1-1-2016

Genes, Transposable Elements, and Small RNAs: Studying the Evolution of Diverse Genomic Components

Michael W. Vandewege

Genes, transposable elements, and small RNAs: Studying the evolution of diverse genomic components

By

Michael W. Vandewege

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 and Plant Pathology

Mississippi State, Mississippi

May 2016

Genes, transposable elements, and small RNAs: Studying the evolution of diverse genomic components

By

Michael W. Vandewege

Approved:

_____
Federico G. Hoffmann
(Major Professor)

_____
David A. Ray
(Committee Member)

_____
Diana C. Outlaw
(Committee Member)

_____
Andy D. Perkins
(Committee Member)

_____
Daniel G. Peterson
(Committee Member)

_____
Kenneth O. Willeford
(Graduate Coordinator)

_____
George M. Hopper
Dean
College of Agriculture and Life Sciences

Name: Michael W. Vandewege

Date of Degree: May 6, 2016

Institution: Mississippi State University

Major Field: Molecular Biology

Major Professor: Federico G. Hoffmann

Title of Study: Genes, transposable elements, and small RNAs: Studying the evolution of diverse genomic components

Pages in Study: 93

Candidate for Degree of Doctor of Philosophy

The evolution of genes and genomes has attracted great interest. The research presented here is an examination of genomes at three distinct levels, protein evolution, gene family evolution, and TE content regulation. First at a genetic level, I conducted an analysis of the salivary androgen-binding proteins (ABPs). I focused on comparing patterns of molecular evolution between the *Abpa* gene expressed in the submaxillary glands of species of New World and Old World muroids and found that in both sets of rodents, the *Abpa* gene expressed in the submaxillary glands appear to be evolving under sexual selection, suggesting ABP might play a similar biological role in both systems. Thus, ABP could be involved with mate recognition and species isolation in New World as well as Old World muroids. Second I examined the largest gene family in vertebrate olfactory receptors (ORs) among birds and reptiles. I found that the number of intact OR genes in sauropsid genomes analyzed ranged over an order of magnitude, from 108 in the lizard to over 1000 in turtles. My results suggest that different sauropsid lineages have highly divergent OR repertoire compositions. These differences suggest that varying rates of gene birth and death, together with selection related to diverse natural histories, have

shaped the unique OR repertoires observed across sauropsid lineages. Lastly, I studied the interactions between transposable elements (TEs) and PIWI-interacting RNAs (piRNAs) among laurasiatherian mammals. piRNAs are predominantly expressed in germlines and reduce TE expression and risks associated with their mobilization. I found that within TE types, families that are the most highly transcribed appear to elicit the strongest ping-pong response. This was most evident among LINEs, but the relationships between expression and PPE was more complex among SINEs. I also found that the abundance of insertions within piRNAs clusters strongly correlated with genome insertions and there was little evidence to suggest that piRNA clusters regulated TE silencing. In summary, the piRNA response is efficient at protecting the genome against TE mobility, particularly LINEs, and can have an evolutionary impact on the TE composition of a genome.

DEDICATION

This dissertation is dedicated to my parents each of whom has sacrificed time and resources 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 Neal Platt, whose assistance was invaluable during my time at MS State. I would also like to thank my current committee members, Drs. Federico Hoffmann, David Ray, Diana Outlaw, Andy Perkins, and Daniel Peterson for their patience and flexibility. Also, I would like to thank my past advisers and committee members, Michael Forstner, Chris Nice, and Noland Martin. I would like to specifically thank my undergraduate adviser(s) Llewellyn Densmore and David “Princess” Rodriguez for seeing potential in me and initiating my career path. Finally I would like to thank my family, Ivana, J.J., and Lilly, for supporting me throughout this entire process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

There is wide variation among eukaryotic genomes reflected in the abundance of DNA and genomic complexity. Eukaryotic genome size can vary from 2.9 megabases (Mb) for some unicellular species to 2900 Mb in humans (Lynch 2007) although the genomes for some land plants and amphibians can be much larger. On average, vertebrates have the largest genomes; despite the large sizes, protein coding DNA only makes up approximately 1-2% of the genomic content, while the remainder is composed of introns, regulatory sequences, transposable elements (TEs), non-coding small RNAs, satellite DNA, and non-coding intergenic space with unknown functions.

How genes and genomes evolve over time has attracted great interest in the scientific community. Evolutionary genomics is a broad field of study that aims to address the increase in genome size and complexity, the rate of nucleotide changes within and among genes, the evolution of gene families, and the presence and evolution of TE content among species. A phenotypic trait can be encoded by one or multiple genes and the evolution of a phenotypic trait is dictated by the underlying genetic code. Since analyzing large datasets has become feasible, great efforts have been made to link genetics to phenotypes and how selection and adaption shapes the genes underlying phenotypes. A relatively famous example is the adult tolerance of lactose among the

descendants of northern Europeans which is associated with two single nucleotide polymorphisms (SNPs) ~14k and 22k upstream of the lactase gene (Poulter et al. 2003).

With the increasing availability of whole genomes, the function of non-coding genomic content has been investigated and debated in recent years (Dunham et al. 2012; Graur et al. 2013). Still, the most commonly studied, and arguably the simplest regions of the genome to study are the protein coding genes. Proteins are considered the major product of genomes and they evolve in a manner that is relatively easy to model (Yang 1998; Yang 2000; Zhang et al. 2005). A single amino acid is coded by a string of three nucleotides (codon) and multiple codons can code for the same amino acid. The redundancy in the genetic code means that any nucleotide change within a codon will either not affect the protein sequence (synonymous mutation) or cause the substitution for a different amino acid (nonsynonymous mutation). The three primary models proposed for protein evolution are neutral evolution, purifying selection, and adaptive (Darwinian) evolution. The null model of gene evolution states that most changes are neutral (Kimura 1983) or nearly neutrally (Ohta 1992), meaning that nucleotide changes occur haphazardly, are often not deleterious and the fixation of a new allele is determined by genetic drift, the strength of which is dependent on population size. However, some gene sequences, like those encoding histone proteins, are highly conserved and exhibit very low rates of change. Among eukaryotic histone H4 genes, the proportion of synonymous differences are at or near saturation, yet the protein sequences are nearly identical, suggesting there is strong selection against any nonsynonymous mutation (Piontkivska et al. 2002). By contrast, genes responsible for immune defense exhibit a higher rate of change at nonsynonymous nucleotide sites than synonymous sites, and this phenomenon

is known as a positive selection (Zhang et al. 1998). Still it should be noted that the roles genetic drift, selection and mutation play on gene and genome evolution is strongly debated among evolutionary biologists. For example, Nei (2013) rejects the claim of natural selection as the major driver of evolution and directs focus toward the importance of mutation on breaking evolutionary constraints. Alternatively, Lynch (2007) proposes that the observed variation in nucleotide composition, TE content, gene birth-and-death rates, and genome size among taxa can be explained by population size and the strength of genetic drift. What is understood is that evolution is a very complicated process where the interactions of selection, population size, mutation, recombination, etc. have varying strengths on shaping gene function and genome content, and genes and genomes have mostly independent evolutionary trajectories.

In this dissertation I used the genomes of mammals, birds and reptiles to study three different aspects of genomic evolution: 1) evolutionary patterns of a single gene among species, 2) the expansion and contraction of the largest gene family in vertebrates and 3) the evolutionary consequences of TE mobility and host defense interactions on genomic content.

In Chapter II, I explored the influence sexual selection may play on protein evolution by studying a protein that has been described as a mate recognition hormone in mice. In rodents, androgen-binding protein is secreted into the saliva and transferred to the pelt during grooming, and evidence suggesting this protein is used for conspecific and mate recognition in the common house mouse (*Mus musculus*) (Laukaitis et al. 1997; Karn and Nachman 1999; Talley et al. 2001; Karn et al. 2002; Bímová et al. 2005; Bímová et al. 2011). I further hypothesized that this protein would be used as a

conspecific recognition protein among all mouse-like rodents (Muroids) and examined the evolution of this protein in both Old and New World Muroids. This study confirmed that several amino acid sites were evolving under positive selection, supporting the hypothesis that this protein is used rodent-wide for mate recognition.

At a higher level, gene duplication is the major mechanism for the origin of new genes and gene functions (Ohno 1970; Zhang 2003). Genes are duplicated primarily through three processes: tandem gene duplication, segmental duplication, and whole genome duplication. If a new gene survives the duplication process intact, the gene copy is considered to be released from functional constraint and permitted to evolve neutrally (Ohno 1970). This evolutionary process will commonly result in a pseudogene; however, a duplicated gene can also develop a novel function and be retained in the genome (Force et al. 1999; Lynch and Conery 2000; Hurley et al. 2005; True and Carroll 2002; Nei and Rooney 2005).

Among vertebrates, the olfactory receptor (OR) gene family is the largest gene family. ORs are cellular membrane G protein-coupled receptors (GPCRs) that communicate environmental cues, such as those signaling food and mates, to the brain (Buck and Axel 1991; Krautwurst et al. 1998; Mombaerts 1999; Fredriksson et al. 2003). Evidence suggests that environmental and niche pressures shape the repertoire of OR genes among genomes (Steiger et al. 2008; Niimura 2009; Hayden et al. 2010; Garrett and Steiper 2014; Hayden et al. 2014; Niimura et al. 2014). The second portion of this research (Chapter III) focused on evolution of the OR family among birds and reptiles (Sauropsids). I used several recently released Sauropsid genomes to investigate how the OR repertoire varies among a vastly diverse group of tetrapods. The OR repertoire ranged

over an order of magnitude (100-1000) and the content among the OR subfamilies was diverse among even relatively closely related species. Yet, in one group of sauropsids, the crocodilians, OR evolution has effectively been stable since radiating from the common ancestor. I suggest that these differences are result of a combination of gene birth and death coupled with selection.

While Chapters II and III primarily investigated protein coding gene evolution, the last chapter focused on non-coding TEs. TEs are segments of DNA that are capable of mobilizing and accumulating within a genome, and often occupy more than half of vertebrate genomes. There are two major classes of TEs. Class I TEs are retrotransposons which mobilize via the reverse transcription of an RNA intermediate (Luan et al. 1993). By contrast, Class II elements (DNA transposons) do not require an RNA intermediate and will directly mobilize. In most vertebrates, Class I elements make up the majority of the TE content. For example, 41% of the human genome is derived from Class I TEs while only 2.8% is derived from DNA transposons (Lander et al. 2001) (de Koning et al. (2011) suggested TE content makes up 70% of the human genome). Because TE mobilization is generally deleterious, there are mechanisms that exist to prevent transposition (Aravin et al. 2007; Jacobs et al. 2014). One of the most understood TE silencing pathway involves PIWI proteins and a class of small RNAs, known as PIWI interacting RNAs (piRNAs) (Brennecke et al. 2007; Carmell et al. 2007; Houwing et al. 2007; Aravin et al. 2008). PIWI proteins are expressed during gametogenesis and their functions have been most studied in mouse testis and *Drosophila* ovaries. Among vertebrates, PIWIs act to silence TEs via two pathways. The primary method is through direct TE transcript digestion (known as the ping-pong cycle) (Siomi et al. 2011) which

appears to occur throughout most of spermatogenesis. In the vertebrate ping-pong model, a PIWI protein, specifically MILI, binds with a piRNA that is anti-sense and complementary to a TE transcript. Through complementary base-pairing, the complex pairs with a TE target and an endonuclease domain cleaves the TE transcript (Aravin et al. 2007). The cleavage of TE transcripts results in a pool of sense piRNAs which are loaded onto MILI and the cycle continues. Another PIWI protein (MIWI2) is also linked to the methylation of TE loci during the early stages of gonad development (Carmell et al. 2007; Kuramochi-Miyagawa et al. 2008; Moralo et al. 2014), however it is unclear how MIWI2 marks TE loci for subsequent methylation. The immediate effects of MILI and MIWI2 deficiency leads to the activation of retrotransposon in the male germline, an arrest of gametogenesis, and complete sterility in male mice (Aravin et al. 2007; Carmell et al. 2007). However, the long term genomic consequences of TE and PIWI interactions are generally unknown, therefore I started to investigate how TE content is (or is not) shaped by the PIWI/piRNA defense. There are several testable predictions of the ping-pong cycle, so I started with the most basal prediction, that the most expressed TE families will be the most represented in the pool of ping-pong piRNAs. I used three laurasiatherian genomes with unique TE compositions to test this hypothesis. I found strong linear relationships between LINE expression and piRNA expression in all three species. By contrast, the second most abundant group of elements, the SINEs, did not fit this pattern and I found major deviations among the three species regarding how PIWIs responded to SINE expression.

Each chapter represents a standalone unit of research at various stages of completion. For this reason, each chapter follows the formatting requirements for the

publishing journals. Chapter II is published in the Journal of Molecular Evolution titled: Evolution of the ABPA subunit of androgen binding protein expressed in the submaxillary glands in New and Old World rodent taxa. Chapter III has been accepted at Genome Biology and Evolution entitled: Contrasting patterns of evolutionary diversification in the olfactory repertoires of reptile and bird genomes. Lastly, Chapter IV is under revision also at Genome Biology and Evolution entitled: Transposable element targeting by piRNAs in Laurasiatherians with distinct transposable element histories.

