of Class II TE activity in Eptesicus and Myotis occurred during this same window, at ~26 mya, just prior to the estimated vespertilionid divergence. After the Eptesicus/Myotis divergence TE activity in general has been declining in Eptesicus and relatively few DNA transposon-derived miRNAs localize to that terminal branch (~1.4 miRNAs/my), leading us to conclude that the rise and fall of the miRNA origination rate in Eptesicus is explained by the rise and fall of Class II TEs leading up to and since the vespertilionid radiation.

**DNA transposon-derived miRNAs target genes expressed in Eptesicus testes**

Mammalian miRNAs often regulate mRNA expression through targeted binding of the 3’ untranslated region (UTR; Lai, 2002). To investigate the potential regulatory
roles of testes miRNAs, we identified candidate targets by mapping miRNAs to the annotated *Eptesicus* genome draft. Subsequently, we identified potential target genes that were expressed in the testes using gene predictions as described in the Materials and Methods. Forty-six miRNAs appear to target genes expressed in testes. We then grouped the miRNAs into categories based on their age. Four expressed genes appear to be targeted by miRNAs that arose prior to the Boreotherian divergence. Three of these are involved in signal transduction (GPR135, GNG10 and TAC3) while the fourth (BCO2) mediates beta-carotene oxidation. Only four expressed miRNAs with identifiable targets arose after the Boreotherian split, but prior to the divergence of Chiroptera. Two (PVRL3 and TNP1) play a direct role in spermatogenesis, while the others (TOP1 and RECQL4) function in telomere maintenance.

Using DAVID (Huang et al., 2009) to categorize the 38 expressed targets of miRNAs unique to vesper bats reveals enrichment for two ontological terms: transcription (22 genes, P value = 6.8E-6, FDR = 0.01) and ubiquitin dependent protein catabolic processes (seven genes, P value = 1E-4, FDR = 0.1). Of the 22 genes affecting transcription, eleven are targeted by miRNAs derived from DNA transposons, while two of the seven genes involved in ubiquitin dependent processes arise from DNA transposons. The remaining miRNAs are derived from retrotransposons or non-TE derived sequences.

The introduction of DNA transposon-derived miRNAs coincides with the rapid diversification of vesper bats

Vespertilionidae radiated rapidly into ~400 species (and ~48 genera) after splitting from their close relatives, Cistugidae, ~34 mya (Miller-Butterworth et al., 2007;
Lack and Van Den Bussche, 2010; Meredith et al., 2011) making it the second most species rich mammalian family (Simmons, 2005). Lineage-through-time plots indicate an increased rate of diversification (Lack and Van Den Bussche, 2010) when comparing vespertilionid bats to the phyllostomid bats, which accumulated lineages more gradually but did not experience a similar Class II expansion (Thomas et al., 2011; Pagan et al., 2012). In fact, the increased rate of cladogenesis is most apparent between 20 and 30 mya, a period that coincides with the massive initial expansion and eventual peak of DNA transposon, primarily Helitron, activity (Fig. 3.3). Of the top 20 TE families that have contributed testis miRNAs, twelve are DNA transposons and ten of these belong to the Helitron superfamily. Of the top ten, five of the top miRNA sources are Helitrons. These data strongly suggest that the unique DNA transposon activity observed in vespertilionid bats has contributed uniquely and significantly to the repertoire of small regulatory RNAs.

**How might the burst of miRNAs have impacted vespertilionid bat diversity?**

Multiple authors have suggested that TEs may impact diversification (Furano et al., 1994; Jurka et al., 2011; Oliver and Greene, 2011; Jurka et al., 2012; Oliver and Greene, 2012), and in fact, a connection between DNA transposon activity and early primate evolution has been previously noted (Pace and Feschotte, 2007; Zeh et al., 2009; Oliver and Greene, 2011). Additional authors have suggested a link between the unique TE activity in vespertilionid bats and their rapid diversification (Pritham and Feschotte, 2007; Ray et al., 2008; Zeh et al., 2009; Thomas et al., 2011; Oliver and Greene, 2012; Mitra et al., 2013) but no mechanism has yet been supported by experimental data.

We note the following observations: 1) vespertilionid bats experienced increased rates of diversification centering 20-30 mya, 2) the same bats experienced a massive increase in
lineage-specific DNA transposon activity during the same period, 3) a large proportion of miRNAs are derived from the lineage-specific DNA transposons, and 4) those miRNAs are enriched for terms related to transcriptional regulation. We also note one additional feature in the history of these bats and their evolution. Approximately 33-34 mya there was a rapid and significant shift in Earth’s climate, the Eocene-Oligocene transition. It was during this transition that the climate changed from warm tropical conditions to a more temperate regime, with ice sheets covering the poles (Zanazzi et al., 2007; Liu et al., 2009). Organisms with plastic genomes and the ability to rapidly evolve to fit the changing climate would have been better prepared to respond to such challenges.

We show that the genome of the ancestral vespertilionid was successfully invaded by Class II TEs just prior to and during this period and that the invasion resulted in the introduction of a large number of miRNAs with the potential to influence regulatory pathways. Given the importance of transcriptional regulation to evolution and the number of transcription regulators that appear to be targeted by DNA transposon derived miRNAs, it is reasonable to hypothesize that the recently accumulated miRNAs led to changes in gene regulation that were acquired in a lineage specific manner and that could influence the adaptive radiation of vespertilionid bats.

All this being said, we must point out that vespertilionid bats are behaviorally, morphologically and karyotypically highly conserved (Simpson, 1944; Corbet and Hill, 1991; Fitch and Ayala, 1994; Koopman, 1994; Nowak, 1999; Simmons, 2005) and it is this characteristic homogeneity that has made relationships within the family difficult to resolve (Lack and Van Den Bussche, 2010). The same authors suggest that vespertilionid bats diversified and dispersed rapidly, obviating the need for morphological or behavioral
specialization. Such a scenario is supported by a recent study of South American *Myotis* from Larson et al. (2011). Prior to that study, 15 species of South American *Myotis* were recognized. Using nuclear and mitochondrial markers, the authors suggest that the actual number of species may be as high as 34. One „species”, *M. nigricans*, exhibited at least 12 species-level lineages (i.e. cytochrome-b variation greater than 5%) and many of these were paraphyletic, suggesting that a single group of essentially identical specimens actually represents multiple cryptic species.

By contrast, the phyllostomid bats are one of the most morphologically diverse mammalian families, but contain comparable numbers of lineages (genera) to Vespertilionidae, There are many aspects to phenotypic diversity beyond morphology, however. In fact, gene expression is often considered a quantitative trait on par with morphology (Cheung and Spielman, 2002). In applying this understanding, it is difficult to say that one family of bats is more phenotypically diverse than another. Due to the complex and rapid acquisition and expression of novel TE-derived miRNAs in vespertilionids, we hypothesize that the opportunity for genetic incompatibilities (isolating mechanisms) was increased. As the opportunity for such incompatibilities increased, species diversification would likely follow but might not have been accompanied by morphological or behavioral variation. Indeed, the lack of morphological diversity in spite of increased species diversity seems to implicate a significant role for genetic rather than morphological or behavioral isolating mechanisms in vespertilionids.
Conclusions

The expansion of novel miRNAs via DNA transposon activity proposed here provides several avenues for future research and a number of testable hypotheses. For example, there has been a recent increase in activity by hAT and piggyBac DNA transposons in *Myotis* (2007; Pace et al., 2008; Ray et al., 2008; Mitra et al., 2013). The activity of these two families is independent of the peak of Helitron activity shared by all vesper bats and is not found in *Eptesicus*. Stadelmann et al. (2007) found that genus *Myotis* experienced a species radiation that includes over 100 species during the last 10-15 my and this coincides with the rapid rise in hAT activity (Ray et al., 2008). By contrast *Eptesicus* consists of only 23 species having emerged over the same period (Roehrs et al., 2010). If our hypotheses are correct, one would expect to find evidence of an increase in hAT-derived miRNAs over the same time. Furthermore, we would expect these miRNA to be limited to genus *Myotis* and not be shared by other vesper bats. Mitra et al. (2013) point out that a single New World *Myotis* clade has radiated very recently, within the last 1-2 my. This radiation is coincident with the expansion of a piggyBac family (*piggyBat*) and evidence to date suggests that *piggyBat* is restricted to New World members of the genus (Ray et al., 2008; Mitra et al., 2013). Like most DNA elements piggyBacs and hATs contain palindromes forming hairpins that could lead to novel miRNAs.

All of these scenarios fit well with the recently proposed TE-Thrust hypothesis of Oliver and Greene (Oliver and Greene, 2011; Oliver and Greene, 2012), which suggests that “[l]ineages with an abundant and suitable repertoire of TEs have enhanced evolutionary potential and, if all else is equal, tend to be fecund, resulting in species-rich
“adaptive radiations.” Indeed, other eukaryotic clades that are more prone to DNA transposon activity (especially from Helitrons) may be have experienced phenomena similar to what we observe in Vespertilionidae. For example, the green anole (*Anolis*), which exhibits an incredible array of mostly young TEs including both retrotransposons and DNA transposons (Carmell et al., 2002; Novick et al., 2009; Novick et al., 2010; Alfoldi et al., 2011; Novick et al., 2011) is particularly diverse with over 400 recognized species having come about through rapid speciation. Future studies of *Anolis* might provide further insight into the relationship between DNA transposon activity and miRNA origitation as well as TE-Thrust in general.