References

Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764

Aravin AA., Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. Mol. Cell 31:785–799

Bímová B, Karn RC, Piálek J. 2005. The role of salivary androgen-binding protein in reproductive isolation between two subspecies of house mouse: Mus musculus musculus and Mus musculus domesticus. Biol. J. Linn. Soc. 84:349–361

Bímová BV, MacHolán M, Baird SJE, Munclinger P, Dufková P, Laukaitis CM, Karn RC, Luzynski K, Tucker PK, Piálek J. 2011. Reinforcement selection acting on the European house mouse hybrid zone. Mol. Ecol. 20:2403–2424

Brennecke J, Aravin AA., Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell 128:1089–1103

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–187

Carmell M a., Girard A, van de Kant HJG, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. 2007. MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. Dev. Cell 12:503–514

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitve elements may comprise over two-thirds of the human genome. PloS Genet. 7:e1002384

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a., Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol. Pharmacol. 63:1256–1272

Garrett EC, Steiper ME. 2014. Strong links between genomic and anatomical diversity in both mammalian olfactory chemosensory systems. Proc. Biol. Sci. 281:20132828

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall R a., Elhaik E. 2013. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of encode. Genome Biol. Evol. 5:578–590

Hayden S, Bekaert M, Crider T a., Mariani S, Murphy WJ, Teeling EC. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. 20:1–9

Hayden S, Bekaert M, Goodbla A, Murphy WJ, Dávalos LM, Teeling EC. 2014. A cluster of olfactory receptor genes linked to frugivory in bats. Mol. Biol. Evol. 31:917–927

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov D V., Blaser H, Raz E, Moens CB, et al. 2007. A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. Cell 129:69–82

Hurley I, Hale ME, Prince VE. 2005. Duplication events and the evolution of segmental identity. Evol. Dev. 7:556–567

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516:242–245

Karn RC, Nachman MW. 1999. Reduced nucleotide variability at an androgen-binding protein locus (Abpa) in house mice: evidence for positive natural selection. Mol. Biol. Evol. 16:1192–1197

Karn RC, Orth A, Bonhomme F, Boursot P. 2002. The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. Mol. Biol. Evol. 19:462–471

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press

Krautwurst D, Yau K, Reed RR. 1998. Identification of ligands for olfactory receptors. Cell 95:917–926

Kuramochi-Miyagawa S, et al. 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. 22:908–917

Lander ES, al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921

Laukaitis CM, Critser ES, Karn RC. 1997. Salivary androgen-binding protein (ABP) mediates sexual isolation in Mus musculus. Evolution 51:2000–2005

Luan DD, Horman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposons. Cell 72:595–605

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lynch M. 2007. The origins of genome architecture. Sunderland, MA: Sinauer Associates

Moralo A, Falciatori I, Hodges E, Aravin AA, Marran K, Rafii S, McCombie WR, Smith AD, Hannon GJ. 2014. Two waves of de novo methylation during mouse germ cell development. Genes Dev. 28: 1544–1549

Mombaerts P. 1999. Seven-transmembrane proteins as odorant and chemosensory receptors. Science 286:707–711

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39:121–152

Nei M. 2013. Mutation-Driven Evolution. Oxford: Oxford University Press

Niimura Y, Matsui A, Touhara K. 2014. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res. 24:1485–1496

Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol. Evol. 1:34–44

Ohno S. 1970. Evolution by Gene Duplication. New York, NY: Springer-Verlag

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23:265–286

Piontkivska H, Rooney AP, Nei M. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol Biol Evol. 19:689–697

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM. 2003. The causal element for the lactase persistenence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet. 67:298–311

Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. Nat. Rev. Mol. Cell Biol. 12:246–258

Steiger SS, Fidler AE, Valcu M, Kempenaers B. 2008. Avian olfactory receptor gene repertoires: evidence for a well-developed sense of smell in birds? Proc. Biol. Sci. 275:2309–2317

Talley HM, Laukaitis CM, Karn RC. 2001. Female preference for male saliva: implications for sexual isolation of Mus musculus subspecies. Evolution 55:631–634

True JR, Carroll SB. 2002. Gene co-option in physiological and morphological evolution. Annu. Rev. Cell Dev. Biol. 18:53–80

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51:423–432

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. U. S. A. 95:3708–3713

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298

CHAPTER II

EVOLUTION OF THE ABPA SUBUNIT OF ANDROGEN-BINDING PROTEIN EXPRESSED IN THE SUBMAXILLARY GLANDS IN NEW AND OLD WORLD RODENT TAXA

Introduction

Speciation can be driven by the isolation of two populations through geographic, temporal, ecological, or behavioral barriers (see Coyne and Orr [2004] for an extended discussion). In many mammals olfaction is a dominant sensory modality and chemical cues can be used to convey information about individuality. In the common house mouse (*Mus musculus*) subspecies complex, the salivary androgen-binding proteins (ABPs) are hypothesized to be a component of such a cue, as they are thought to mediate mate recognition (Laukaitis et al. 1997; Talley et al. 2001). ABPs are secretoglobins (Klug et al. 2000; Laukaitis et al. 2005; Laukaitis and Karn 2005; Mukherjee and Chilton 2000) present in the saliva following expression within the submaxillary, sublingual, and parotid glands (Dlouhy et al. 1986; Laukaitis et al. 2005). The putative biological function of mouse salivary ABPs are that of a pheromone, mediating mate selection resulting in assortative mating and incipient reinforcement at the edges of the house mouse hybrid zone in Europe (Bímová et al. 2005; Bímová et al. 2011).

ABP is secreted into the saliva and transferred to the pelage and environment after grooming, allowing an animal to mark territory with a biochemical signal. From a

genomic standpoint, the mouse ABP system was thought to consist of three single copy genes, *Abpa*, *Abpb*, and *Abpg*, encoding for the three separate subunits, ABPA, ABPB, and ABPG, respectively (Dlouhy et al. 1987; Karn and Laukaitis 2003). Genomic comparisons, however, revealed a more complex pattern (Emes et al. 2004; Laukaitis et al. 2008; Karn and Laukaitis 2009). Most mammals have a single *Abpa* gene encoding an ABPA protein and a single *Abpbg* gene encoding an ABPBG protein in the *Abp* locus, however, rat and house mouse possess multiple paralogs in the corresponding location (Laukaitis et al. 2008; Karn and Laukaitis 2009). In the case of house mouse there are 64 *Abp* paralogs (30 *Abpa* and 34 *Abpbg* genes) over a 3 megabase (Mb) region, whereas the rat genome has 6 paralogs (3 *Abpa* and 3 *Abpbg* paralogs). Analyses of intron sequences suggest these expansions occurred independently in these two species (Laukaitis et al. 2008). In house mouse, the *Abpa27* paralog has been the most studied gene in the *Abp* locus. The translated gene ABPA27 forms a dimer with either ABPBG26 or ABPBG27 via disulfide bridges and is expressed in the submaxillary glands of both males and females. Evidence suggests the ABPA27 subunit plays a significant role in conspecific recognition and mate selection (Bímová et al. 2005; Hwang et al. 1997; Karn et al. 2002; Laukaitis et al. 1997; Talley et al. 2001). Interestingly, recent studies also suggest that additional *Abpbg* paralogs secreted into the saliva of house mouse, specifically *Abpbg26* and *Abpbg27* show patterns of molecular variation suggesting adaptive evolution (Laukaitis et al. 2012).

Protein coding genes associated with reproductive and chemosensory roles have a tendency to evolve rapidly and display signatures of positive selection (Duret and Mouchiroud 2000; Kosiol et al. 2008; Park et al. 2011; Swanson and Vacquier 2002;

Torgerson et al. 2002). In line with these expectations, comparisons among different *Mus sp.* alleles revealed high rates of nonsynonymous to synonymous substitutions, and comparative analyses among lineages within the *Mus* species complex have shown that distinct *Abpa27* alleles are fixed in the different subspecies of *M. musculus*, and molecular evolution analyses of mouse *Abpa27* sequences detected strong signals of Darwinian selection (Emes et al. 2004; Hwang et al. 1997; Karn and Nachman 1999; Karn et al. 2002). Therefore the combination of controlled mate choice experiments and the observed genetic signatures consistent with positive selection suggests that ABPs probably play a role in olfactory communication, assortative mating, and incipient reinforcement of reproductive isolation (Bímová et al. 2005; Bímová et al. 2011; Karn and Dlouhy 1991; Karn et al. 2002; Laukaitis et al. 2005). The expression and evolution of ABP has been extensively studied in *Mus* (Dlouhy et al. 1987; Emes et al. 2004; Karn et al. 2010; Laukaitis et al. 2008; Laukaitis and Karn 2005; Karn and Laukaitis 2009). However, few data have been presented for other rodent genera*:* Karn and Dlouly (1991) extracted and identified ABP in the saliva for New World rodents, and Laukaitis et al. (2008) identified two putative paralogs of *Abpa* in *Apodemus sylvaticus*.

In the present study, my objective was to explore whether the patterns of variation in the *Abpa* genes that encode for the ABPA subunit expressed in the submaxillary glands are similar in New and Old Word rodents. In particular, I focused on comparing patterns of molecular evolution between the *Abpa27* gene in house mouse, with their functional counterparts in other Old and New World rodent species. To do so, I analyzed partial sequence data from mRNA transcripts isolated from submaxillary glands of selected New and Old World muroids and used maximum likelihood methods to characterize patterns

of molecular evolution. My results indicate that the *Abpa* genes expressed in the submaxillary glands are evolving rapidly in both rodent groups, and that a similar set of codon positions appears to be under positive selection in the two systems. Given that proteins involved in species isolation tend to evolve rapidly, these findings would suggest that ABPs might play similar roles related to maintaining species boundaries in both sets of species.

Materials and Methods

Data Collection

Representatives of New World muroids were collected from West Texas and Ecuador and Old World muroid representative rodents were collected from northern Ukraine. Voucher specimens are stored in the Natural Science Research Laboratory collections at the Texas Tech Museum. The submaxillary gland was removed from the animal and stored at -70° C. Isolation of mRNA, cDNA preparation, PCR amplification of expressed *Abpa*, and sequencing of amplicons followed Wickliffe et al. (2002). In addition, I obtained additional known expressed *Abpa* sequences from *Spermophilus tridecemlineatus*, *Mus sp.*, and *Rattus norvegicus* from Karn et al. (2002) and Laukaitis et al. (2008). *Mus spicilegus* was not included from Karn et al. (2002) given the exact sequence identity to *M. macedonicus.* In the case of heterozygous individuals, h*aplotypes were resolved* using PHASE version 2.1(Stephens and Donnelly 2003*)*. Previously unpublished *Abpa* sequences were deposited in GenBank under the accession numbers JX275970–JX275986.

Sequence Alignment and Analysis

Partial *Abpa* cDNA sequences were aligned by codons with MUSCLE (Edgar 2004) as implemented in MEGA 5 (Tamura et al. 2011). I explored alignment sensitivity by comparing the MUSCLE alignment to the results of ClustalW (Thompson et al. 1994), MAFFT (Katoh et al. 2005), and PRANK (Löytynoja and Goldman 2005). Alternative alignment strategies yielded the same alignment, which was used for downstream analyses. The intraspecific number of haplotypes, nucleotide diversity, and the uncorrected average pairwise number of differences for nucleotide and amino acid changes were determined in MEGA 5. Intra-generic number of nonsynonymous and synonymous changes were counted within *Apodemus*, *Mus*, *Peromyscus*, and *Reithrodontomys*. Once redundant alleles were removed from the alignment, I reconstructed phylogenetic relationships using maximum likelihood as implemented in Treefinder version March 2011 (Jobb et al. 2004), and I evaluated support for the nodes with 1,000 bootstrap pseudoreplicates. I used the 'propose model' tool of Treefinder to select the best-fit models of nucleotide substitution, with an independent model at each codon position in nucleotide analyses. Model selection was based on the Akaike information criterion with correction for small sample size.

We then estimated patterns of molecular evolution using the maximum likelihood approach described by (Goldman and Yang 1994) as implemented in CODEML in PAML (Yang 2007). In order to estimate the putative role of negative and positive selection, I compared the rate of nonsynomous substitution per nonsynonymous site (d_N) to the rate of synonymous substitution per synonymous site (d_S). The d_N/d_S ratio (also labeled as ω) can be used to measure the selective regime of a given codon, as similar

rates of nonsynonymous and synonymous substitution ($\omega \approx 1$) are indicative of neutral evolution, an excess of synonymous mutations ($\omega < 1$) is indicative of purifying, or negative selection, and an excess of nonsynonymous mutations ($\omega > 1$) is indicative of positive Darwinian selection, or adaptive evolution. I compared models that allow ω to vary among codons in the alignment (M0 vs. M3, M1a vs. M2a, M7 vs. M8, and M8a vs. M8). In all cases, I used likelihood ratio tests (LRTs) to compare nested sets of model (Yang 1998). Bayesian Empirical Bayes (BEB) was used to calculate the posterior probabilities for sites under positive selection in models M2a and M8 (Yang et al. 2005). These analyses were performed separately for New World and Old World muroids, because orthology between *Abpa27* in *Mus* and New World taxa cannot be guaranteed. To account for the potential problem generated by intraspecific polymorphism, these analyses were repeated using only one randomly selected representative from each species. Phylogenetic trees were reconstructed from each subset under the same parameters as the full dataset.

Residues were determined to be under selection if a position was predicted to be under positive selection with posterior probability >0.9 in one model and >0.5 in one other model. I predicted the *Mus* ABPA27 amino acid structure was using Phyre2 (Kelley and Sternberg 2009). Residues under selection in New World rodents and Old World rodents were mapped onto the structure. Swiss-PDBviewer (Guex and Peitsch 1997) was used to manipulate the protein structure and the 3D image was rendered with POV-Ray (www.povray.org). Relative solvent accessibilities of residues were calculated based on criteria given in (Emes et al. 2004). Relative accessibility values were divided into three categories, buried (<9% relative accessibility), intermediate (9 –35% relative

accessibility) and exposed (>35% relative accessibility) (Emes et al. 2004; Rost and Sander 1994).

PolyPhen-2 (Polymorphism Phenotyping ver. 2.0; http://genetics.bwh.harvard.edu/pph2/index.shtml), an *in silico* tool for predicting structural and functional effects of amino acid substitutions on proteins, was used to explore the possible impacts of amino acid variants that appeared to be under strong, positive selection within and among New World and Old World rodent lineages (Adzhubei et al. 2010). Default options were used for all analyses.