The data presented here explain the acquisition of novel miRNAs via lineage specific TE activity and indicate a mechanism that may have impacted the radiation of a highly diverse mammalian clade. The expansion of DNA transposons and the resulting introduction of miRNA loci in the ancestral vespertilionid genome correspond with the rapid diversification of this group. The miRNAs that resulted could have provided the raw material to generate novel regulatory pathways, as evidenced by the targeting of expressed testis genes. Introduction of novel gene expression phenotypes could have promoted the rapid diversification of this family during a period of rapidly changing climate.

**Materials and Methods**

**TE landscape characterization in *Eptesicus* and *Myotis***

We characterized the TE landscape of *E. fuscus* (EptFus1.0, GenBank accession ALEH00000000) using RepeatModeler and a combination of existing tools and custom perl scripts (Smit and Hubley, 2008-2010, and available upon request; Smith and Ray,
The methods used were similar to those described previously (Alfoldi et al., 2011; Consortium, 2012). Briefly, RepeatModeler was used to identify potential TE family consensus sequences. We then used BLAST to query the entire WGS draft (Altschul et al., 1990). Hits of 100 bases or more were extracted along with up to 1000 bases of flanking sequence using custom perl scripts. MUSCLE (Edgar, 2004) was used to align the extracted hits with the consensus and 50% majority rules consensus sequences were constructed. We then examined the 5’ and 3’ ends of the alignments. Consensus sequences were considered „complete“ when single copy sequence could be identified at both ends of the alignment in the component sequences. If single copy sequences were not identifiable, the process was repeated. The resulting library of elements was submitted to CENSOR (Kohany et al., 2006), BLASTN and BLASTX to identify potential designations, which were confirmed through the identification of key sequence landmarks such as terminal inverted repeats, target site duplications, coding sequences, etc. All elements have been submitted to RepBase.

The TE landscape of *M. lucifugus* has been previously characterized (Pritham and Feschotte, 2007; Ray et al., 2007; Ray et al., 2008). We combined our novel library of *Eptesicus* TEs with elements previously identified in *Myotis* and other bats (Pritham and Feschotte, 2007; Ray et al., 2007; Pace et al., 2008; Ray et al., 2008; Thomas et al., 2011; Pagan et al., 2012) as well as elements characterized independently and available from RepBase. Using the combined library, the *Eptesicus* and *Myotis* genome drafts were analyzed using a locally implemented version of RepeatMasker to estimate the TE content of the genome and to generate genome-wide TE annotations.
The TE landscapes of dog and horse have been previously characterized as described at http://www.repeatmasker.org/species/canFam.html and http://www.repeatmasker.org/species/equCab.html. RepeatMasker output files were obtained from these sites.

To calculate approximate periods of activity for all taxa, we restricted ourselves to RepeatMasker hits that spanned at least 50% of any given consensus sequence. To estimate divergences, we used a modified version of the calcDivergenceFromAlign.pl script that is included in the RepeatMasker package to calculate Kimura 2-parameter distances between each insertion and its respective consensus (Pagan et al., 2010). The -noCpG option was invoked. We applied the mutation rate estimated by Ray et al. (2008), 2.366 x 10^{-9} substitutions per site/per million years.

Small RNA isolation and sequencing

A horizontal cross section from the testes of a wild caught E. fuscus was taken and tissues were snap-frozen in liquid nitrogen. Testis samples from dog and horse were collected as a byproduct of veterinarian preformed castrations. Total RNA was extracted in TRIzol following the manufacturers recommended protocol. Small RNAs were isolated, indexed, and prepared for sequencing using the Illumina TruSeq Small RNA Sample Preparation kit. Small RNAs were sequenced on a single Illumina HiSeq lane (1 x 50 nt reads). 3” adapter sequences were removed and reads less than 15 bp were discarded. Every base call within reads was required to have a Phred quality scores greater than 25. Reads falling short of this threshold were removed. Each of the above steps was accomplished via FASTX-toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/).
miRNA prediction and evolution

To detect miRNAs we used a local installation of the miRanalyzer package and completed all searches against EptFus1.0, canFam3.1, and the equCab2 assemblies for the *Eptesicus*, dog and horse genomes, respectively. The miRanalyzer platform is a multi-step approach to miRNA discovery. In the first step, reads were mapped to all known species-specific pre-miRNAs from the current release of miRBase (Release 19; (Kozomara and Griffiths-Jones, 2011)). After this comparison, all reads mapped to known pre-miRNAs were removed and the remaining reads were mapped to the *ab initio* transcriptome libraries from Ensembl (Release 70; Flicek et al., 2012) or to the Augustus-based gene predictions of the *E. fuscus* assembly. Any reads mapping to the transcriptome were considered to be degraded mRNA and removed. In the final step, miRanalyzer, via Bowtie, mapped all reads to the genome draft. The region surrounding these reads was analyzed for the potential to form stable pre-miRNA hairpins. In each step of our analyses we chose conservative parameters. All mapping steps did not allow any mismatches, and miRNA predictions were required to meet a probability score of 0.95 and positively predicted by four of five miRNA models in miRanalyzer. In some instances, miRanalyzer predicted multiple pre-miRNAs from a single sequence read. These multiple hairpins were usually offset by a single nucleotide. In these cases, one pre-miRNA was kept while other overlapping pre-miRNAs were excluded from additional analyses. Predicted pre-miRNA loci were intersected with RepeatMasker TE annotations using Bedtools (Quinlan and Hall, 2010). MiRNAs were considered TE-derived if the mature miRNA overlapped a TE by 10 nt or more. When a region was
annotated as multiple TEs or when a hairpin miRNA intersected two TEs, the TE overlapping the miRNA the most was retained.

To infer likely miRNA origination periods, we identified the presence of the mature miRNA predicted by miRanalyzer in 22 mammalian genomes using the Mapmi package (Guerra-Assuncao and Enright, 2010). By default Mapmi uses the RepeatMasked version of Ensemble genomes. To circumvent this feature, we used a local installation of Mapmi and supplied non-RepeatMasked genomes (Ensemble release 70). After the querying each genome with all predicted, mature miRNAs an estimated origination period was inferred for each. MiRNA origination was mapped onto a reduced version (22 taxa) of the mammalian phylogeny provided by Meredith et al. (2011) using the Dollo parsimony function in Phylip.

**De novo gene prediction in *Eptesicus***

Gene predictions on EptFus1.0 were accomplished using the Augustus (Stanke and Waack, 2003). In addition to the assembled genome scaffolds, two reference species, *Pteropus vampyrus* and *Myotis lucifugus* were incorporated into the Augustus analysis. Tophat version 2.0.6 and Bowtie version 2.0.5 were used to align the RNAseq reads to the assembled genome scaffolds. These alignments were then used by Augustus to improve gene predictions. Note that the supplementary scripts and programs required that are packaged with Augustus were used from Augustus version 2.7, when Augustus itself was run, version 2.5.5 was used. The reference genomes from the *P.vampyrus* and *M lucifugus* were added in the second iteration of running Augustus. The sequences were downloaded from the Ensembl 70 database, and alignments were created using the genblastg version 1.38 with the `-gff` argument to generate alignments in GFF format.
The results were concatenated to the hints file generated from the RNA-Seq data and the first iteration of Augustus predictions. For the final run of Augustus, the '-UTR=on' argument was added to generate UTR predictions.

**miRNA target prediction**

Mammalian miRNAs regulate mRNA expression through targeted binding within the 3' untranslated region (UTR) of testis transcripts as per our Augustus gene predictions. Perfect complementarity within the seed region of the miRNA (nt 2-8) is necessary for miRNA binding. To identify potential miRNA binding sites, miRNAs were mapped to the 3'UTRs using Bowtie with the following parameters: the first base on the 5' end was skipped, the seed was 7 bp long, no mismatches were allowed within the seed, and up to two mismatches were allowed in the remainder of the miRNA.

**Acknowledgements**

We thank the Broad Institute Genomics Platform and Genome Sequencing and Analysis Program, Federica Di Palma, and Kerstin Lindblad-Toh for making the data for *Eptesicus fuscus* and *Myotis lucifugus* available. The mammal phylogeny was kindly provided by R. Meredith. This work was supported by the National Science Foundation [EPS-0903787 (FGH), DBI-1147029 (CJS), MCB-0841821 (DAR) and DEB-1020865 (DAR)]. Additional support was provided by the College of Agriculture and Life Sciences at Mississippi State University. Tissues were provided by the College of Veterinary Medicine at Mississippi State University and Jeremiah Dumas. RNA-Seq and small RNA reads have been deposited in the Gene Expression Omnibus archive (GSE47593).
Table 3.1  Summary of miRNA data for *Eptesicus*, dog and horse.