Results

Description of Data

We obtained sequences corresponding to 213 bp of the 279 nucleotides of coding sequence of *Abpa* transcripts from submaxillary glands from 34 rodent specimens. Twenty-seven samples were collected from the New World species, 9 *Peromyscus leucopus*, 6 *Peromyscus maniculatus*, 3 *Reithrodontomys fulvescens*, 1 *Akodon aerosus*, 6 *Sigmodon hispidus*, 2 *Oligoryzomys microtis*; and six samples were collected from the Old World, 1 *Microtus oeconomus,* 1 *Apodemus agrarius*, 1 *Apodemus flavicolis* and 3 *Apodemus sylvaticus*. The *Apodemus* sequences were previously isolated and described in Wickliffe et al. (2002). Even though *Microtis oeconomus* was collected in northern Ukraine, the genus is more closely related to New World muroid rodents. Therefore the *M. oeconomus Abpa* cDNA sequence was compared to the remaining New World muroid rodents in this study. The *Abpa* sequence of *Spermophilus tridecemlineatus* reported by Laukaitis et al. (2008), which corresponds to a single-copy gene in the spetri2 assembly of the squirrel genome, was included as an outgroup for phylogenetic analyses. All

alignment strategies yielded similar results, with no gaps or premature stop codons found among cDNA derived sequences. Intraspecific pairwise comparisons of the number of nucleotide substitutions ranged from 0 to 14 and from 0 to 12 when comparing amino acid sequences (Table 1). Despite the relatively small sample sizes, I found that *A. sylvaticus*, *P. leucopus*, *P. maniculatus*, and *R. fulvescens* were polymorphic for the *Abpa* gene in this sample, whereas there was no sequence variation among the 6 specimens of *Sigmodon hispidus* (Table 1).

We then estimated phylogenetic relationships among the different alleles using maximum likelihood, excluding redundant haplotypes. In the resulting tree the New World and Old World muroid sequences fell in reciprocally monophyletic clades, as expected given current estimates of organismal phylogeny (Fig. 1; Jansa and Weksler 2004; Steppan et al 2004). In general, the sequences of given genera were monophyletic, with the exception of the rat *Abpa* paralogs (Fig. 1). Phylogenies based on intronic sequence suggest that the presence of multiple paralogs of *Abpa* in rat and mouse derive from lineage-specific expansions (Laukaitis et al. 2008). My results differ from those of Laukaitis et al. (2008) as the rat paralogs did not form a monophyletic clade. However, an approximately unbiased topology test (Shimodaira 2002) could not discriminate between the maximum-likelihood tree and a tree where the rat paralogs were constrained to monophyly. This issue warrants further attention once intronic sequences from the *Abp* genes of a wider selection of Old World muroid rodents become available.

Patterns of molecular evolution

As a starting point, I compared nonsynonymous and synonymous substitution rates in a pairwise manner. I detected an excess of nonsynonymous substitutions relative

to synonymous substitutions in intrageneric and intraspecific comparisons within the Old World and New World muroid groups. This is especially noticeable in comparisons within *R. fulvescens*. In line with these results, estimates of d_N/d_S (=ω) for the whole fragment sequenced were high for both datasets, 0.9 for New World muroids and 1.18 for Old World rodents. In general, estimates of d_N/d_S averaged over all sites in excess of 0.5 have been considered as suggestive of positive Darwinian selection (Swanson et al. 2004).

We then compared different models of molecular evolution that explore variation in ω among codons in a tree-based approach. I first explored whether there was evidence of variation in ω among codons. I compared models M0, which assigns the same value of ω to all codons with M3, a model that groups codons in three separate classes with independent estimates of ω for each class. There was significant variation in ω, as the LRT rejected M0 in favor of M3 in both cases (Table 2). I then looked for evidence of positive Darwinian selection among samples by comparing models that allow a class of sites to have $\omega > 1$, M2a and M8, with the corresponding models that restrict all ω estimates to be ≤ 1, M1a, M7, and M8a. In all comparisons, M1a vs. M2a, M7 vs. M8, and M8a vs. M8, the models that allow a class of sites to have $\omega > 1$ were favored by the corresponding LRTs (Table 2). Under the criteria given, 5 and 12 sites were found to be evolving under positive selection in New World and Old World muroids, respectively (Table 2). Similar results were obtained when I reran the analyses using only one representative from each species (Table 2). Most of the sites were found to be in exposed regions of the ABPA27 structure in both the New World and Old World groups (Fig. 2). I then focused on amino acid variants at sites 68 and 69 (90 and 91 in the accessioned

amino acid sequence for *M. musculus* [http://www.ncbi.nlm.nih.gov/protein/NP_033726.1]) as these sites exhibited marked positive selection in the New World and somewhat in the Old World muroids. *In silico* simulations with PolyPhen-2 suggest that changes in either of these sites do not appear to alter the structure or putative function of the *Abpa*.

Discussion

In house mouse the ABP heterodimers present in the saliva appear to play a behavioral role in reproductive isolation (Laukaitis et al. 1997; Talley et al. 2001; Laukaitis et al. 2012). Most of these studies have focused on the *Abpa27* gene, which is expressed in the submaxillary glands in both male and female *Mus musculus* (Laukaitis et al. 2005). In the *M. musculus* subspecies complex, several lines of evidence suggest salivary ABP plays an important role in assortative mating. First, alternative alleles are fixed in each *M. musculus* subspecies at the *Abpa27* locus (Hwang et al. 1997; Karn and Dlouhy 1991). Second, in laboratory experiments, female mice preferred to associate and mate with males of their own *Abpa27* genotype significantly more often than with males carrying a different allele (Laukaitis et al. 1997; Talley et al. 2001). Lastly, comparisons of evolutionary rates among *Abpa27* alleles in different species of mice revealed a large excess of nonsynonymous substitutions over synonymous substitutions consistent with positive Darwinian selection (Hwang et al. 1997; Karn and Nachman 1999; Karn et al. 2002).

Because the rat and mouse *Abpa* repertoires derive from largely independent sets of duplications, resolving orthology for the expressed genes and predicted proteins can be difficult. However, the evidence at hand indicates that all of the *Abpa* genes in mammals

derive from the single copy ancestral *Abpa* gene present in the common ancestor of placental mammals. The most recent assessments of variation in the gene complement of the *Abp* gene family (Emes et al. 2005, Laukaitis et al. 2008, and Karn and Laukaitis 2009) indicate that the duplications that gave rise to the presence of multiple *Abpa* genes in mouse genome are specific to the mouse lineage (see Fig. 1), therefore, I inferred that all mouse *Abpa* paralogs are co-orthologs with the ancestral single copy *Abpa* gene. I have no evidence of the presence of multiple *Abpa* paralogs in New World rodents; thus, I assumed that *Abpa* gene from New World rodents is an ortholog of the ancestral single-copy gene. Because of this shared ancestry and tissue-specific patterns of expression, I assumed that *Abpa27* genes of house mouse were functionally equivalent to the *Abpa* expressed in the submaxillary gland of other rodents. To ensure this, I extracted and sequenced cDNA from the salivary glands in New World rodent group using primers designed to amplify *Abpa27* in *Mus* (Hwang et al. 1997). If the proposed role of the APBA subunit of New World rodents is to act as a mate choice pheromone, as the ABPA27 subunit does in *Mus*, I expected to observe similar patterns of molecular evolution in New and Old World rodents. In this case, I addressed this question by comparing patterns of molecular evolution between Old World and New World muroids.

These results suggest that both the New and Old World muroid *Abpa* sequences share the signature of positive Darwinian selection. In both sets of sequences I found high rates of nonsynonymous substitutions within and among genera, consistent with positive selection. In particular, all analyses indicate that a similar subset of positions appear to be under positive selection within these two separate rodent lineages. These results suggest that changes in a few key residues might play a significant role in the evolution of this

protein. Also, a prediction of solvent accessibility is consistent with previous results where the majority of sites under positive selection are mostly in exposed regions of the protein (Fig. 2). Because there is strong evidence that *Abpa27* plays an important role in speciation among Old World muroids, I speculate that ABPA subunit analyzed might be playing a similar role among the New World rodents studied.

Interestingly, the patterns of intraspecific variation I identified are not entirely consistent with those reported for the *Mus musculus* subspecies complex. The signal of positive selection detected among New World muroids is not as strong as the one detected among Old World rodents. There were five sites under selection inferred among New World rodents compared to the 12 sites detected in the Old World muroids. Additionally, despite limited sample size, I found significant levels of intraspecific variation in species of the genera *Apodemus*, *Peromyscus* and *Reithrodontomys*. Thus, the data would suggest that the putative role of the *Abpa* gene in New World muroids might be slightly different relative to the *M. musculus* complex where different alleles segregate with different subspecies (Bímová et al. 2005; Bímová et al. 2011). The fact that the *Abpa27* paralog is fixed in the members of the two subspecies *M. m. musculus* and *M. m. domesticus* and the ABPA27 subunit is apparently involved with reinforcing reproductive barriers in the Central Europe hybrid zone between these two subspecies, could account for the higher rate of molecular evolution in this system relative to New World rodents (Bímová et al. 2005; Bímová et al. 2011).

In rodents, chemical cues and pheromones play a central role in individual recognition as evidenced by the approximately 1000 functional olfactory receptors and 212 vomeronasal receptors found in the mouse genome (Shi and Zhang 2009; Zhang and

Firestein 2002). Currently, the primary hypotheses proposed to explain patterns of ABP evolution are related to assortative mating, and the coevolution of the ABP pheromone and vomeronasal (VNO) receptors of the V1R receptor family (Karn et al. 2010). As suggested by Karn et al. (2010), it may be that these two systems, ABP and V1R receptors, are coevolving to promote and maintain species boundaries. It would be interesting to identify the V1Rs involved in this putative interaction, and evaluate whether positive selection is also acting on their evolution.

Acknowledgments

We thank R. Baker, R. Chesser, B. Rodgers, M. Bondarkov, and S. Gaschak for access to collecting sites in northern Ukraine as part of another research project. I also want to thank two anonymous reviewers whose comments and suggestions greatly improved the quality of this manuscript. All field and laboratory studies were conducted with appropriate permits and approved institutional protocols and in accord with the NIH Guidelines for the Care and Use of Laboratory Animals. Research was partially funded by Texas Tech University funding to CJP and directed funding for the DOE (CJ Phillips, R. Chesser, R. Baker, co-PIs). FGH acknowledges grant support from the National Science Foundation (EPS-0903787).

Table 2.1 ABPA sequence information for included taxa.

Species	*N*	*h*	π	*distance (nuc)*	*distance (aa)*
A. sylvaticus	4	4	0.043	9.3 (6-14)	7.5 (4-12)
P. leucopus	12	7	0.008	1.8 (0-4)	0.8 (0-2)
P. maniculatus	7	3	0.009	2.0 (0-4)	1.5 (0-3)
O.microtis	2	1	0	0	0
R. fulvescens	3	3	0.048	10 (1-15)	6.7 (1-10)
S. hispidus	6	1	0	0	0

Abpa sequence statistics for intraspecific samples. The number of samples sequenced (*N*), number of unique haplotypes (*h*) and nucleotide diversity (π) are listed along with the average (and range in parentheses) pairwise number of differences (*distance*) for both nucleotide (*nuc*) and amino acid (*aa*) sequences.

Table 2.2 Site models positive selection tests.

Group	*n*	2(ΔL)			M2a sites under selection	M8 sites under selection
		M2a vs M1a	M3 vs M0	M8 vs M8a		
New World	17	14.54*	67.06**	14.34*	10 11 47 **68 69**	7 <u>10</u> <u>11</u> 12 16 23 32 33 40 43 <u>47</u> 61 62 65 66 67 **68 69**
New World (single sequence)	7	8.18*	51.20**	8.08*	11 40 47 62 **68** 69	7 10 <u>11</u> 15 16 18 23 25 26 27 29 30 32 33 <u>40</u> 43 **47** 55 57 58 61 <u>62</u> 66 **68** <u>69</u>
Old World	16	18.96**	38.76**	18.90**	<u>12</u> 14 **18** 23 27 30 **32 33** 34 36 <u>43</u> 45 <u>55</u> 59 60 68 **69**	7 **12** <u>14</u> **18** <u>23</u> 27 30 **32 33** <u>34</u> 36 **43** <u>45</u> 48 51 <u>55</u> 59 <u>60</u> 68 **69**
Old World (single sequence)	13	17.23**	31.12**	17.09**	<u>12</u> 14 **18** 23 30 **32 33** 36 43 45 <u>55</u> 59 60 68 69	**12** 14**18** <u>23</u> 27 30 **32 33** 36 43 **45** 51 <u>55</u> <u>59</u> <u>60</u> 68 <u>69</u>

Tests of positive selection for *n* sequences among sites for the New World and Old World groups. Sites listed had a posterior probability >0.5. Underlined positions were predicted to be under positive selection with a posterior probability 0.9 – 0.949 and bold sites had a posterior probability >0.95. *$p < 0.001$, **$p < 0.0001$