<table>
<thead>
<tr>
<th></th>
<th><em>Eptesicus</em></th>
<th>Dog</th>
<th>Horse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previously identified miRNAs $^1$</td>
<td>-</td>
<td>323</td>
<td>341</td>
</tr>
<tr>
<td>Predicted testis miRNAs</td>
<td>661</td>
<td>212</td>
<td>388</td>
</tr>
<tr>
<td>Previously identified miRNAs sequenced from testis</td>
<td>-</td>
<td>229</td>
<td>295</td>
</tr>
<tr>
<td>Total miRNAs analyzed from each taxon</td>
<td>661</td>
<td>535</td>
<td>729</td>
</tr>
</tbody>
</table>

### Branches of interest

<table>
<thead>
<tr>
<th></th>
<th>A - <em>Eptesicus</em></th>
<th>A - Dog</th>
<th>A - Horse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total miRNAs</td>
<td>396</td>
<td>184</td>
<td>339</td>
</tr>
<tr>
<td>Non-TE derived miRNAs</td>
<td>154</td>
<td>140</td>
<td>284</td>
</tr>
<tr>
<td>TE derived miRNAs</td>
<td>242</td>
<td>44</td>
<td>55</td>
</tr>
<tr>
<td>Class II derived miRNAs</td>
<td>142</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>% Class II derived miRNAs</td>
<td>35.9%</td>
<td>2.2%</td>
<td>5.6%</td>
</tr>
<tr>
<td>miRNA origination rate (per my)</td>
<td>4.8</td>
<td>2.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Non-TE miRNA origination rate (per my)</td>
<td>1.9</td>
<td>1.7</td>
<td>3.5</td>
</tr>
<tr>
<td>TE-derived miRNA origination rate (per my)</td>
<td>3.0</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Class II-derived miRNA origination rate (per my)</td>
<td>1.7</td>
<td>0.05</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C - <em>Eptesicus</em></th>
<th>C - E</th>
<th>D - E</th>
<th>F - Dog</th>
<th>B - Horse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total miRNAs</td>
<td>385</td>
<td>209</td>
<td>-</td>
<td>174</td>
<td>323</td>
</tr>
<tr>
<td>Non-TE derived miRNAs</td>
<td>149</td>
<td>32</td>
<td>-</td>
<td>131</td>
<td>268</td>
</tr>
<tr>
<td>TE derived miRNAs</td>
<td>236</td>
<td>177</td>
<td>-</td>
<td>43</td>
<td>55</td>
</tr>
<tr>
<td>Class II derived miRNAs</td>
<td>142</td>
<td>106</td>
<td>106</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>% Class II derived miRNAs</td>
<td>36.9%</td>
<td>50.7%</td>
<td>-</td>
<td>2.3%</td>
<td>5.9%</td>
</tr>
<tr>
<td>miRNA origination rate (per my)</td>
<td>5.8</td>
<td>5.0</td>
<td>-</td>
<td>3.7</td>
<td>4.0</td>
</tr>
<tr>
<td>Non-TE miRNA origination rate (per my)</td>
<td>2.2</td>
<td>0.8</td>
<td>-</td>
<td>2.8</td>
<td>3.3</td>
</tr>
<tr>
<td>TE-derived miRNA origination rate (per my)</td>
<td>3.5</td>
<td>4.3</td>
<td>-</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>Class II-derived miRNA origination rate (per my)</td>
<td>2.1</td>
<td>2.6</td>
<td>5.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

---

1 - From mirBase

2 - MRCA = Most Recent Common Ancestor, in this case Variamana. Date taken from global mean provided in Meredith et al. 2011.

3 - Time estimates for all nodes are global means from Meredith et al. 2011 except E - *Eptesicus* and D - E. F - *Eptesicus* was were calculated using TimeTree (*Eptesicus* vs. *Myotis*). D - E is the difference between the dates on nodes D and E.
Figure 3.1  Recent DNA transposon insertions are more common in *Eptesicus* and *Myotis* than in dog and horse.

Genome contributions (% genome) from TE classes over the past 40 my.  Class I elements (retrotransposons) are shaded blue and Class II elements (DNA transposons) are shaded red.  Genome coverage estimates were obtained by RepeatMasking the genome drafts with custom TE libraries for the two bats, and RepBase-derived libraries, dog and horse.  Only hits spanning at least 50% of the consensus were considered.  Divergences from the consensus were calculated as described in the Materials and Methods and ages were calculated by applying a mutation rate of 2.366 x 10^-9.
Figure 3.2  miRNA evolution along a mammalian phenogram.

Phenogram illustrating relationships and approximate divergence times for relevant taxa. Divergence times are as described in Meredith et al. (2011). Broken branches indicate taxa for which we do not have whole genome data, yielding holes in our knowledge with regard to TE complement, but which are relevant to the analyses (see the text). Total miRNA originations are identified below each branch for *Eptesicus* (red), dog (blue) and horse (green) (and on selected internal branches). Numbers of TE-derived miRNAs are found above each branch based on their origin. Class I (retrotransposon) derived miRNAs are shown in brackets and Class II (DNA transposon) derived miRNAs are outside the brackets. miRNAs that originated prior to the common ancestor of bats, carnivores and cetartiodactyls based on queries against 19 additional mammal genomes are considered "Shared" and would have originated on an undepicted branch leading to node A. miRNAs predicted by miRanalyzer but that cannot be assigned to a node are considered unplaced.
Figure 3.3  Recent transposable element history of *Eptesicus fuscus*.

Line graph illustrating contributions to the *Eptesicus* genome by TEs over the past ~60 my (just prior to the estimated divergence of *Eptesicus* from *Artibeus*). As with Figure 1, shades of blue indicate retrotransposons and shades of red indicate DNA transposon superfamilies.
References


Hellen, E.H. and Brookfield, J.F., 2013. The diversity of class II transposable elements in mammalian genomes has arisen from ancestral phylogenetic splits during ancient waves of proliferation through the genome. Molecular Biology and Evolution 30, 100-8.


CHAPTER IV
TARGETED CAPTURE OF VES SINE INSERTIONS FOR PHYLOGENETICS IN
MYOTIS

Introduction

Transposable elements (TEs) are repeats within eukaryotic genomes that have the ability to mobilize and/or replicate themselves. While values can vary greatly among taxa, significant portions of most eukaryotic genomes are derived from TEs. For example, ~85% of the corn (*Zea mays*) genome is derived from TEs (Tenaillon et al., 2011), while TEs comprise less than 3% of the pufferfish (*Takifugu rubripes*) genome (Aparicio et al., 2002). In mammals, estimates of TE content range from ~40-60% (Lander et al., 2001; Waterston et al., 2002; Gentles et al., 2007; de Koning et al., 2011). By comparison protein coding regions make up less than 2% of the human genome (Lander et al., 2001)

TEs can be subdivided into two major classes based on their mobilization mechanism. Class I TEs (retrotransposon) mobilize through a copy and paste mechanism where the parent retrotransposon is transcribed and re-integrated into the genome at another location via reverse transcription. Retrotransposons are the dominant TE class in mammalian genomes and include several superfamilies of LINEs (Long INterspersed Elements, SINEs (Short INterspersed Elements), LTRs (Long Terminal Repeat retrotransposons), and ERVs (Endogenous RetroViruses). Class II TEs are referred to as
DNA transposons and include a wide array of superfamilies including Tcl/Mariners, hATs, and piggyBacs. Helitrons and Mavericks are also included in this class but utilize distinct mobilization mechanisms. Other than the bat family Vespertilionidae, including the genus *Myotis*, class II TEs have been essentially inactive in mammal genomes for the last ~40 my (Pritham and Feschotte, 2007; Ray et al., 2007; Ray et al., 2008; Thomas et al., 2011; Pagán et al., 2012).

Retrotransposon have been used as phylogenetic markers for some time (Ryan and Dugaiczyk, 1989; Minghetti and Dugaiczyk, 1993; Murata et al., 1993; Murata et al., 1996; Nikaido et al., 1999) and are being used with increasing frequency (for examples see: Ray et al., 2005a; Xing et al., 2005; Xing et al., 2007; Hallström et al., 2011; Suh et al., 2011; Haddrath and Baker, 2012; Matzke et al., 2012; McLain et al., 2012; Meyer et al., 2012). SINEs in particular are well suited for this purpose (Okada, 1991; Hillis, 1999; Ray et al., 2006). These non-autonomous retrotransposons are 100-500 bp and reach copy numbers ranging from several thousand to over a million in mammalian genomes. Of these insertions, each one can serve as an independent phylogenetic marker, reflecting the evolutionary history of the genome they occupy. For example, SINE insertions that occurred in a parent lineage would be shared among all daughter lineages. As the daughter lineages diverge, their daughter taxa would accumulate SINE insertions at additional, lineage-specific loci. Though each lineage harbors a unique SINE insertion pattern, lineages more recently diverged from one another will share more orthologous SINE insertions than those more distantly related. Retrotransposon insertions have been successfully implemented to differentiate individuals (Wang et al., 2006), populations (Ray et al., 2005a; Witherspoon et al., 2006; Witherspoon et al., 2013), species (López-
Giráldez et al., 2005) and even to resolve difficult nodes at the taxonomic levels beyond family (Nikaido et al., 2007; Churakov et al., 2010).

SINE based phylogenies have three major advantages (Ray et al., 2006) over DNA sequence phylogenies. First, unlike DNA sequence based markers, any single SINE insertion locus can have only two character states, presence (1) or absence (0), making interpretation of results relatively simple. Second, the ancestral state for a locus is known to be the absence of a SINE insertion. This reduces the number of assumptions required during analyses since any putative ancestral taxon would lack a certain subset of insertions, depending on the ancestor. Third, nucleotides in non-related taxa can have the same character state due to back mutation or selection rather than homology (i.e. homoplasy). Since the precise insertion or excision of a SINE element is rare, homoplasious insertions are significantly less likely than homoplasious nucleotide mutations. Some care can be used to clarify potentially insertion patterns where homoplaspy is suspected (Ray et al., 2005b; Ray et al., 2006; Xing et al., 2007; Li et al., 2009b; Chen et al., 2011; Han et al., 2011; McLain et al., 2012; Meyer et al., 2012).

Though SINEs offer valuable perspectives on difficult phylogenies, identifying informative insertions in non-model taxa has been difficult. More commonly, a reference genome would be mined for potentially informative SINE insertions. Primers were then developed in the regions flanking the insertion in the reference genome. These primers would then be used to amplify the SINE insertion locus in other taxa. Amplicons from these reactions would be electrophoresed and scored based on the band size. Taxa sharing the SINE insertion would yield larger bands than taxa lacking the SINE insertion. While this method has produced well-resolved phylogenies, support is necessarily biased
along nodes leading to the reference genome. If the reference genome was a basally diverging taxa, the remainder of the tree would lack any support. In addition, large numbers of SINE insertions needed to be analyzed to find phylogenetically informative insertions. Recently developed sequencing technologies, laboratory methods and the increasing availability of reference genomes provide a framework for addressing these shortcomings.

*Myotis* (Chiroptera: Vespertilionidae), is one of the most species rich, mammalian genera comprising over 100 species distributed worldwide (Simmons, 2005). Despite being one of the more commonly studied mammalian groups, many phylogenetic questions have been difficult to resolve (Hoofer and Bussche, 2003; Kawai et al., 2003; Bickham et al., 2004; Stadelmann et al., 2004; Stadelmann et al., 2007; Evin et al., 2008; Carstens and Dewey, 2010; Lack et al., 2010; Larsen et al., 2012). In the past, *Myotis* were classified into subgenera based on ecoptypes which were defined by morphology, ecology, and behavior. As DNA-sequence based phylogenetic resources became more abundant it was clear that morphology-based classifications resulted in polyphyletic assemblages. Instead relationships among species were better characterized by geography (Ruedi and Mayer, 2001; Hoofer and Bussche, 2003; Stadelmann et al., 2007). Mitochondrial DNA-based analyses suggest that 15 million years ago (mya) *Myotis* diverged from *Kerivoula* and split into old world (OW) and new world (NW) clades followed by further subdivision into Nearctic (NA) and neotropical (NT) clades by ~10 mya. The only exception to a strict division between NW and OW clades was the placement of *M. gracilis* and *M. brandti*, western Asian species, within the NW *Myotis* between the NA and NT clades (Stadelmann et al., 2007).
The availability of the *M. lucifugus* draft genome has led to extensive study of its transposable elements (Pritham and Feschotte, 2007; Ray et al., 2007; Ray et al., 2008; Thomas et al., 2011; Pagán et al., 2012). Beyond the recent DNA transposon activity mentioned above, Ves SINES have been active over the last 60 my in bats (Borodulina and Kramerov, 1999; Kawai et al., 2002; Pagán et al., 2012) thus spanning the *Myotis* divergence. Here we describe a protocol to identify potentially informative SINE insertions in the genus *Myotis*. The protocol is modified from the ME-Scan protocol, a sequence-capture method targeting Alu insertions in humans (ME-Scan; Witherspoon et al., 2010; Witherspoon et al., 2013). The availability of the *Myotis lucifugus* genome, the rapid species radiation characteristic of this clade, discordance between morphology and phylogeny, and the general difficulty in resolving phylogenetic questions make our modified protocol a potentially valuable resource. We have identified over 796 thousand Ves SINE insertions in six NW and one OW species of *Myotis* and validated a subset of these insertions to validate the method and demonstrate its phylogenetic utility.

**Materials and Methods**

**Specimens examined**

SINE libraries were generated for six NW (*Myotis M. auriculus, M. dominicensis, M. lucifugus, M. M. occultus, M. simus, M. vivesi*) and one OW (*M. M. horsfieldii*) taxa (Table 1) using an adaptation of the ME-Scan protocol (Witherspoon et al., 2010; Witherspoon et al., 2013). Modifications were made to target the relatively young Ves3ML SINE subfamily, which is specific to vesper bats (Ray, In prep.). Taxa were chosen based on their phylogenetic relationships as well as DNA availability, quality, and quantity.
**Library preparation and Ves enrichment**

A summary of the biochemical pipeline is shown in Figure 1A-H. Genomic DNA (gDNA) was extracted from tissue samples using a standard phenol-chloroform / ethanol precipitation protocol. For each sample 10 µg of gDNA was fragmented to an average size of 1 kilobase (Kb) on a Covaris S220 Focused-ultrasonicator using the following parameters: peak incident power, 105 watts; duty factor, 5.0%; cycles per burst, 200; time, 40 seconds; temperature, 7ºC (Figure 4.1A). Fragmented gDNA was purified and concentrated using Qiagen QIAquick PCR purification columns (Qiagen, Germantown, MD) using the recommended protocol.

Fragmented gDNA was then prepped for Illumina sequencing using the End Repair, dA Tailing and Adapter Ligation modules from the NEBNext DNA Library Prep Master Mix set (New England Biolabs, Ipswich, MA). During the adapter ligation step, custom, indexed adapters were ligated to each library for species identification (Figure 4.1B). Adapter and index sequences are presented in Table 4.2. After ligation, each indexed library was quantified using a NanoDrop ND-1000 spectrophotometer and all seven libraries were pooled into a single library so that each species was equally represented. All subsequent steps were performed on this combined pool.

The combined pool of indexed gDNA fragments were enriched for Ves insertions by binding a biotinylated probe that is complementary to the consensus Ves3ML subfamily element (Figure 4.1C). The biotinylated probe targets a 20 bp region beginning at the 59th nucleotide of the Ves3ML consensus sequence. A 5 cycle PCR was used to bind the Ves biotinylated probe to the gDNA fragments under the following thermal conditions: an initial denaturation of 98ºC for 30 seconds; 5 cycles of 98ºC for 10
seconds, 65°C for 30 seconds, 72°C for 30 seconds; and a final 72°C 5 minute extension.

Reaction concentrations were as follows 6 µL (~150 ng) of pooled gDNA fragments, 1 µl [10 µM ] Ves.btin.bp59 (biotinlyated probe), 1 µL [10 µM] Illumina P7 primer, 4 µL [5x] NEB Phusion HF buffer, 0.4 µL [10 mM] dNTPs, and 0.2 µL NEB Phusion Taq in a 20 µL reaction. Fragments between 500 and 600 bp were size selected via electrophoresis (Figure 4.1D) on a 15 cm, 2% agarose gel run at 80 volts for 4 hours and purified using the Qiagen gel extraction kit. Ves fragments bound by the biotinylated probe were magnetically isolated from gDNA fragments using streptavidin-coupled Dynabeads (Figure 4.1E) resulting in a library of Ves-enriched fragments between 500-600 bp, each containing a species specific indexed adapter. The Ves fragments were then amplified for an additional 20 cycles (Figure 4.1F) with the standard Illumina P7 primer and another composite Ves + random sequence + Illumina P5 primer (Table 4.2). This Ves PCR primer was designed so that the first 20 bp were complementary to Ves, the next 3-6 bp alternated purines and pyrimidines, and the remaining 58 bp contained the P5 Illumina sequence. Amplification of the Ves fragments was done in three 25 µL reactions with 2 µL bead-bound Ves fragments, 1 µl [10 µM] mixture of all 8 composite Ves PCR composite primer, 1 µL [10 µM] Illumina P7 primer, 5 µL [5x] NEB Phusion HF buffer, 0.5 µL [10 mM] dNTPs, and 0.4 µL NEB Phusion Taq using the previously mentioned cycling conditions. After amplification, bead bound Ves fragments were magnetically removed, and the PCR aliquot was electrophoresed at 40 volts for 5 hours on a 15 cm 2% agarose gel (Figure 4.1G). Ves fragments ~550 bp from all three amplified samples were purified from a gel excision using a Qiagen QiaQuick Gel Extraction column to produce the final sequencing library. Nanomolar concentration was calculated using a Qubit
fluorometer. The final library of Ves fragments were sequenced on a single Illumina HiSeq 2000 through a commercial core lab (Figure 4.1H). The sequencing protocol allowed for 100 nt paired-end reads.