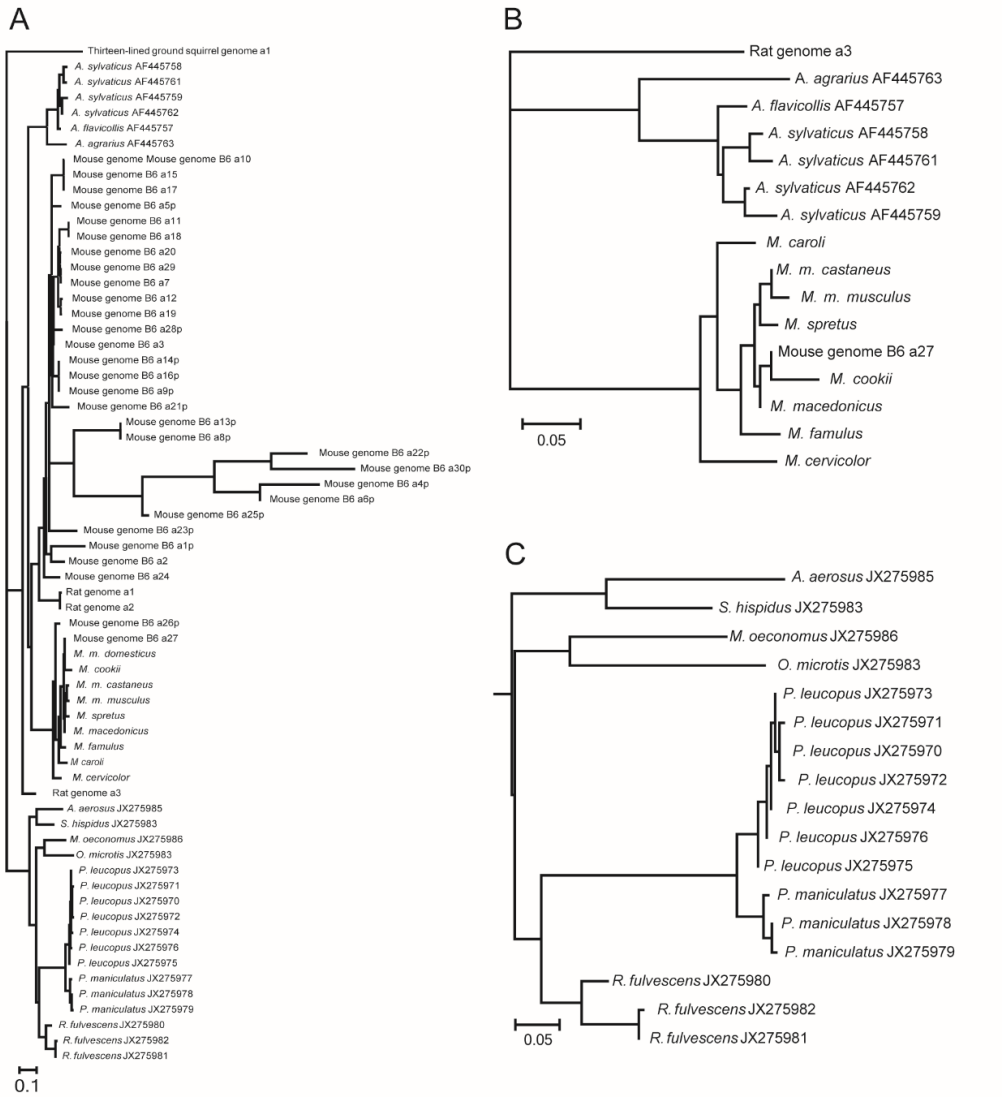

Figure 2.1 ABPA phylogenetic trees

Phylogenetic reconstructions from the 210 bp fragment of *Abpa*. A) Maximum likelihood phylogeny for all newly isolated *Abpa* cDNA sequence, and previously published Old World orthologs and paralogs, excluding redundant haplotypes. B) Maximum likelihood phylogeny for the Old World *Abpa* sequences included in the molecular evolution analyses and C) maximum likelihood phylogeny for the New World *Abpa* sequences included in the molecular evolution analyses.

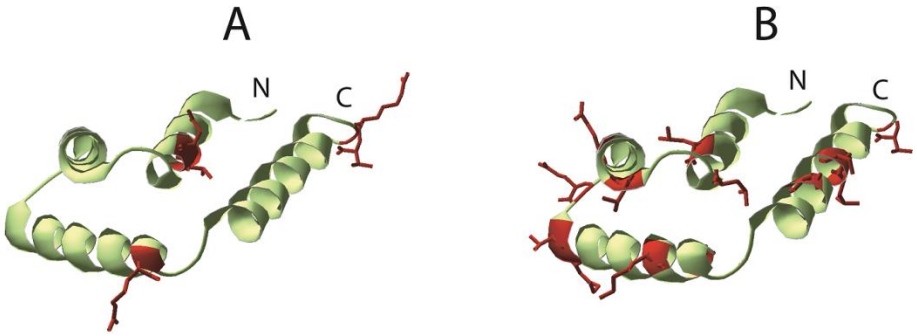

Figure 2.2 Sites under positive selection in ABPA

Sites found to be under positive selection for New World muroids (A) and Old World muroids (B) mapped to the *Mus* ABPA27 subunit. Codons and their side chains predicted to be under positive selection are colored red.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat. Methods 7:248–249

Bímová B, Karn RC, Piálek J (2005) The role of salivary androgen-binding protein in reproductive isolation between two subspecies of house mouse: *Mus musculus musculus* and *Mus musculus domesticus*. Biol. J. Linn. Soc. 84:349–361

Bímová BV, Macholán M, Baird SJE, Munclinger P, Dufková P, Laukaitis CM, Karn RC, Luzynski K, Tucker PK, Piálek J (2011) Reinforcement selection acting on the European house mouse hybrid zone. Mol. Ecol. 20:2403–2424

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates Sunderland, MA

Dlouhy SR, Nichols WC, Karn RC (1986) Production of an antibody to mouse salivary androgen binding protein (ABP) and its use in identifying a prostate protein produced by a gene distinct from Abp. Biochem. Genet. 24:743–763

Dlouhy SR, Taylor BA, Karn RC (1987) The genes for mouse salivary androgen-binding protein (ABP) subunits alpha and gamma are located on chromosome 7. Genet. 115:535–543

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17:68–74

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797

Emes RD, Riley MC, Laukaitis CM, Goodstadt L, Karn RC, Ponting CP (2004) Comparative evolutionary genomics of androgen-binding protein genes. Genome Res. 14:1516–1529

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. Electrophor. 18:2714–2723

Hwang JM, Hofstetter JR, Bonhomme F, Karn R (1997) The microevolution of mouse salivary androgen-binding protein (ABP) paralleled subspeciation of *Mus musculus*. J. Hered. 88:93 –97

Jansa SA, Weksler M (2004) Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. Mol. Phylogen. Evol. 31:256–276

Jobb G, Von Haeseler A, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4:18

Karn RC, Dlouhy SR (1991) Salivary androgen-binding protein variation in *Mus* and other rodents. J. Hered. 82:453–458

Karn RC, Laukaitis CM (2003) Characterization of two forms of mouse salivary androgen-binding protein (ABP): Implications for evolutionary relationships and ligand-binding function. Biochem. 42:7162–7170

Karn RC, Laukaitis CM (2009) The mechanism of expansion and the volatility it created in three pheromone gene clusters in the mouse (*Mus musculus*) genome. Genome Biol. Evol. 1:494–503

Karn RC, Nachman MW (1999) Reduced nucleotide variability at an androgen-binding protein locus (*Abpa*) in house mice: evidence for positive natural selection. Mol. Biol. Evol. 16:1192–1197

Karn RC, Orth A, Bonhomme F, Boursot P (2002) The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. Mol. Biol. Evol. 19:462–471

Karn RC, Young JM, Laukaitis CM (2010) A candidate subspecies discrimination system involving a vomeronasal receptor gene with different alleles fixed in *M. m. domesticus* and *M. m. musculus*. PLoS One 5:e12638

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518

Kelly LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the Phyre server. Nat. Protoc. 4:363–371

Klug J, Beier HM, Bernard A, Chilton BS, Fleming TP, Lehrer RI, Miele L, Pattabiraman N, Singh G (2000) Uteroglobin/Clara Cell 10-kDa Family of Proteins: Nomenclature Committee Report. Ann. N. Y. Acad. Sci. 923:348–354

Kosiol C, Vinař T, Da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. PLoS Genet. 4:e1000144

Laukaitis CM, Critser ES, Karn RC (1997) Salivary androgen-binding protein (ABP) mediates sexual isolation in *Mus musculus*. Evol. 51:2000–2005

Laukaitis CM, Dlouhy SR, Emes RD, Ponting CP, Karn RC (2005) Diverse spatial, temporal, and sexual expression of recently duplicated androgen-binding protein genes in *Mus musculus*. BMC Evol. Biol. 5:40

Laukaitis CM, Heger A, Blakley TD, Munclinger P, Ponting CP, Karn RC (2008) Rapid bursts of androgen-binding protein (Abp) gene duplication occurred independently in diverse mammals. BMC Evol. Biol. 8:46

Laukaitis CM, Karn RC (2005) Evolution of the secretoglobins: A genomic and proteomic view. Biol. J. Linn. Soc. 84:493–501

Laukaitis CM, Mauss C, Karn RC (2012) Congenic strain analysis reveals genes that are rapidly evolving components of a prezygotic isolation mechanism mediating incipient reinforcement. PloS one 7:e35898

Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. 102:10557–10562

Mukherjee AB, Chilton BS (2000) The uteroglobin/Clara cell protein family. Ann. N. Y. Acad. Sci. 923:1–358

Park SH, Podlaha O, Grus WE, Zhang J (2011) The microevolution of V1r vomeronasal receptor genes in mice. Genome Biol. Evol. 3:401–412

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins: Struct., Funct. Bioinform. 20:216–226

Shi P, Zhang J (2009) Extraordinary diversity of chemosensory receptor gene repertoires among vertebrates. Results Probl. Cell Differ. 47:1–23

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. 73:1162–1169

Steppan SJ, Adkins RM, Anderson J (2004) Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Syst. Biol. 53:533–553

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nature Rev. Genet. 3:137–144

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genet. 168:1457–1465

Talley HM, Laukaitis CM, Karn RC (2001) Female preference for male saliva: implications for sexual isolation of Mus musculus subspecies. Evol. 55:631–634

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680

Torgerson DG, Kulathinal RJ, Singh RS (2002) Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. Mol. Biol. Evol. 19:1973–1980

Wickliffe JK, Lee VH, Smith E, Tandler B, Phillips CJ (2002) Gene expression, cell localization, and evolution of rodent submandibular gland androgen-binding protein. Eur. J. Morphol. 40:257–260

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118

Zhang X, Firestein S (2002) The olfactory receptor gene superfamily of the mouse. Nat. Neurosci. 5:124–1

CHAPTER III

CONTRASTING PATTERNS OF EVOLUTIONARY DIVERSIFICATION IN THE OLFACTORY REPERTOIRES OF REPTILE AND BIRD GENOMES

Introduction

In vertebrates, the ability to detect odors is mediated by olfactory receptors (ORs), a type of transmembrane G protein-coupled receptor (GPCR) that mediates interactions between the cell and its surroundings. Structurally, GPCRs have seven α-helical transmembrane domains bound to a G-protein, and the binding of extracellular ligands triggers conformational changes that, in turn, lead to intracellular signaling cascades (Fredriksson et al. 2003). Vertebrate ORs belong to the rhodopsin-like group of GPCRs, which includes receptors that mediate the detection of hormones, neurotransmitters, and photons (Fredriksson et al. 2003). Vertebrate ORs are primarily expressed in the olfactory epithelium of the nasal cavity, where they bind odorants, and transmit the resulting nerve impulse to the brain (Buck and Axel 1991; Mombaerts 1999). The OR repertoires of amniote vertebrates are dominated by two major groups of ORs, Class I ORs, which appear to have a higher affinity for hydrophilic ligands, and Class II ORs, which generally bind hydrophobic ligands (Saito et al. 2009).

Genomic surveys have revealed that ORs represent the largest vertebrate gene family (Zhang and Firestein 2002), and indicate that the numbers and diversity of ORs vary widely among vertebrates, even between closely related taxa (Niimura and Nei

2005b; Nei et al. 2008). There is debate regarding the relative influence of different evolutionary forces in shaping OR repertoires. Nei et al. (2008) suggests that OR evolution is largely a neutral process, whereas multiple comparative studies report that similarities among OR repertoires reflect shared ecology and anatomy rather than phylogenetic relatedness (Hayden et al. 2010; Garrett and Steiper 2014; Hayden et al. 2014; Khan et al. 2015). Consistent with the prominent roles of ecology and anatomy, the size of the OR repertoire has been previously related to reliance on olfaction. There are ~800 OR genes in the human genome, half of which appear to be pseudogenes, whereas there are more than 1,000 intact OR genes in the mouse genome and ~2,000 intact ORs in the elephant (Glusman et al. 2001; Zhang and Firestein 2002; Niimura et al. 2014). Further, although clear links between particular ORs and specific chemical ligands are largely missing, multiple studies have linked features of the OR repertoires to ecological adaptation and lineage-specific specialization (Steiger et al. 2009; Hayden et al. 2010, 2014; Garrett and Steiper 2014; Niimura et al. 2014; Khan et al. 2015).

Most of the comparative studies of the OR repertoires of tetrapods have focused on mammals because of the greater availability of mammalian genome drafts (Zhang and Firestein 2002; Niimura and Nei 2003, 2007; Hayden et al. 2010, 2014; Matsui et al. 2010; Niimura et al. 2014), with a recent study comparing bird OR repertoires as a notable exception (Khan et al. 2015). Sauropsids are the sister group of mammals, and include Rhynchocephalia (tuatara), Squamates (snakes and lizards), Testudines (turtles and tortoises) and Archosaurs (crocodilians, dinosaurs, and birds), and with the exception of birds, have been largely absent from OR studies. Multiple genomes from representatives from this group have been released recently (Castoe et al. 2013; Wan et

al. 2013; Wang et al. 2013; Green et al. 2014) and offer an opportunity to explore the evolution of OR repertoires in amniote vertebrate lineages other than mammals. Therefore, the primary goal of this study was to investigate patterns of diversification of sauropsid OR repertoires using these recently released genomes.

Prior studies based on the genomic analyses of the green anole, chicken, and zebra finch suggest that squamates have smaller OR repertoires than most mammals (Steiger et al. 2009) and that gene loss played a prominent role in the evolution of avian OR repertoires (Khan et al. 2015). Similarly, the OR repertoires of birds appear to be small relative to most other amniotes yet include an expansion of OR subfamily 14 (Lagerström et al. 2006; Steiger et al. 2008, 2009; Khan et al. 2015). The phyletic extent of this expansion has not, however, been fully resolved. Further, it is not known whether snakes, which rely heavily on their sense of smell and chemoreception abilities (Cooper Jr 1991; Stone and Holtzman 1996; Shine and Mason 2001; LeMaster and Mason 2002; Clark 2007), do indeed have a reduced OR repertoire like that observed in the green anole. Similarly, the OR repertoires in crocodiles and turtles, which invaded semi-aquatic niches independently, have yet to be thoroughly analyzed and compared. Because Class I ORs are thought to be primarily involved in detecting aquatic-borne odorants and are particularly abundant in turtles (Wang et al. 2013), I was interested in evaluating whether semi-aquatic crocodilians may have also experienced an expansion of Class I ORs. To address these questions I analyzed patterns of OR gene gain and loss from a sample of sequenced sauropsid genomes. These results indicate that different sauropsid lineages have diverse OR repertories that range from few to several hundred genes derived from

lineage-specific combinations of expansions, losses, and differential retention of ancestral genes.

Material and Methods

Data sources

We queried the genomes for putative ORs from the following representative sauropsid species: green anole (*Anolis carolinensis*), Burmese python (*Python morulus*), Chinese softshell turtle (*Pelodiscus sinensis*), painted turtle (*Chrysemys picta*), American alligator (*Alligator mississippiensis*), Indian gharial (*Gavialis gangeticus*), saltwater crocodile (*Crocodylus porosus*), chicken (*Gallus gallus*), and zebra finch (*Taeniopygia guttata*). I included duckbilled platypus (*Ornithorhynchus anatinus*) as an outgroup. Although many of these genome drafts have been previously surveyed for ORs, I re-annotated these genomes to benchmark the accuracy of my OR prediction approach, and to provide a consistent basis for the annotation of ORs across genomes for comparative analyses. Further, in many of these cases only OR numbers were reported, therefore I sought to provide more detail regarding subfamily designations and comparative evolutionary histories among OR repertoires which has yet to be conducted.

OR prediction

To identify putative ORs I implemented a bioinformatic pipeline similar to the one described in Niimura and Nei (2007). Briefly, I conducted TBLASTN searches of the specified genomes excluding hits with an e-value greater than 1e-10. These searches were conducted using as queries a set of known ORs from the green anole, African clawed frog (*Xenopus tropicalis*), chicken, and zebra fish (*Danio rerio*) from Niimura (2009), and

human ORs from Niimura and Nei (2003). Hits shorter than 150 bp were discarded. I extracted the best BLAST hits identified by the smallest e-value from non-overlapping regions, plus 999 bp in the upstream and downstream flanking sequences, using modules in BEDTOOLS (Quinlan and Hall 2010) and custom Python scripts. Putative OR genes were considered intact if there was an uninterrupted open reading frame with no gaps ≥ 5 amino acids in the 7 transmembrane domains or conserved regions, and an appropriate stop codon. Newly discovered intact ORs were added to the amino acid query and the TBLASTN search was conducted a second time to discover potentially undetected pseudogenes and truncated genes using a cutoff of 1e-20. The best hits, plus 99 bp upstream and downstream, were extracted. ORs were considered pseudogenes if the longest open reading frame (ORF) was shorter than 250 amino acids, there were gaps of five or more amino acids in the transmembrane domains or conserved regions, frame-shift mutations, or premature stop codons. OR sequences located at the end of a scaffold or interrupted by scaffold gaps, but otherwise apparently intact, were considered truncated. Truncated ORs were validated by alignment to functional genes using MAFFT 7.127 (Katoh and Toh 2008) and visually inspected for premature stop codons and gaps within conserved regions. Predicted OR amino acid sequences were mapped back to their corresponding genome to annotate their precise coordinates and orientation.

Class I and II ORs diverged and diversified early in tetrapod evolution (Niimura 2009). Mammalian OR genes have been historically classified into 18 subfamilies, four Class I subfamilies (51, 52, 55, 56) and 14 Class II subfamilies identified from the human genome (Glusman et al. 2000). However, Hayden et al. (2010) determined that several of the previously classified Class II subfamilies were not monophyletic among all mammals

and subsequently defined new groups by identifying monophyletic lineages of ORs (1/3/7, 2/13, 4, 5/8/9, 11, 6, 10, 12, 14). I used BLASTP to group intact ORs into putative subfamilies based on human ORs and the classifications of Hayden et al. (2010). I then verified and corrected the putative BLASTP-based assignments based on the inferred phylogenetic tree of the full OR dataset (see below). I assigned pseudogenes to OR subfamilies in the following manner. I created a database of all of the annotated amino acid sequences and used BLASTX to query the pseudogene nucleotide sequences against the protein database. I used a cutoff of 1e-10 and allowed 10 target sequences per query sequence. The subfamily annotation that was most frequent among the 10 hits was assigned to the pseudogene.

Analyses

After annotation, I used CAFÉ (De Bie et al. 2006) to reconstruct the OR repertoires from the number of intact Class I and Class II genes to identify ancestral OR gene copy number states given the gene gain and loss in each lineage. The CAFÉ method assumes equal probability of birth (duplication) and death (deletion / pseudogenization). Divergence times for each node in the CAFÉ analyses were taken from TimeTree (Hedges et al. 2006).

We estimated the evolutionary relationships of OR sequences based on amino acid alignments. In all cases, I aligned the amino acid sequences of intact ORs using EINSI parameters in MAFFT 7.127. I created a full alignment of all intact ORs and also separate alignments of OR sequences for the birds, crocodilians, turtles, and squamates, and estimated phylogenetic relationships using Fastree2 (Price et al. 2010), which is specifically designed to calculate “approximately-maximum-likelihood” phylogenetic

trees on extremely large alignments such as those generated from aligning thousands of ORs here. Nodal support was estimated from 1,000 bootstrap replicates. The resulting tree was used to infer and date gene duplication events based on a phylogeny-aware algorithm (Huerta-Cepas and Gabaldón 2011) as implemented in ETE v2 (Huerta-Cepas et al. 2010). This method is complementary to CAFÉ, which does not consider the topology of the gene tree.

In most vertebrates studied to date, OR genes are spatially clustered (Giglio et al. 2001; Niimura and Nei 2005a). Thus, it was of interest to investigate how ORs were organized and distributed across various sauropsid genomes. To do so, I analyzed spatial clustering patterns of genetically linked OR genes using BEDTOOLS to locate genomic clusters of ORs in each genome in my analysis, even though establishing the exact boundaries of OR clusters was difficult for most genome drafts. OR clusters can be several Mb long yet many of the unmapped scaffolds containing ORs were shorter than 1 Mb due to the overall shorter scaffold sizes of some genome assemblies I analyzed. Due to this limitation, I defined clusters as three or more OR genes that are separated by less than 100 kb of one another. Clusters that were within 10 kb of a scaffold end were considered incomplete.

Results and Discussion

I first compared results from my bioinformatic pipeline on updated drafts of the green anole, zebra finch, and chicken with the original reports. I found that gene counts were similar between anoCar1 and anoCar2, that galGal4 had more ORs than galGal3, and that these counts were very similar to those in the zebra finch and softshell turtle reported in Wang et al. (2013). The annotation of the python genome yielded more ORs

than previous estimates (Dehara et al. 2012; Castoe et al. 2013). Overall, these comparisons suggest that my pipeline generates results that are generally comparable to those from previous studies, and in some cases more inclusive. Thus, I inferred the characterization of the OR repertoires of painted turtle, python, gharial, American alligator, and saltwater crocodile represent robust estimates of the diversity and size of the OR gene family in these genomes.

OR repertoires vary among major sauropsid groups

Quantitative comparisons of ORs across genomes indicate that sauropsids evolved extensive variation in the size of the OR repertoires, as the number of intact genes in the genomes analyzed ranged over an order of magnitude, from 108 in the green anole to 1180 in the Chinese softshell turtle. Similarly, the number of pseudogenes ranged from 33 in the green anole to 538 in the American alligator (Table 1) and the number of truncated but putatively coding genes ranged from 1 in the green anole to 598 in the python. The abundance of these truncated genes did not appear to be related to the overall contiguity of genome assembly, since the crocodile and gharial genomes had shorter scaffold N50s yet fewer truncated genes (Table 1).

Intriguingly, the two squamates in this study, the python and the green anole, diverged ~160 million years ago (Evans 2003; Castoe et al. 2009) and exhibit the largest difference in the number of ORs between species within a major sauropsid lineage, 481 in the python to 108 in the anole (Table 1). This difference is probably higher, as the number of truncated genes in the python genome (Table 1, fig. 1A) suggests the number of intact genes in the python genome is likely higher than my current estimate. Despite

these numerical differences, both species have repertoires dominated by Class II ORs (fig. 1A) with similar subfamily proportions in the two species (fig. 1B).

The two testudines in this study, the Chinese softshell and the painted turtle, diverged ~ 170 mya (Pyron 2010). The turtle genomes contained the largest numbers of intact ORs among sauropsids, and included several hundred Class I genes (fig. 1A), primarily from subfamily 52 (fig. 1B). This class of ORs is thought to mediate detection of water-borne odorants (Saito et al. 2009). Compared to the Chinese softshell turtle, the painted turtle genome contained a higher fraction of truncated ORs (24% vs 3% in the softshell turtle, fig. 1A) and pseudogenes (~50% vs. ~30% in the softshell turtle).

The two extant groups of archosaurs, birds and crocodilians, show marked differences in their OR repertoires. Chicken and zebra finch had the second smallest number of ORs, with 200 and 250 ORs respectively, almost all of which belonged to subfamily 14 (fig. 1B). By contrast, crocodilian genomes encode more than twice the number of intact ORs, between 465 and 597 (Table 1), derived from multiple subfamilies (fig. 1B). It is notable that although the three crocodilian species diverged ~90 MYA (Roos et al. 2007), they have similar OR repertoires in terms of gene numbers (fig. 1A) and subfamily composition (fig. 1B), further illustrating suggestions that crocodilian genomes have remained remarkably static and conserved over many millions of years (Green et al. 2014).

OR pseudogenization

If there has been no gene gain and pseudogenes are retained in the genome, there should be a negative correlation between the number of intact ORs and number of pseudogenes. To test this prediction, I explored the number of pseudogenes and their

distribution across OR subfamilies. I calculated the proportion of pseudogenes by dividing the number of pseudogenes by the total number of genes, excluding truncated genes because they cannot be classified confidently. The analyses indicate that the overall proportion of pseudogenes was not correlated with the number of intact genes (fig. 2A). However, I did find a significant positive correlation between the proportion of pseudogenes per subfamily and the proportion of intact genes per subfamily ($r^2 = 0.87$, $p < 0.0001$, fig. 2B). These two observations together suggest that the pseudogenes present reflect the composition of the OR repertoire, but that the current abundance of pseudogenes is not determined by the abundance of intact genes. The fraction of pseudogenes can change when genes or pseudogenes are deleted from the genome (Niimura et al. 2014) and although several groups have suggested that the proportion of pseudogenes relative to the total number of genes is related to olfactory ability (Kishida et al. 2007; Hayden et al. 2010; Kishida and Hikida 2010), the results here are not consistent with this. In agreement with Niimura et al. (2014), my analyses suggest that the fraction of pseudogenes is a poor indicator of olfactory ability.

Genomic organization of OR genes

In most mammalian genomes, OR genes are arranged in gene clusters composed of closely related genes and orthologous clusters are often shared among relatively distantly related species, such as the human and the mouse (Niimura and Nei 2005a). Similarly in sauropsids, the proportion of ORs in clusters ranged from 42 to 90%, in the zebra finch and softshell turtle respectively, and the number of OR gene clusters per genome ranged from 5 to 139 across species (Table 2). Some of these clusters were composed of a single subfamily, but most clusters (such as the largest cluster in the

painted turtle) contained multiple OR subfamilies (fig. 3A). As expected, genome drafts with lower scaffold N50s exhibited smaller clusters and greater abundances of incomplete clusters (Table 2). True OR cluster sizes are likely larger than my estimates (due to the fragmentary nature of assemblies), and ultimately the majority of ORs may be located in a small number of clusters, as in the green anole where almost 85% of ORs were found in only five clusters (Table 2). For example, five scaffolds contained the majority of subfamily 51 ORs in the softshell turtle (fig. 3B), and the entire length of these five contigs is composed almost exclusively of these ORs (fig. 3C). The combined length of these five contigs is approximately 1.5 Mb. Because almost all of the subfamily 51 ORs are scattered among these five contigs and each of these contigs is incomplete at the 5' and 3' end, I suspect these five contigs may represent a single contiguous cluster, containing the majority of the subfamily 51 ORs. The largest single human OR locus contains ~130 ORs and the largest mouse locus contains ~250 ORs (Niimura and Nei 2005a). By comparison, the largest clusters in my study were observed in the anole, platypus, painted turtle and softshell turtle (assuming a single contiguous cluster) and contained 75, 93, 108, and 147 total ORs, respectively. These comparisons should be considered preliminary, as the actual cluster sizes will likely change as future more contiguous genome assemblies become available. Regardless of the exact numbers of clusters per genome, observed patterns of OR clustering across genomes suggest that tandem duplication is a primary source of novel ORs, as suggested by Niimura and Nei (2005a).

Evolution of OR repertoires

We tracked gene gain and loss for Class I and Class II ORs on the species tree using maximum likelihood estimates of OR repertoire size (fig. 1C). These analyses suggest that the relatively small OR repertoires of the chicken, the zebra finch and the anole are the result of multiple gene losses, and that ancestors of both the birds and the green anole likely contained larger OR repertoires. Interestingly, the analysis of the anole and the two birds yield drastically different patterns of gene loss. The anole lost ORs from all subfamilies while retaining repertoire diversity, whereas the two birds analyzed lost almost all ORs belonging to all subfamilies except subfamily 14. Khan et al. (2015) previously demonstrated that birds generally have diverse OR repertoires. The notable exceptions were chicken, zebra finch and the little egret, as more than 90% of the ORs in these genomes were made of subfamily 14 ORs suggesting that almost all OR subfamilies were lost independently in these three species.