**Data processing**

Expected read pairs contained one read that spanned ~17-30 bp of the Ves insertion and up to 83 bp of gDNA, designated as the ‚Ves’ read. The other read, the ‚Flanking’ read, consisted of 100 bp of gDNA ~400 bp away from the insertion site. Both the Ves and Flanking reads were deconvoluted into species specific files based on the 6 bp index identified in the Flanking read using Sabre (https://github.com/najoshi/sabre). Any reads where the sequence index did not perfectly match the template index were excluded from further analyses. After reads were deconvoluted, the Ves reads for each sequence pair was checked for complementarity to the 5” 34 bp of the consensus Ves3ML using fastx_clipper from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Regions complementary to this portion of Ves3ML were clipped, and any Ves reads lacking complementarity were discarded. In addition, any reads with an average quality Phred+33 quality score less than 15 were removed. Processing the Ves and Flanking reads independently necessarily resulted in files where the number of sequences and sequence order were different. A custom perl script was used to re-organize the Ves and Flanking files so that orphaned reads were placed in a separate file and the complementarity of reads in both files was restored.

Paired reads were mapped to the 7x Myotis lucifugus (Myoluc2.0: GCA_000147115.1 ) genome using BWA (Li and Durbin, 2009) with the default options. Reads were initially mapped independently, then combined using BWA sampe based on
an average insert size (-a) of 400 bp. Using SAMtools (Li et al., 2009a) read pairs were filtered so that only those mapping in the proper orientation (-f 0x002) and within ~400 bp of its mate were kept. For these read pairs, the Ves insertion site was designated as the immediate nucleotide position on the 5’ end of the Ves read. Comparisons of shared Ves insertions sites between species showed that most sites were designated at the same position or within a small 1-2bp window (Figure 4.2). Based on the precision in read mapping determined a posteriori, reads within a ±2 bp window were merged into the nucleotide position most commonly designated as the Ves insertion site. All custom scripts used for data analysis are publically available at https://github.com/nealplatt/bioinformatics.

Validation of Ves insertions via PCR and Sanger sequencing

To verify the ability of our method to capture polymorphic Ves insertions in Myotis, 20 loci were chosen for validation via PCR. Insertions sites tested were chosen at random after meeting the criteria listed. First, no other TE insertions (as annotated in the M. lucifugus genome) fell within a ±1 Kb window of the Ves insertion site. Some insertions were recovered by only one Ves/Flanking read pair in one of the seven taxa, while others were supported by greater sequencing depth, or in multiple species. These poorly supported reads, i.e. insertions identified less than 10 times, were excluded from PCR validation. Finally validated insertions could be within 1 Kb of distinct insertions identified in any of the other species. For those loci that meet the requirements, the insertion position plus 500 bp of flanking sequence were extracted from the M. lucifugus genome. Using BatchPrimer3 (You et al., 2008) primers were designed to flank the expected Ves insertion site plus 100 bp on the 5’ and 3’ ends so that when amplified the
size of each locus would vary by ~200 bp based on the presence or absence of the Ves
insertion (Table 4.3).

The predicted insertion site was then amplified in seven *Myotis* taxa. For certain
samples the available DNA was exhausted during library preps, for these taxa con-
specific samples were used. PCR reaction conditions were as follows, 1.5 µL [10x]
buffer, 0.7 µL RediLoad loading dye, 1 µL Taq, 0.9 µL of MgCl₂ [25 mM], 0.3 µL of
each primer [10 µM], 0.3 µL of dNTPs [10 mM], 15 ng of gDNA and water to 15 µL.
Samples were amplified via an initial 94°C denaturation period of 1 minute, then 32
cycles of a 45 second 94°C denaturation, 45 second at 52°C annealing, and a 45 second
72°C extension periods, followed by a final 3 minute 72 ºC extension. Amplicons were
visualized on a 15 cm, 1.5% agarose gel and electrophoresed at 140 volts for 60 minutes.
After visual inspection a minimum of one sample with a Ves insertion was purified using
Qiagen QIAquick PCR purification columns. Samples were bi-directionally sequenced
using internal primers binding to the 5” and 3” regions of the Ves.

**Results and Discussion**

Currently, SINE based phVylogenetics is infrequently used due to the difficulty in
generating large data sets in non-model taxa. Our goal was to identify Ves insertions, in
taxa for which no reference genome is available, to be used in future phylogenetic work.
To do this we isolated and sequenced Ves containing DNA fragments from six NW and
one OW *Myotis* species on a large scale. After sequencing, insertion positions were
identified and analyzed bioinformatically to validate the protocol.

Sequencing the Ves libraries on a single Illumina HiSeq lane yielded over 103.5
million read pairs. All but ~297.4 thousand read pairs were viable (i.e. binned into
species specific files based on the 6 bp index at the beginning of each Flanking read). Despite individually tagged gDNA libraries being pooled in equal proportions, the number of reads per taxon varied by more than 2.5 orders of magnitude; from ~5.8 thousand (M. lucifugus) to over 23.4 million (M. M. auriculus; Table 4.4), though all taxa other than M. lucifugus had ~10 million or more read pairs.

Based on the design of the Ves PCR composite primer our libraries, ~30 bp of Ves sequence was expected at the beginning (5’) of each Ves read. This expected Ves sequence was identifiable in 99.6 to 99.8% of Ves reads from all read pairs. More than three-quarters (75.6%) of all read pairs were able to be mapped to the M. lucifugus genome, with success ranging from 59.3% (M. M. horsfieldii) to 86.2% (M. M. occultus; Table 4.4). The mapping ability, and high Ves recovery rate in Myotis combined with high Alu recovery rates (94.3%) in humans (Witherspoon et al., 2010) indicates that the ME-Scan method is highly efficient for sequencing large numbers SINE insertions under a range of conditions and should be extendable to SINE families in other mammals.

**Precision and accuracy of Ves identification**

To use the Ves insertions identified in the sequencing and computational steps above, the insertion position must be identified with a high degree of precision and accuracy. This is particularly important for shared or polymorphic insertions whose inferred presence or absence could be used for phylogenetics. A direct relationship exists among distance between insertion sites, and our ability to identify them as unique insertion events. Two SINE insertions separated by 100 -1,000 bp are easily identifiable as unique insertions. However, when one reduces that distance to 1-10 bp, their uniqueness is more difficult to resolve. Insertions within a narrow window could be
homoplastic or the result of multiple reads from the same insertion mapping imprecisely. To address this question, the distance to the nearest neighboring insertion for each insertion position was calculated in all seven taxa. While two insertions separated by 10 bp in different species could be due to imprecise read mapping or homoplasy, insertions separated by 10 bp from the same individual are almost certainly due to imprecise mapping or sequencing mistakes. When closest neighbors were calculated within each species, we found that almost all *Ves* reads supported insertion sites within a very narrow window (Figure 4.2). In all seven taxa 99.5% of *Ves* reads with a ±100bp of each other had a nearest neighbor within ±10bp. Further, 96.7 – 98.7% all reads from the same ±100bp window were supported by other reads at the same position (±0 bp). Based on these results, insertion positions within each taxon were identified based on the position with the largest number of reads within any continuous stretch of identified insertion positions. For example, if *Ves* reads identified chr1:100-102 as potential insertion positions in *M. M. auriculus*, but chr1:100 was supported by 10 *Ves* reads, chr1:101 by 1 *Ves* read, and chr1:102 by 140 *Ves* reads, chr1:102 was designated as the true insertion position.

Once insertion positions had been identified within each species, these positions were compared among species. Similar to the within species results, insertion positions among species tended to be identified with a high degree of accuracy within a ±10 bp window (Figure 4.3). Though parallel insertions cannot be completely ruled out, the mapping accuracy is strong evidence that identified insertion positions within a narrow window are likely shared insertions acquired from the common ancestor of both taxa. On the other hand, even if parallel insertion events cannot be ruled out, these events are
expected to occur at such a low frequency that they would not be expected to substantially impact our result (Ray et al., 2006). Combined, these two semi-independent lines of evidence (precision within species and accuracy among species) suggest that reads mapping within a narrow window are very likely to have originated from the same ancestral insertion, an assumption that is important for phylogenetic inquiry.