Reconstructions of the OR repertoire in the common ancestor of sauropsids suggest it had 51 Class I and 344 Class II ORs. Ancestral nodes also had hundreds of Class II receptors, ranging from 263 to 520, and tens of Class I receptors, from 14 to 58, with the exception of the common ancestor of softshell and painted turtle, which had an estimated 274 Class I receptors. Turtles are notable because they are the only group analyzed to have gained Class I ORs at a greater rate than Class II ORs (fig. 1C). The crocodilian ancestor is estimated to have gained approximately 100 ORs since diverging from birds, and interestingly, the number of ORs in the three crocodilians has apparently remained remarkably similar to the number inferred for their ancestor.

To investigate patterns of OR gene gain and loss among subfamilies in more detail, I estimated phylogenetic relationships among all 4,991 intact OR genes (fig. 4). Class I and Class II ORs formed highly supported monophyletic clades (fig. 4A). Most mammalian-defined subfamilies generally formed monophyletic groups that included mammalian and sauropsid representatives (fig. 4A). Exceptions to this pattern include subfamily 5/8/9, which is paraphyletic in my analysis. The 5/8/9 subfamily is represented by three relatively distant clades (fig. 4A) that each includes representatives of all sauropsids (fig. 1B, 4B). Because of the difficulties in resolving a tree with many more sequences than sites in the alignment, I restricted the primary focus to strongly supported monophyletic OR subfamilies, such as groups 51, 52, and 14 (fig. 4A).

In most cases, it was uncommon for ORs from a species or lineage to form a monophyletic group within a subfamily suggesting that these subfamilies had expanded prior to radiation of these species. Subfamily 14 was an exception, as almost all ORs in this subfamily formed species or lineage-specific clades (fig. 4B) suggesting that the same ancestral gene expanded independently multiple times in different lineages. Within this subfamily, chicken and zebra finch ORs are the most remarkable, as they are reciprocally monophyletic, stemming exclusively from species specific expansions (fig. 5C; Khan et al. 2015), with a long branch leading to their common ancestor gene (fig. 4B).

When I constructed independent phylogenetic trees for each major sauropsid group, I was able to visualize the distinct patterns of gene gain and loss that produced current OR repertoires (fig. 5). Phylogenetic tree characteristics tended to be fundamentally different among the four groups. Turtles and crocodilian genomes are both

notable for evolving slowly (Shaffer et al. 2013; Green et al. 2014), yet the OR repertoires of turtles show extensive evolutionary dynamics with multiple species-specific expansions (fig. 5B) while crocodilian OR repertoires have apparently experienced little change in gene number and diversity (fig. 5D). The conservative nature of crocodilian OR repertoires is exemplified by noticeably short terminal branches among orthologous gene copies (fig. 5D). These observations collectively suggest that not only have crocodilians not experienced substantial change in the number and diversity of OR genes, but have also experienced fewer amino acid changes among orthologous OR genes since extant crocodilians diverged from their common ancestor.

ORs subfamilies in the last common ancestor of sauropsids

Consistent with previous analyses of OR repertoires in birds (Khan et al. 2015), these phylogenies indicate that at least six OR subfamilies (51, 52, 14, 4, 12, and 1/3/7) formed monophyletic groups among sauropsids, suggesting that these subfamilies began diversifying in the common ancestor of sauropsids. In addition, the majority of the predicted ORs in subfamilies 11, 10, 6, and 2/13 were placed in monophyletic groups as well, indicating that these subfamilies were also present in the last common ancestor of sauropsids. The most interesting case is subfamily 5/8/9, which is split into three weakly supported clades in this study (fig. 4A). Thus, my analyses suggest that the OR classification derived from mammals is largely applicable to sauropsids, and that the 11 major groups emerged prior to divergence of these mammals and sauropsids.

The Role of Natural Selection in Shaping OR repertoires

The relative role selection played in shaping OR repertoires is a matter of debate. Early studies suggest that variation in OR repertoires is largely independent of selection (Niimura and Nei 2007; Nei et al. 2008). However, more recent comparative studies among mammals (Hayden et al. 2010, 2014) and birds (Khan et al. 2105) suggest that OR repertoires reflect ecological adaptations and have in part been shaped by natural selection. Khan et al. (2015) also show that olfactory acuity, as reflected by the size of the olfactory bulb, is correlated with the size of the olfactory repertoire.

These results provide new and intriguing evidence consistent with a role of natural selection in shaping OR repertoires. I found independent expansions of subfamilies associated with detection of waterborne odorants in the two aquatic groups studied: subfamily 2/13 expanded in crocodiles, which has been linked to chemoreception in aquatic mammals and birds (Hayden et al. 2010; Khan et al. 2015), and Class I ORs in turtles, which are hypothesized to primarily bind waterborne odorants (Saito et al. 2009; Wang et al. 2013). Additional support for natural selection was observed in comparisons between squamate reptile OR repertoires. The green anole is an arboreal insectivore that relies on visual cues for social interactions (Leal and Fleishman 2004) and has the lowest number of functional ORs, despite having high OR subfamily diversity. By contrast, the python, which like most snakes, has poor hearing and vision and relies heavily on chemoreception to locate prey and mates, has at least five times as many putatively functional OR genes as the anole. Thus, the difference in size of squamate OR repertoires points towards a correspondence between OR size and the relative dependence on chemosensory information. While not conclusive, examples from sauropsid OR

repertoires are at least consistent with natural selection playing a role in shaping OR repertoires, and suggests that the diversity of OR repertoires and natural history of sauropsid species may provide a rich model system for more detailed tests of this hypothesis.

Conclusions

Sauropsids represent an ecologically and phenotypically diverse set of tetrapods that include the closest living relatives to mammals, and recently available genomes of representative members of sauropsid lineages provide new opportunities to study the patterns of OR diversification in the group. These results indicate that most sauropsids have diverse and relatively large OR repertoires that derive from a complex diversity of lineage-specific patterns of gene birth and death, and the differential retention of OR duplicates. I found that gene loss has played a prominent role in the evolution of the repertoires of birds and lizards. In contrast, turtles have experienced notable gains of class I ORs, and the common ancestor of crocodilians gained multiple ORs. Unlike other lineages, however, the crocodilian repertoire has remained nearly constant since the diversification of crocodilian lineages. Overall sauropsids have undergone numerous major life history and ecological transitions that are likely to have resulted in changes in the dependence of various lineages on olfaction and on OR repertoires.

Acknowledgements

This work was supported by the National Science Foundation [DEB-1355176 (FGH and DAR), MCB-0841821 (DAR), and MCB-1052500 (DAR) as well as funding

from the College of Agriculture and Life Sciences and the Institute for Genomics, Biocomputing and Biotechnology at Mississippi State University.

Table 3.1 Summary of OR gene annotations from each genome.

Genome	Intact (I)	Pseudogenes (P)	Truncated (T)	Total (I+P+T)	%Truncated (T/I+T)	%Intact (I/I+P)
Platypus	270	351	35	656	11	44
Green Anole	108	33	1	142	0.9	77
Python	481	319	598	1398	55	60
Softshell Turtle	1180	533	40	1753	3	69
Painted Turtle	842	942	279	2063	24	47
Crocodile	592	331	66	989	10	64
Gharial	597	389	153	1139	18	61
Alligator	465	538	74	1077	14	46
Zebra Finch	190	306	45	541	19	38
Chicken	266	173	83	522	24	61

For each species I calculated the total number of intact, pseudogenes and truncated genes.

Table 3.2 Summary of OR gene clusters.

Genome	Clusters	Intact	5' incomplete	3' incomplete	5' and 3' incomplete	% genes in clusters
Platypus	39	11	23	1	4	41
Green Anole	5	5	0	0	0	83
Python	130	16	84	30	0	58
Softshell Turtle	126	30	83	3	10	90
Painted Turtle	115	53	51	6	5	78
Crocodile	122	30	79	2	11	69
Gharial	139	48	67	12	12	58
Alligator	120	0	58	3	15	75
Zebra Finch	52	52	0	0	0	42
Chicken	27	7	18	0	2	57

From each genome draft, I estimated the number of OR clusters, the number that clusters that were not near a contig end (intact), those that were near the 5', 3' or both contig ends, and the percentage of genes found in clusters.

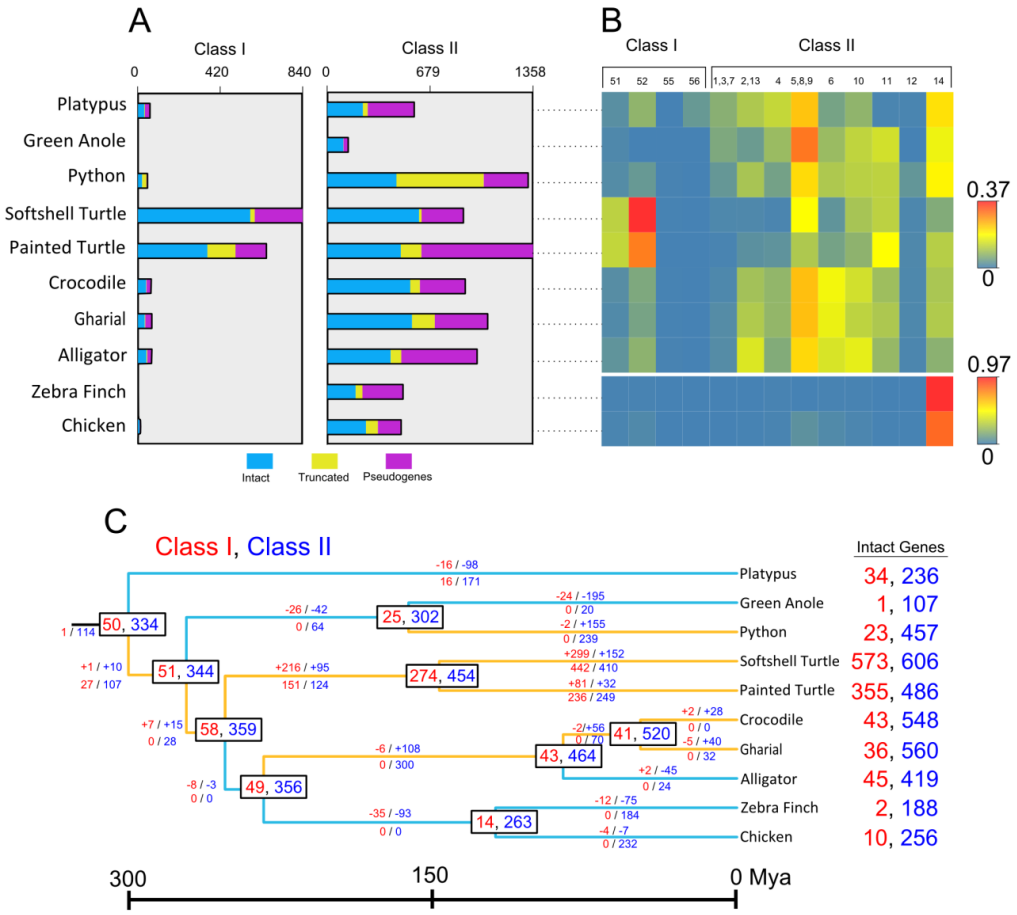

Figure 3.1 OR composition among taxa

A) The comprehensive collection of annotated Class I and Class II genes in each taxon. B) Heat map based on the proportion of intact ORs that belong to OR subfamilies. Avian and non-avian groups were presented on two different scales because more than 85% of avian ORs are in subfamily 14, whereas the highest percentage is 36% in subfamily 52 of turtles. C) Historic Class I and Class II gene numbers in the ancestral nodes and gain/loss along each branch of taxa (CAFÉ analysis, above branches), and the inferred number of past duplication events per OR Class and lineage, based on the gene phylogeny and a species-overlap duplication detection and dating algorithm (Huerta-Cepas and Gabaldón (2011), below branches). Light blue branches are those with an average gene loss per Class and orange branches are those with an average gene gain.

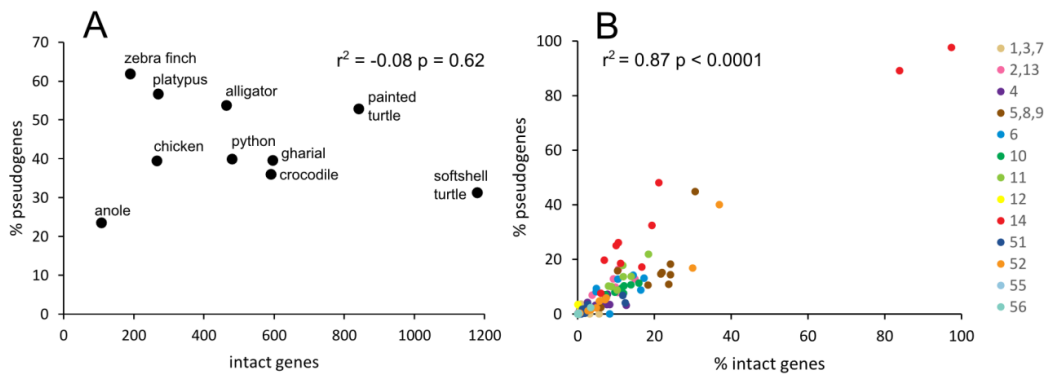

Figure 3.2 Pseudogene frequency

A) The number of intact genes plotted against the percentage of pseudogenes as a proportion of the total number of intact genes and pseudogenes (p/p+i) within the same genome. A simple linear model was applied to the data and there was no significant correlation. B) I plotted the percentage of pseudogenes against the percentage of intact genes for all subfamilies in all species and again applied a simple linear model to the data and found a strong linear relationship between the two metrics.