**Average sequence coverage of Ves**

Between 120,214 (*M. vivesi*) and 143,322 (*M. M. auriculus*) unique Ves insertions were identified, excluding *M. lucifugus* which yielded poor sequencing results (Table 4.4). On average each Ves insertion was sequenced to 129.6x coverage. While *M. M. auriculus* was sequenced 2.5x (13.6 million read pairs) more frequently than *M. simus*, the difference in number of Ves insertions identified was less than 17.8 thousand insertions. Assuming these two taxa, which diverged ~10 mya (Stadelmann et al., 2007), have similar numbers of Ves insertions, this suggests that a saturation point is reached where more sequencing does not significantly increase the number of unique insertions discovered (Figure 4.4A). Our results indicate that after ~5 million read pairs, most Ves insertion sequences are represented. Additionally, when the number of reads pairs is compared to the number of reads per Ves insertion a strong linear relationship is recovered ($R^2 = 0.975$; Figure 4.4B).

These results provide strong evidence that there are only ~142 thousand Ves insertions recoverable using the ME-Scan method and our probe design. There are 544,807 Ves insertions divided among four subfamilies (Ves2_ML = 100,744; Ves2B_ML = 121,261; Ves3_ML = 278,544; Ves4_ML = 44,258) identified in the *M. lucifugus* genome (www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html).
Our probes were designed to bind to positions within Ves that are specific to the Ves3_ML subfamily, yet we are only able to recover ~50% of the expected number of insertions. To determine if our probe design or number of reads was the limiting factor in the number of Ves recovered the Ves portion of the biotinylated probe and Ves PCR composite primers were queried against the *M. lucifugus* genome. To identify a capturable binding site, all portions of the genome complementary to the Ves portion of our biotinylated probe were searched for Ves PCR composite binding sites within a ±50 bp region. Using these criteria a maximum of 164,689 potentially capturable Ves sites exist within *Myotis lucifugus*. If this number is valid, then we captured almost 70% of all potential sites, and are likely sequencing Ves fragments to saturation.

**Validation with PCR and Sanger Sequencing**

To verify the presence of Ves insertions identified by *Ves* reads, 42 loci were amplified in all seven taxa (Figure 4.5). Loci containing a Ves insertion are expected to be ~200 bp larger than loci lacking a Ves. Of the loci tested 32 of 42 amplified in five or more samples. Thirty of the 32 primer pairs produced amplicons of predicted sizes. After verifying the increase in band size that is expected from a Ves insertion, loci were sequenced using internal Ves primers to confirm the ~200 bp insertion was indeed a Ves. For each locus either one or two taxa were selected for sequencing. Sequencing of insertion loci verified the presence of the expected insertions.

**Conclusions**

The benefits of SINE based phylogenies have been discussed extensively in the past (Hillis, 1999; Ray et al., 2006; Ray, 2007). Unfortunately, identifying informative
SINE loci required expensive and time-consuming laboratory techniques or relied on insertions identified in a reference taxon. In these cases, nodes could only be supported along the reference lineage. As new sequencing technologies have become accessible, the ability to identify SINE insertions on a large scale has become feasible. Further, as the number of reference genomes increases, many non-model taxa will have closely related genomes available for comparative bioinformatic analyses. To lay the groundwork for a SINE insertion based phylogeny of New World Myotis, the ME-Scan protocol (Witherspoon et al., 2010; Witherspoon et al., 2013) was modified to target Ves elements. We show that large numbers of insertions can be captured in taxa separated by up to 12 million years (Stadelmann et al., 2007) and that insertion positions can be identified to a narrow window within and among species. In *Myotis*, we identified over 796 thousand SINE insertions in seven taxa. On average each insertion was supported by a large number of reads (129.6x), and each species (excluding *M. lucifugus*) recovered similar numbers of Ves insertions though the number of reads per taxa varied greatly. Though more work remains to turn these Ves insertions into a phylogenetically informative dataset, demonstrating the capacity of this method is an essential first step.

**Acknowledgments**

We would like to thank the following institutions for tissue loans: Natural Science Research Laboratory, Museum of Texas Tech University; Louisiana Museum of Natural Science, Section of Genetic Resources; Museum of Southwestern Biology, University of New Mexico. This research project was supported by grants from the National Science Foundation [MCB-0841821 and DEB-1020865], Sigma Xi Grant-in-Aid [G20130315163915], as well as funding from the Institute for Genomics, Biocomputing
and Biotechnology at Mississippi State University. Finally we would like to thank M. W. Vandewege and F. G. Hoffmann for their insights, advice, and constructive criticism.
Table 4.1  Voucher ID and collecting locality for each specimen examined.

<table>
<thead>
<tr>
<th>Voucher I.D.</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New World</strong></td>
<td></td>
</tr>
<tr>
<td><em>Myotis lucifugus</em></td>
<td>MSB 40815</td>
</tr>
<tr>
<td><em>Myotis M. auriculus</em></td>
<td>MSB 40883</td>
</tr>
<tr>
<td><em>Myotis dominicensis</em></td>
<td>TTU 31510</td>
</tr>
<tr>
<td><em>Myotis simus</em></td>
<td>TTU 46348</td>
</tr>
<tr>
<td><em>Myotis M. occultus</em></td>
<td>MSB 121995</td>
</tr>
<tr>
<td><em>Myotis vivesi</em></td>
<td>MSB 42649</td>
</tr>
<tr>
<td><strong>Old World</strong></td>
<td></td>
</tr>
<tr>
<td><em>Myotis M. horsfieldii</em></td>
<td>M 4424</td>
</tr>
</tbody>
</table>

MSB - Museum of Southwestern Biology, TTU - Natural Science Research Laboratory, M - LSU Museum of Natural Science
<table>
<thead>
<tr>
<th>Species specific adapters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Myotis lucifugus</em> [GCCTAA]</td>
<td></td>
</tr>
<tr>
<td>Tag2-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGCTCTCCGATCTGCCTAA*T-3'</td>
</tr>
<tr>
<td>Tag2-short</td>
<td>5'-TTAGGCAGATCGGAAGAGCGTCGTG-3'</td>
</tr>
<tr>
<td><em>M. auriculus</em> [TGGTCA]</td>
<td></td>
</tr>
<tr>
<td>Tag3-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATCCCTAACAGAGTCAGTCAGTGAGTTCAAGACGTGCTCTCCGATCTTGCTCA*T-3'</td>
</tr>
<tr>
<td>Tag3-short</td>
<td>5'-TGACCAAGATCGGAAGAGCGTCGTG-3'</td>
</tr>
<tr>
<td><em>M. dominicensis</em> [CACTGT]</td>
<td></td>
</tr>
<tr>
<td>Tag4-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATCGGTGACTGGAGTCAGTCAGTCAGTCAAGACGTGCTCTCCGATCTCAGCTGT*T-3'</td>
</tr>
<tr>
<td>Tag4-shor</td>
<td>5'-ACAGTGAGATCGGAAGAGCGTCGTG-3'</td>
</tr>
<tr>
<td><em>M. simus</em> [ATTGGC]</td>
<td></td>
</tr>
<tr>
<td>Tag5-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATCGGTGACTGGAGTCAGTCAGTCAGTCAAGACGTGCTCTCCGATCTCAGCTGT*T-3'</td>
</tr>
<tr>
<td>Tag5-short</td>
<td>5'-CCAAATAGATCGGAAGAGCTGCTG-3'</td>
</tr>
<tr>
<td><em>M. occultus</em> [GATCTG]</td>
<td></td>
</tr>
<tr>
<td>Tag6-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATCGGTGACTGGAGTCAGTCAGTCAGTCAAGACGTGCTCTCCGATCTCAGCTGT*T-3'</td>
</tr>
<tr>
<td>Tag6-short</td>
<td>5'-CAGATCAGATCGGAAGAGCTGCTG-3'</td>
</tr>
<tr>
<td><em>M. vivesi</em> [TCAAGT]</td>
<td></td>
</tr>
<tr>
<td>Tag7-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATCGGTGACTGGAGTCAGTCAGTCAGTCAAGACGTGCTCTCCGATCTCAGCTGT*T-3'</td>
</tr>
<tr>
<td>Tag7-short</td>
<td>5'-ACTTTGAAGATCGGAAGAGCTGCTG-3'</td>
</tr>
<tr>
<td><em>M. horsfieldii</em> [AAGCTA]</td>
<td></td>
</tr>
<tr>
<td>Tag9-full</td>
<td>5'-CAAGCAGAAGACGCATACGAGATCGGTGACTGGAGTCAGTCAGTCAGTCAAGACGTGCTCTCCGATCTCAGCTGT*T-3'</td>
</tr>
<tr>
<td>Tag9-short</td>
<td>5'-TAGCTTTAGATCGGAAGAGCTGCTG-3'</td>
</tr>
<tr>
<td><strong>5' cycle biotinylated probe</strong></td>
<td></td>
</tr>
<tr>
<td>Ves.short.btinPD.bp59</td>
<td>5'-Biosq/GCCCTTGACCGGAATCGAA*C-3'</td>
</tr>
</tbody>
</table>
Table 4.2 (continued)

<table>
<thead>
<tr>
<th>25 cycle PCR primers</th>
<th>5' - AATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ves.head.bp17_YR3</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_RY3</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_YR4</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_RY4</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_YR5</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_RY5</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_YR6</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
<tr>
<td>Ves.head.bp17_RY6</td>
<td>5' - ATGATACGCGCACTACAGTCACTACACTCTCTCTACACGACGCTCTTCCGATCT-3'</td>
</tr>
</tbody>
</table>