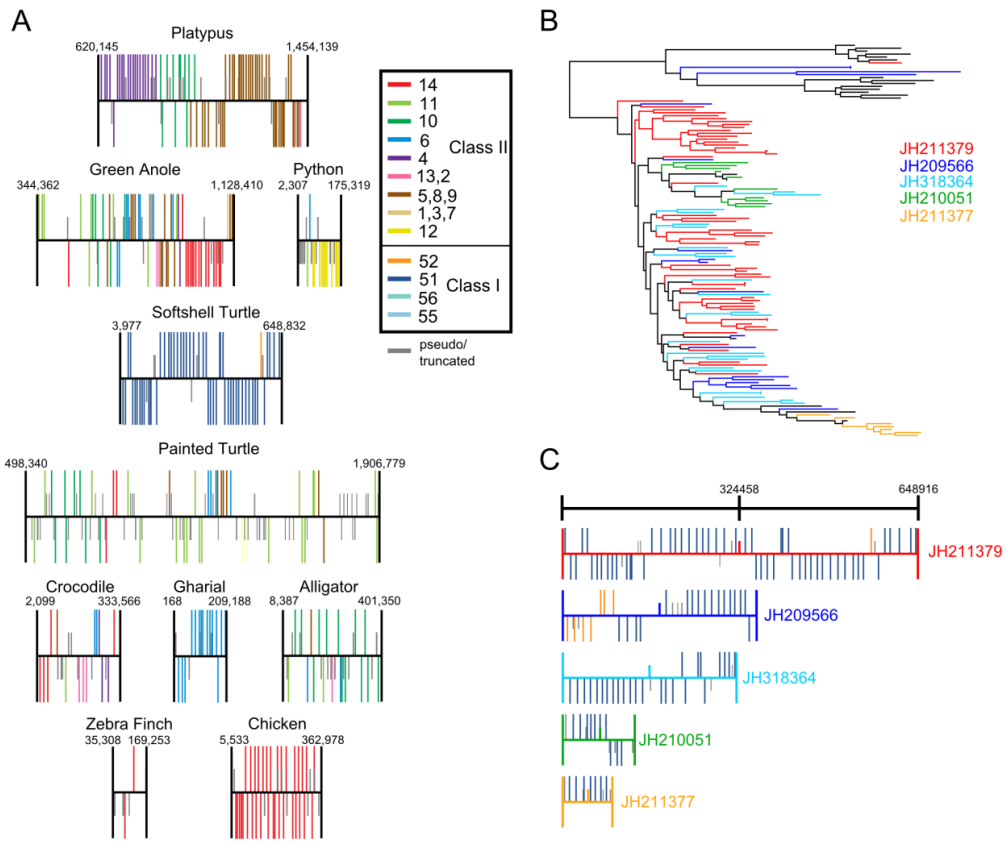

Figure 3.3 OR cluster composition

A) The largest (most numerous) OR gene cluster from each genome draft. Each vertical bar represents a position of an OR. Bars above the horizontal line represent sense oriented genes and bars below the line represent anti-sense oriented genes in relation to the scaffold sequences. Each OR is colored according to the annotated subfamily. Cluster lengths are drawn to scale. B) Neighbor joining tree of the subfamily 51 ORs in the softshell turtle; branches are colored according to the contig each OR was identified on. C). The OR content of each contig presented in the B panel. Contig lengths are to scale, and the gene color scheme is congruent with the legend in panel A.

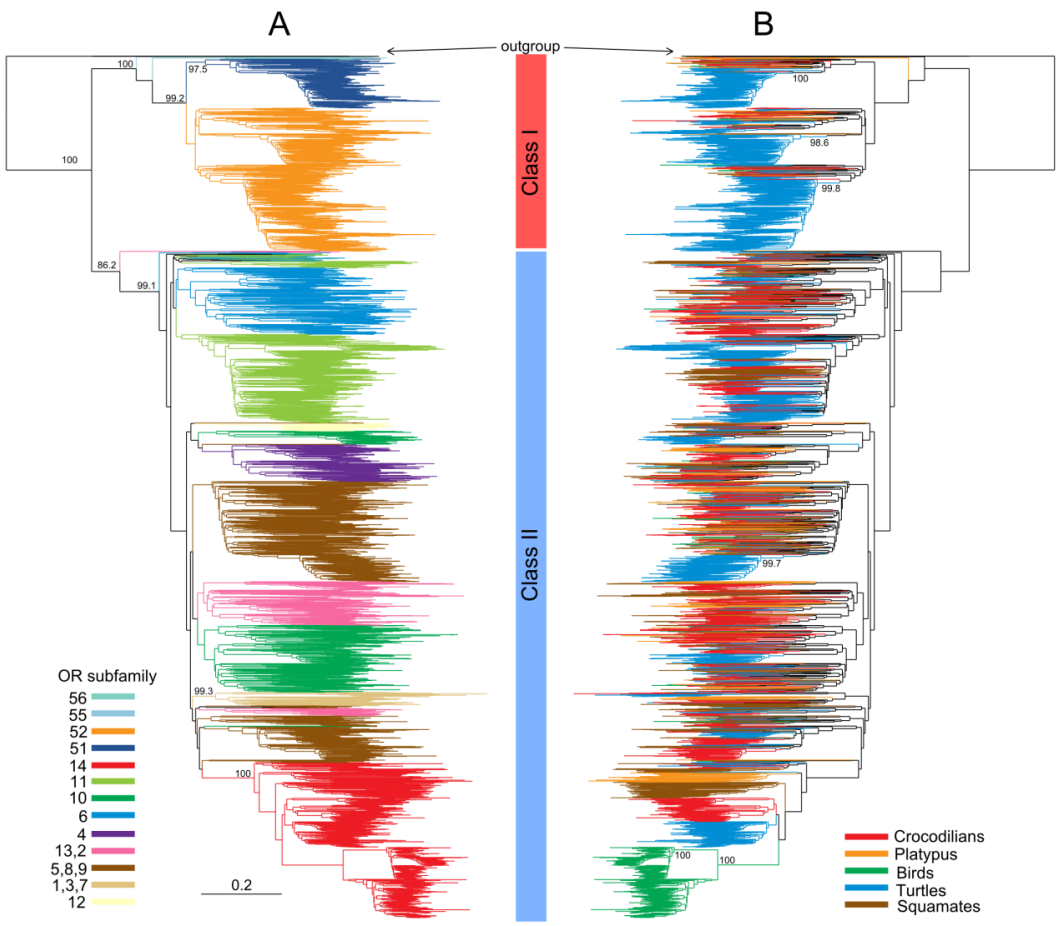

Figure 3.4 Phylogenetic tree estimate of the 4,991 intact ORs

Phylogenetic tree estimate of the 4,991 intact ORs. A) Branches colors are based on the annotated OR subfamily, nodal support is listed for the Classes and high supported subfamilies. B) The same tree presented in A but branches are colored according to the major taxonomic classifications and nodal support is presented for high supported group specific OR expansions.

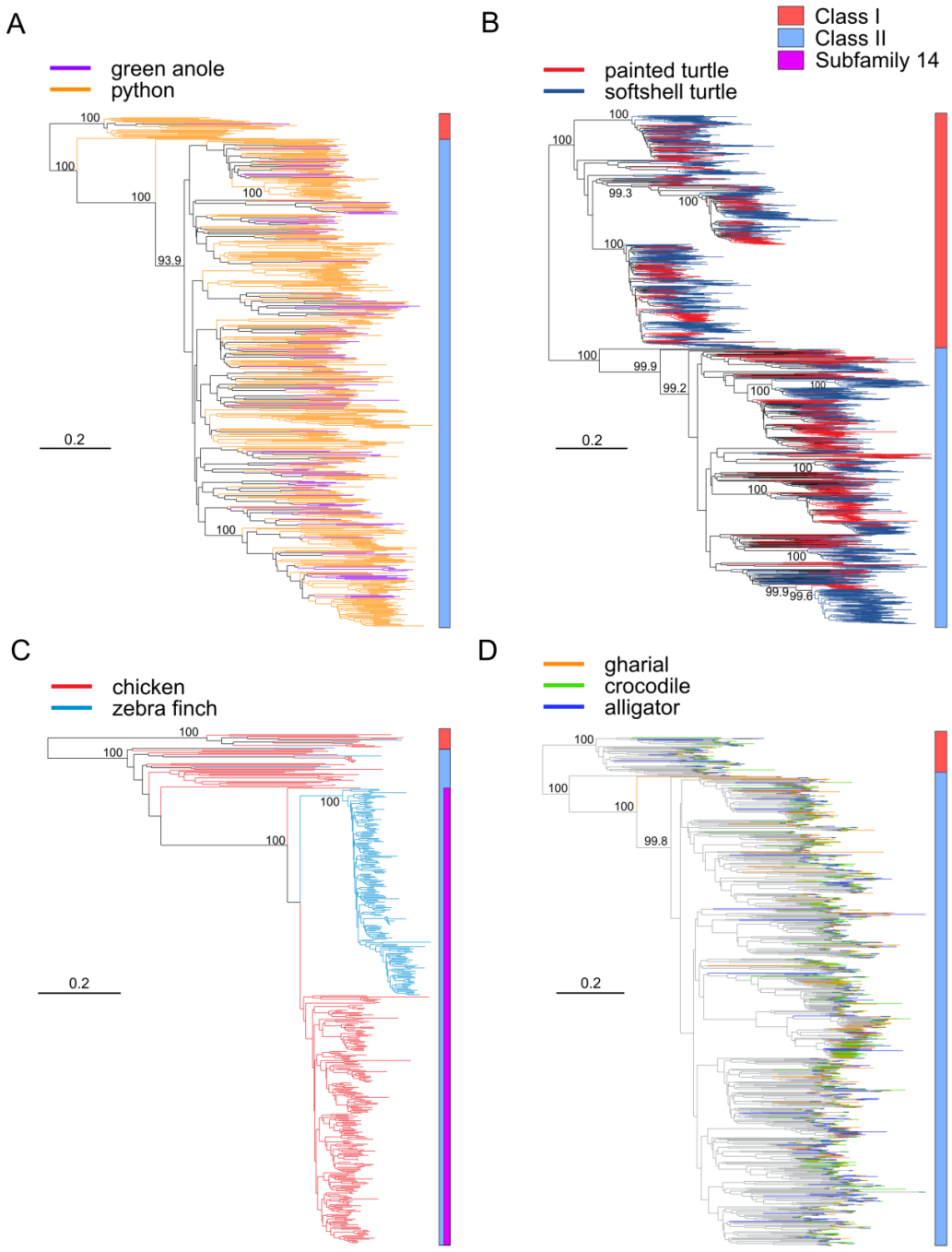

Figure 3.5 OR gene trees among major sauropsid groups

Phylogenetic reconstruction of ORs in each major group of saurposids, including A) squamates B) turtles, C) birds, and D) crocodilians.

References

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–187.

Castoe TA et al. 2009. Dynamic nucleotide mutation gradients and control region usage in squamate reptile mitochondrial genomes. Cytogenet. Genome Res. 127:112–127.

Castoe TA et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc Natl Acad Sci. 110:20645–20650.

Clark RW. 2007. Public information for solitary foragers: timber rattlesnakes use conspecific chemical cues to select ambush sites. Behav. Ecol. 18:487–490.

Cooper Jr WE. 1991. Discrimination of integumentary prey chemicals and strike-induced chemosensory searching in the ball python, Python regius. J. Ethol. 9:9–23.

Dehara Y et al. 2012. Characterization of squamate olfactory receptor genes and their transcripts by the high-throughput sequencing approach. Genome Biol Evol. 4:602–616.

Evans SE. 2003. At the feet of the dinosaurs: the early history and radiation of lizards. Biol Rev. 78:513–551.

Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol. 63:1256–1272.

Garrett EC, Steiper ME. 2014. Strong links between genomic and anatomical diversity in both mammalian olfactory chemosensory systems. Proc R Soc B Biol. 281:20132828.

Giglio S et al. 2001. Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet. 68:874–883.

Glusman G et al. 2000. The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm Genome. 11:1016–1023.

Glusman G, Yanai I, Rubin I, Lancet D. 2001. The complete human olfactory subgenome. Genome Res. 11:685–702.

Green RE et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346:1254449.

Hayden S et al. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. 20:1–9.

Hayden S et al. 2014. A cluster of olfactory receptor genes linked to frugivory in bats. Mol Biol Evol. 31:917–927.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:24.

Huerta-Cepas J, Gabaldón T. 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics 27:28–45.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9:286–298.

Khan I et al. 2015. Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. Mol Biol Evol. 32:2832–2843.

Kishida T, Kubota S, Shirayama Y, Fukami H. 2007. The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: evidence for reduction of the functional proportions in cetaceans. Biol Lett. 3:428–430.

Kishida T, Hikida T. 2010. Degeneration patterns of the olfactory receptor genes in sea snakes. J Evol Biol. 23:302–310.

Lagerström MC et al. 2006. The G protein–coupled receptor subset of the chicken genome. PLoS Comput Biol. 2:e54.

Leal M, Fleishman LJ. 2004. Differences in visual signal design and detectability between allopatric populations of Anolis lizards. Am Nat. 163:26–39.

LeMaster MP, Mason RT. 2002. Variation in a female sexual attractiveness pheromone controls male mate choice in garter snakes. J Chem Ecol. 28:1269–1285.

Mombaerts P. 1999. Seven-transmembrane proteins as odorant and chemosensory receptors. Science 286:707–711.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet. 9:951–963.

Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol. 1:34–44.

Niimura Y, Matsui A, Touhara K. 2014. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res. 24:1485–1496.

Niimura Y, Nei M. 2005a. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. Gene 346:13–21.

Niimura Y, Nei M. 2003. Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci. 100:12235–12240.

Niimura Y, Nei M. 2005b. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc Natl Acad Sci. 102:6039–6044.

Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2:e708.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Pyron RA. 2010. A likelihood method for assessing molecular divergence time estimates and the placement of fossil calibrations. Syst Biol. 59:185–94.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Roos J, Aggarwal RK, Janke A. 2007. Extended mitogenomic phylogenetic analyses yield new insight into crocodylian evolution and their survival of the Cretaceous–Tertiary boundary. Mol Phylogenet Evol. 45:663–673.

Saito H, Chi Q, Zhuang H, Matsunami H, Mainland JD. 2009. Odor coding by a mammalian receptor repertoire. Sci Signal. 2:ra9.

Shaffer HB et al. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14:R28.