A “*” represents a phosphorothioated nucleotide and “Biosq” represents a 5” biotin modification.
<table>
<thead>
<tr>
<th>Locus Name</th>
<th>Chromosome Position</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeScan_PCR:010</td>
<td>GL429770:29724012-29726213</td>
<td>TTCCTCAAGGGGAATTGTA</td>
<td>GAGTGCTCTGGCTCCTTCCTCCTG</td>
</tr>
<tr>
<td>MeScan_PCR:012</td>
<td>GL429771:4642750-4644951</td>
<td>TGCTCCAGATTTTTAACATGTC</td>
<td>CACTCCTCAAGGCTTTTGGTT</td>
</tr>
<tr>
<td>MeScan_PCR:015</td>
<td>GL429774:15381733-1538393</td>
<td>TGCCATAATTGCAATGACG</td>
<td>GAGACCTGGAGAGAACAA</td>
</tr>
<tr>
<td>MeScan_PCR:019</td>
<td>GL429781:3363439-3365640</td>
<td>GCTCTCCTACCCCTAACAAGC</td>
<td>TAGGGCCTGAAAGGTGGTGATG</td>
</tr>
<tr>
<td>MeScan_PCR:020</td>
<td>GL429784:12143391-12145592</td>
<td>TTTTTCTGATCTGCTCCCTT</td>
<td>CACGCTGCTTCAGGTTTCCT</td>
</tr>
<tr>
<td>MeScan_PCR:021</td>
<td>GL429785:6477072-6479274</td>
<td>GCCTCAGCAAGGAAACAGAA</td>
<td>GCTACCAGTTCATGTAGGC</td>
</tr>
<tr>
<td>MeScan_PCR:022</td>
<td>GL429791:5472425-5474626</td>
<td>TCTTCCAGGAAATTCAGGT</td>
<td>GCCAAAACAATGGACAA</td>
</tr>
<tr>
<td>MeScan_PCR:024</td>
<td>GL429799:4765806-4768011</td>
<td>GACAGGCAAGAGACGAGTACG</td>
<td>CAGACACCTGCATTGGATGG</td>
</tr>
<tr>
<td>MeScan_PCR:025</td>
<td>GL429800:3121113-3123314</td>
<td>TCACTCAATGCGTCTCCTT</td>
<td>AGAAATGAGACTGGGCAAA</td>
</tr>
<tr>
<td>MeScan_PCR:030</td>
<td>GL429818:709761-711963</td>
<td>GCTCTAAAAGGGCTCAAA</td>
<td>GAGCCTGTTGTGAGCTAGA</td>
</tr>
<tr>
<td>MeScan_PCR:034</td>
<td>GL429836:3588373-3590574</td>
<td>TCTGGTCTGCTCTTCAAGACC</td>
<td>ATGGGAAAGCAAGAATGCA</td>
</tr>
<tr>
<td>MeScan_PCR:037</td>
<td>GL429836:3588373-3590574</td>
<td>TCTGGTCTGCTCTTCAAGACC</td>
<td>ATGGGAAAGCAAGAATGCA</td>
</tr>
<tr>
<td>MeScan_PCR:040</td>
<td>GL429848:12143391-12145592</td>
<td>TTTTTCTGATCTGCTCCCTT</td>
<td>CACGCTGCTTCAGGTTTCCT</td>
</tr>
<tr>
<td>MeScan_PCR:045</td>
<td>GL429900:3211113-3213314</td>
<td>GCCTCAGCAAGGAAACAGAA</td>
<td>GCTACCAGTTCATGTAGGC</td>
</tr>
<tr>
<td>MeScan_PCR:048</td>
<td>GL429907:810121-812322</td>
<td>AAGTCTCCATTTCCTGCTCA</td>
<td>CAATGGAAAATTTAGCACGA</td>
</tr>
<tr>
<td>MeScan_PCR:050</td>
<td>GL429919:2220553-2222754</td>
<td>TGGGAGATGTTGTGATTTCAAGG</td>
<td>TATTCATTTGCCCCGATT</td>
</tr>
<tr>
<td>MeScan_PCR:052</td>
<td>GL429932:201256-2014457</td>
<td>AGCAGGCTGGAATATGCTA</td>
<td>AGAGGCCTTGTTGATTCAG</td>
</tr>
<tr>
<td>MeScan_PCR:053</td>
<td>GL429933:2231305-2233506</td>
<td>ATTTGGGGATTACGCCCTTC</td>
<td>TGAAGGGGAAACAGATG</td>
</tr>
<tr>
<td>MeScan_PCR:055</td>
<td>GL429955:491279-493923</td>
<td>GTTCTGCAAGAATCGCTCC</td>
<td>AGTTCTCCCTGCTGCCTTAG</td>
</tr>
<tr>
<td>MeScan_PCR:058</td>
<td>GL429962:1550604-1552805</td>
<td>AGAGGAAACGCTCTGAGTCA</td>
<td>CAGGGCTTCTGAGCCTCAA</td>
</tr>
<tr>
<td>MeScan_PCR:059</td>
<td>GL429964:630373-632574</td>
<td>CAGAATTCTAGCCCATGGA</td>
<td>AATGTGAAACGGTTTTCAGT</td>
</tr>
<tr>
<td>MeScan_PCR:066</td>
<td>GL430035:845756-847957</td>
<td>CAGGGAATCAAGAGAAGCA</td>
<td>GCAGGCTGACATCCATTAGT</td>
</tr>
<tr>
<td>MeScan_PCR:073</td>
<td>GL430135:15531-17732</td>
<td>TCGGACAAAGCAGCCTGTA</td>
<td>GTGTTTTTCCCCATGCACTC</td>
</tr>
<tr>
<td>MeScan_PCR:081</td>
<td>GL430293:55619-57820</td>
<td>CAAACCAGGAAAGACT</td>
<td>AATTCGTTTGGCAAATGTT</td>
</tr>
</tbody>
</table>

Chromosome position is given based on the *Myotis lucifugus* (Myoluc2.0: GCA_000147115.1) genome.
Table 4.4  Basic sequence statistics for each taxon during the data analysis.

<table>
<thead>
<tr>
<th></th>
<th><em>M. M. auricus</em></th>
<th><em>M. dominicensis</em></th>
<th><em>M. M. horsfieldi</em></th>
<th><em>M. lucifugus</em></th>
<th><em>M. M. occultus</em></th>
<th><em>M. simus</em></th>
<th><em>M. vivesi</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of read pairs</td>
<td>23,422,455</td>
<td>18,459,538</td>
<td>17,806,029</td>
<td>5,790</td>
<td>20,257,386</td>
<td>9,802,564</td>
<td>13,496,084</td>
</tr>
<tr>
<td>Read pairs with expected Ves</td>
<td>23,380,075</td>
<td>18,417,016</td>
<td>17,765,391</td>
<td>5,770</td>
<td>20,218,366</td>
<td>9,766,555</td>
<td>13,465,245</td>
</tr>
<tr>
<td>Percent of total</td>
<td>99.82</td>
<td>99.77</td>
<td>99.77</td>
<td>99.65</td>
<td>99.81</td>
<td>99.63</td>
<td>99.77</td>
</tr>
<tr>
<td>Read pairs mapping to reference</td>
<td>18,385,217</td>
<td>14,570,546</td>
<td>10,557,215</td>
<td>3,841</td>
<td>17,457,290</td>
<td>6,398,914</td>
<td>10,810,003</td>
</tr>
<tr>
<td>Percent of total</td>
<td>78.49</td>
<td>78.93</td>
<td>59.29</td>
<td>66.34</td>
<td>86.18</td>
<td>65.28</td>
<td>80.1</td>
</tr>
<tr>
<td>Number of Ves identified</td>
<td>143,322</td>
<td>138,002</td>
<td>127,490</td>
<td>533</td>
<td>141,435</td>
<td>125,559</td>
<td>120,214</td>
</tr>
<tr>
<td>Average coverage per Ves</td>
<td>128.3x</td>
<td>105.6x</td>
<td>82.8x</td>
<td>7.2x</td>
<td>123.4x</td>
<td>51.0x</td>
<td>89.9x</td>
</tr>
</tbody>
</table>
Figure 4.1 Preparation and sequencing of Ves enriched libraries.

(A) Genomic DNA is fragmented to an average size of 1 Kb. (B) The fragmented, genomic DNA is end-repaired, and dA tailed before indexed adapters are added. (C) Individual libraries are pooled together and DNA fragments containing Ves are bound with a biotinylated Ves probe using a 5 cycle PCR reaction. (D) The entire library is size sorted along a gel. Fragments 500-600 bp in length are isolated (E) The size-selected fragments are enriched for Ves by binding the Ves biotinylated probe to streptavidin-coated magnetic beads. (F) A final amplification of 15-20 cycles is used to amplify the library of Ves-enriched fragments for sequencing. (G) A final library of 550 bp Ves fragments is isolated via gel electrophoresis and purified for sequencing. (H) Fragments are sequenced so that one read contains the species specific index plus DNA sequence up to 300-400 bp away from the Ves insertion (flanking read). The second read contains a portion of the Ves insertion plus ~80 bp of the Ves insertion site (Ves read).
Figure 4.2  Precision of estimated Ves insertion sites within each species.