Shine R, Mason RT. 2001. Courting male garter snakes (Thamnophis sirtalis parietalis) use multiple cues to identify potential mates. Behav Ecol Sociobiol. 49:465–473.

Steiger SS, Fidler AE, Valcu M, Kempenaers B. 2008. Avian olfactory receptor gene repertoires: evidence for a well-developed sense of smell in birds? Proc R Soc B Biol. Sci. 275:2309–2317.

Steiger SS, Kuryshev VY, Stensmyr MC, Kempenaers B, Mueller JC. 2009. A comparison of reptilian and avian olfactory receptor gene repertoires: Species-specific expansion of group γ genes in birds. BMC Genomics 10:446.

Stone A, Holtzman DA. 1996. Feeding responses in young boa constrictors are mediated by the vomeronasal system. Anim Behav. 52:949–955.

Wan Q-H et al. 2013. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. Cell Res. 23:1091–1105.

Wang Z et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet. 45:701–706.

Zhang G et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346:1311–1320.

Zhang X, Firestein S. 2002. The olfactory receptor gene superfamily of the mouse. Nat Neurosci. 5:124–133.

CHAPTER IV

TRANSPOSABLE ELEMENT TARGETING BY PIRNA IN LAURASIATHERIANS WITH DISTINCT TRANSPOSABLE ELEMENT HISTORIES

Introduction

Transposable elements (TEs) are selfish DNA sequences that have the ability to invade and propagate in host genomes, and are classified as either DNA transposons or retrotransposons based on their mechanism of mobilization and cycle of replication. Retrotransposons (Class I TEs) mobilize exclusively through 'copy-and-paste' mechanisms, they transcribe an RNA intermediate that is reverse transcribed and inserted into a new genomic location. In mammalian genomes the most common retrotransposons are Long INterspersed Elements (LINEs), Short INterspersed Elements (SINEs), and Long Terminal Repeat elements (LTRs). DNA transposons (Class II TEs) do not use an RNA intermediate and may mobilize either through a 'cut-and-paste' mechanism (TIR elements), by excising themselves from one locus and reinserting into a novel one (Kapitonov and Jurka 2007), or by “copy-and-paste” mechanisms (e.g. Helitrons and Mavericks).

TEs are major components of vertebrate genomes, and in the case of mammals, TEs can account for up to 70% of the genomic content (de Koning et al. 2011), most of which is derived from retrotransposon insertions (Yohn et al. 2005). TEs can be an important source of variation within and among species. In addition to increasing genome

size, TE insertions can disrupt gene reading frames or alter gene expression by inserting within or close to a gene, promote genomic deletions and reorganize genome structure via TE-mediated non-homologous recombination (Gilbert et al. 2002; Liu et al. 2003; Callinan et al. 2005; Han et al. 2005; Sen et al. 2006). Because of these potential impacts, TE mobilization is generally considered deleterious and their unrestricted proliferation can have profound biological effects. As a result, the question of how organisms control TE mobilization has attracted great interest.

Data from multiple metazoans indicate that proteins in the PIWI and Argonaute gene families, referred to as PIWI proteins from here onwards, and PIWI-interacting RNAs (piRNAs), a class of small noncoding RNAs predominantly expressed in germlines, play a major role reducing TE expression and mobilization (Aravin, Hannon, et al. 2007; Aravin, Sachidanandam, et al. 2007; Brennecke et al. 2007; O'Donnell and Boeke 2007; Saito and Siomi 2010). piRNAs are the most abundant class of small RNAs expressed in testis and range in size from approximately 24 to 32 nucleotides (Aravin et al. 2006; Girard et al. 2006; Aravin et al. 2008; Höck and Meister 2008). piRNA and PIWI proteins associate in complexes that are involved in epigenetic and post-transcriptional repression of TEs (Siomi et al. 2011).

Two distinct populations of piRNAs have been described in mammals, pre-pachytene and pachytene, which differ in their expression, biogenesis, and genomic origins (Aravin, Sachidanandam, et al. 2007; Li et al. 2013). Expression of pre-pachytene piRNAs begins in pre-meiotic and early prophase 1 spermatogonia whereas expression of pachytene piRNAs starts in the pachytene stage of prophase 1 through sperm maturation (Aravin, Sachidanandam, et al. 2007; Aravin et al. 2008; Reuter et al. 2011). Both classes

of piRNAs are present in mature testes; however pachytene piRNAs greatly outnumber pre-pachytene piRNAs (Li et al. 2013). That being said, pachytene piRNAs are derived from unannotated (non-TE) regions of the genome and do not appear to be involved in TE silencing. Instead, evidence suggests pachytene piRNAs regulate and eliminate gene transcripts from the cytoplasm in a manner similar to the miRNA pathway (Guo et al. 2014). By contrast, the pre-pachytene population of piRNAs appears to be heavily involved with post-transcriptional silencing of TEs via a feed-forward amplification loop known as the "ping-pong" cycle. In the mammalian ping-pong cycle, Aravin et al. (2008) proposed that a primary piRNA is derived from a TE transcript. This primary piRNA becomes bound to its PIWI counterpart, directs the complex to a complementary anti-sense TE transcript, and through the slicer activity of the PIWI protein, directs the cleavage of the bound transcript producing a secondary anti-sense piRNA. The secondary piRNA is then loaded onto a new PIWI protein and the cycle is repeated, amplifying the pool of both primary and secondary piRNAs while reducing the threat of TE transcripts. This is in stark contrast to the *Drosophila* ping-pong cycle where anti-sense piRNAs are derived from transcribed piRNA clusters and subsequently bind sense TE transcripts (Brennecke et al. 2007).

The evolutionary relationships between TE families and regulatory piRNAs have been examined in a few model species. For example, Kelleher and Barbash (2013) found limits of the piRNA response and suggested the most deleterious TE families are not the most abundant in the piRNA pool. By contrast, the relationship between piRNAs and TE families in mammals is effectively unknown. Lukic and Chen (2011) and Mouier (2011) found a correlation between the age of TE families and piRNA density in humans and

mice, respectively. However, there has not been a thorough investigation into which TE parameters illicit the strongest piRNA response. Because there is such a contrast between the TE silencing pathway in flies and mammals, I feel a broader investigation was needed.

The goal of this research is to better understand the relationships between TE abundance at the genome and transcriptome level and piRNA abundance among mammals. To do so, I compared genome-wide TE composition, TE expression, and the strength of the piRNA response elicited by TEs in three laurasiatherian mammals with very distinct TE landscapes. The three species diverged from one another within a relatively short period, approximately 80 million years ago (Meredith et al. 2011), and the combination of distinct TE loads and similar evolutionary divergences allowed us to explore the piRNA response to TE related variables within the context of the ping-pong model. Briefly, analyses indicate that TE expression was a strong predictor of the level of piRNA response, in agreement with predictions of the ping-pong model, and suggest that the level of piRNA response may modulate the relative contribution of the different TEs to the genome.

Materials and Methods

Sample collection and library prep

We collected discarded testicular tissue from one adolescent dog and one adolescent horse after the animals were sedated and neutered by licensed veterinarians from the College of Veterinary Medicine at Mississippi State University. A wild caught adult big brown bat, *Eptesicus fuscus,* was sacrificed in accord with IACUC standards to collect testis tissue. In each case, a cross section of testis was snap frozen in liquid

nitrogen immediately following castration and stored at -70°C prior to RNA isolation. I isolated total RNA using Trizol ® (Invitrogen, USA) according to the manufacturer's specifications. Small RNA libraries were prepped using the Illumina TruSeq small RNA kit © and 1 x 50 bp reads were sequenced on the Illumina HiSeq2000 platform. Directional RNASeq libraries were prepped using the Illumina TruSeq v2 kit and 2x100 bp reads were sequenced on a single lane of a HiSeq2000.

TE composition and expression

We masked the dog (CanFam3.1) and horse (EquCab2) genomes using RepeatMasker 4.0.5 using the '-species dog' and '-species horse' parameters, respectively. The big brown bat genome was acquired from NCBI (EptFus1.0, GenBank accession ALEH00000000, 1.806 gigabases). Contigs were first masked with RepeatMasker with the '-species Chiroptera' option then secondarily masked with a *de novo* repeat library constructed from the *Eptesicus* genome draft (Platt et al. 2014). To estimate genetic distances, I used the calcDivergenceFromAlign.pl script included with RepeatMasker to calculate Kimura two-parameter (Kimura 1980) distances between each insertion and its respective consensus sequence. The option –noCpG was invoked to exclude highly mutable CpG sites from distance calculations. I calculated the total number of insertions, total number of bases (as a proportion of the genome), average insertion length, and the median genetic distance among insertions for each TE family from the RepeatMasker output. Novel TEs insertions, especially among retrotransposons, are expected to be identical to the source element, and the consensus sequence of a given subfamily is inferred to be the best estimate of the sequence of the source element for that subfamily. Within the framework of the master element model proposed by Brookfield

and Johnson (2006), the distance between an element of a given subfamily and the corresponding consensus provides an estimate of the age of that insertion, and the median distance among insertions of a given subfamily provide an estimate of the peak of accumulation in that subfamily. Thus, insertions with high similarity to the corresponding consensus, i.e. low genetic distances, are assumed to have occurred in the recent past, whereas insertions with low similarity (high genetic distance) are thought to be older.

We estimated the relative expression of TE families by mapping RNA-Seq reads to the TE consensus sequences representing families found in each genome. For each species, I mapped approximately 30 million RNASeq reads to the consensus elements using the default parameters of RSEM (Li and Dewey 2011), which used Bowtie to initially map reads. The default parameters allow two mismatches in a seed region of the first 25 bases of an alignment, then unlimited mismatches in the remainder of the sequence alignment. Expression estimates were measured in transcripts per million (TPM).

piRNA processing and cluster annotation

Prior to small RNA mapping, I clipped barcodes, removed reads that had bases with Phred quality score <25, and removed identical reads using modules in the fastx toolkit. I also removed low complexity small RNA sequences using a custom python script. I mapped piRNA-like (pilRNA) sequences 24-32 bases long to the complete genomes using Bowtie allowing one mismatch in the alignment (Langmead et al. 2009). pilRNAs that mapped to only one locus were reported. A cluster was defined as a group of at least 50 pilRNAs where contiguous pilRNAs were separated by less than 1,500 bases (Beyret et al. 2012). Only clusters longer than 10Kb and that had a normalized

small RNA count of 10 pilRNAs / cluster length (in thousands) / number of mapped sequences (in millions) were analyzed for TE insertions. I calculated the same TE parameters within clusters as I did for the whole genome (see above).

Ping-pong piRNA expression

pilRNA sequences were mapped to a library of consensus sequences representing the TE families annotated in each genome. I mapped pilRNAs to the consensus elements, allowing three mismatches, and allowed pilRNAs to map to all possible loci. I identified the ping-pong signature by partitioning mapped reads into putative primary or secondary pilRNAs. pilRNAs that had a U in the first position and did not have an A in the 10^{th} position were considered 'primary' pilRNAs, whereas those pilRNAs that had an A in the 10^{th} position and did not have a U in the first position were classified as 'secondary' pilRNAs. Pairs of primary and secondary pilRNAs that overlapped at the first 10 nucleotides were assumed to have resulted from the ping-pong cycle. Ping-pong pilRNA expression was estimated for each TE family by summing the number of ping-pong pilRNAs for each element and dividing the pilRNA counts by the length of the consensus sequence (in thousands of bp) and the number of ping-pong pilRNA that mapped to the entire consensus libraries for each species (in millions). I refer to this metric as ping-pong pilRNA expression (PPE) throughout and consider it as a proxy for the strength of the piRNA response against a given TE because the abundance of ping-pong pairs would indicate where PIWI proteins are most concentrated.

Statistical analyses

Because they have different mechanisms of transposition, the major types of elements (LINEs, SINEs, LTRs, and DNA transposons) were analyzed independently within each species. I log + 1 transformed all variables (both dependent and independent) associated with TE families and first performed simple linear regression between PPE and all independent variables. I then used bi-directional stepwise regression analyses using Akaike's Information Criteria to choose the best sub-model from a full model that included all independent variables to explain the most PPE variation. To explore the relevance of each independent variable in the chosen sub-model, I used the lmg method available in the R package relaimpo which averages sequential sums of squares over the ordering of regressors.

Results

We sought to better understand the interplay between TEs and piRNAs among mammals by comparing TE activity and piRNA repertoires in dog, horse and the big brown bat. Patterns of TE activity are often inferred based on the relative abundance of TEs in a given genome. However, for the purpose of this study, it was critical to distinguish between genome-wide patterns of TE accumulation and levels of TE expression, which are related but need not necessarily be the same. The first reflect historical patterns of TE deposition and retention, whereas the second reflect the TE stress currently challenging a given genome. Both of these factors could impact piRNA production, as the abundance of a given TE in the genome could directly relate to its potential as a source of primary piRNAs, and TEs that are actively transcribed are expected to contribute more to the pool of piRNAs in the ping-pong cycle.

SINE, LINE, LTR, and DNA transposon insertions are grouped into discrete families based on overall similarity and are often represented by a single consensus sequence. This consensus sequence is considered the best estimate of the mobilizing elements for that particular family. There were 745, 787, and 976 distinct TE families annotated by RepeatMasker in the dog, horse, and bat genomes, respectively, corresponding to 150-159 LINE, 20-26 SINE, 283-430 LTR, and 280-376 DNA transposon families. For each separate TE family I calculated 1- the number of insertions, 2- the relative age of the family, 3- the average length of insertions (in two categories, among all insertions and only those within piRNA clusters), and 4- the abundance of transcripts. There were clear differences in patterns of TE accumulation and expression among the three genomes that may allow us to tease apart what drives the production of TE related ping-pong piRNAs when the types of TE families, insertion numbers, expression and genomic proportion vary.

Genomic TE composition and properties

Among retrotransposons, LINEs occupied the most genomic space in all three species, accounting for ~10% of the genome, followed by SINEs, ~3 to 8%, and LTR retrotransposons, ~