For each species, Ves insertion sites were called based on the mapping location of the *Ves* insertion read. In some instances within each species, *Ves* reads map to approximately the same location (±10bp). To determine the precision of the estimated Ves insertion site related to all other estimated insertion sites within a ± 10 bp window, each the distance to the nearest neighbor was calculated for each Ves insertion site. Each species is shown along with the number of comparisons made (n).
Figure 4.3  Accuracy of estimated Ves insertion sites shared among species.

Estimated Ves insertion sites within each species (A – Myotis M. auriculus; B – M. dominicensis; C – M. M. horsfieldii; D – M. lucifugus; E – M. M. occultus; F – M. simus; G – M. vivesi) were compared to the six other species examined. For estimated insertion sites that fell within ±100 bp of each other between species, the distance was calculated. Orthologous insertions are expected to occur at the same position, while independent insertions would occur randomly.
Figure 4.4  Sequence statistics per Ves insertion.

(A) After ~5 million reads the number of Ves insertions identified tended to stabilize so that more sequencing did not identify additional insertions, proportionally. (B) The average coverage of each Ves insertion is directly related to the number of read pairs.
Figure 4.5  Results of Ves insertion panel on seven species of *Myotis* with three different loci.

The names of each locus and its location in the *M. lucifugus* genome (Myoluc2.0: GCA_000147115.1) are given. Loci with a Ves insertion are expected to be ~200 bp larger than those lacking an insertion.
References


Murata, S., Takasaki, N., Saitoh, M., Tachida, H. and Okada, N., 1996. Details of retropositional genome dynamics that provide a rationale for a generic division: The distinct branching of all the Pacific Salmon and Trout (Oncorhynchus) from the Atlantic Salmon and Trout (Salmo). Genetics 142, 915-926.


Ray, D., In prep. SINE evolution in vesper bats.


Transposable elements (TEs) are ubiquitous sequence repeats with the capability of replicating themselves within a genome. TEs comprise a significant portion of eukaryotic genomes (Lander et al., 2001; Tenaillon et al., 2011). The primary objective of this research was to use transposable elements (TEs) to better understand genome-level evolution and the contribution of TEs to species diversity.

In Chapter II it was determined that overall TE activity has been steadily declining in *S. tridecemlineatus*. More specifically, the typically ubiquitous L1 activity of mammals has decreased drastically within the last 26 MY. Only three L1 insertions with intact ORF1 sequences were readily identifiable and no intact ORF2 sequences were identified. The last L1 and SINE insertions date to ~5.3 MYA and 4 MYA, respectively, based on genetic distance values between the most recently inserted elements. Based on our inability to computationally identify recently inserted L1 or SINE elements we suggest that *S. tridecemlineatus* is at least experiencing a quiescence of non-LTR retrotransposon activity and such activity has likely been eliminated.

While most TE activity is neutral or deleterious, it may be that TE activity is necessary for long-term evolvability and that species lacking TE activity are more prone to extinction (Oliver and Greene, 2009; Schaack et al., 2010). As more data is generated it will be possible to correlate long-term TE activity with species diversity. The non-LTR
retrotransposon extinction in \textit{S. tridecemlineatus} is only the fourth TE extinction observed in mammals is an important data point to guide our understanding of TE dynamics.

In Chapter III we found that DNA transposons contributed significantly to the miRNA repertoire of \textit{Eptesicusfuscus}. More specifically we were able to temporally isolate the acquisition of these miRNAs. The vesper bat radiation, which includes \textit{E. fuscus}, began around 36 million years (my) ago and resulted in the second most species rich mammalian family (>400 species). Coincident with that radiation was an initial burst of DNA transposon activity that has continued into the present in some species. Such extensive and recent DNA transposon activity has not been observed in any other extant mammal. Deep sequencing of the small RNA fraction from \textit{E. fuscus}, as well as in dog and horse revealed that, while the rate of microRNA (miRNA) origination is similar in all three taxa, 61.1\% of post-divergence miRNAs in \textit{Eptesicus} are derived from TEs compared to only 23.9\% and 16.5\% in the dog and horse, respectively. Not surprisingly given the retrotransposon bias of dog and horse, the majority of TE-derived miRNAs are associated with retrotransposons. In \textit{Eptesicus}, however, 58.7\% of the TE-derived and 35.9\% of the total miRNAs arose not from retrotransposons but from bat-specific DNA transposons. Notably, we observe that the timing of the DNA transposon expansion and the resulting introduction of novel miRNAs coincides with the rapid diversification of the family Vespertilionidae. Furthermore, potential targets of the DNA transposon-derived miRNAs are identifiable and enriched for genes that are important for regulation of transcription.
The vespertilionid radiation was one of the most significant species radiations among extant mammals. We proposed that lineage specific DNA transposon activity lead to the rapid and repeated introduction of novel miRNAs, potentially influencing the diversification of Vespertilionidae. This observation raises additional questions. For example, are there other instances of bursts of TE activity being correlated with species radiations? Are TEs significant contributors to miRNAs in general? Can we assign specific function to the miRNAs deposited during the vespertilionid radiation? Future work will focus on such questions.

In Chapter IV we modified the sequence-capture protocol, ME-Scan, (Witherspoon et al., 2010; Witherspoon et al., 2013) to identify large numbers of Ves SINE insertions in Myotis for use in a retrotransposon-based phylogeny. A well-supported phylogeny of Myotis has been difficult to resolve. Suggested relationships based on morphology were found to be paraphyletic using mitochondrial sequence data (Findley, 1972; Ruedi and Mayer, 2001; Stadelmann et al., 2007). Despite repeated attempts (Stadelmann et al., 2007; Larsen et al., 2012), little of the mitochondrial phylogeny has been confirmed by nuclear sequence data. By adapting the ME-Scan protocol we were able to identify and sequence over 796 thousand insertions from the recently active Ves3ML SINE subfamily in seven species of Myotis. For six of the seven species we were able to identify reads originating from over 76% of all capturable Ves insertion sites. Further, reads identifying Ves insertion sites were accurate within, and among, all taxa analyzed. Our results suggest that the modified ME-Scan method is accurate at identifying insertions within a species, and able to identify shared insertions.
among species. A handful of loci were validated via amplification and sequencing to confirm our results and test for phylogenetic utility.

The above work was exploratory and developmental. Future efforts will focus on using the loci identified in Chapter IV to generate a Ves-based phylogeny of more than 30 New World *Myotis* taxa. Over two hundred loci have been selected based on data presented herein and work is ongoing. Furthermore we believe that this technique will be broadly applicable to the phylogenetic community. Retrotransposon insertions have several advantages over traditional DNA sequence based markers. Our method offers a cost-effective way for targeted sequencing of retrotransposon insertions that is currently unfeasible using traditional Sanger sequencing. Taken together, we believe that our method will make large-scale phylogenomic studies more feasible in the near future.

Each of chapters II-IV is being, or will be, expounded on. Currently efforts are being made to better understand the driving force behind the non-LTR retrotransposon extinction in *S. tridecemlineatus* identified in Chapter II. This work is being pursued by a lab mate and was recently awarded NSF funding. In order to strengthen and confirm our findings from Chapter III-that DNA transposons have significantly contributed to lineage specific miRNAs in vespertilionid bats—we have gathered and are analyzing 10x more data than was used to generate our initial hypotheses. Finally, phylogenetically informative loci identified in Chapter IV are being used to generate a phylogeny of new world *Myotis* in over 48 samples more than 30 taxa.

The primary objective of this dissertation was to use TEs to better understand genome-level evolution and the contribution of TEs to species diversity. During this process we observed a non-LTR retrotransposon extinction event in *S. tridecemlineatus*,
identified DNA transposons as a major source of miRNAs in vespertilionid bats, and developed a method to use Ves SINE insertions as phylogenetic markers in *Myotis*. Prior to this work it was believed that TE extinctions in mammals is rare, TEs were ignored as potential miRNA progenitors, and retrotransposon-based phylogenies focused primarily model organisms. While most research will continue to focus on genes, non-coding RNAs, gene expression, *etc* the work presented here shows that TEs play a significant role at the genotypic (Chapter II) and phenotypic (Chapter III) level. Additionally TEs can be used to trace the evolutionary history of a species (Chapter IV).

In conclusion, TEs represent a powerful evolutionary maker that is underutilized. The work presented here used TEs to address three distinct questions in three distinct ways. When combined with the increasing availability of genomic level data, access to next-generation sequencing platforms, and increasing advances in the fields of bioinformatics and data mining, TEs will be used more frequently to better understand the evolutionary history of species.
References


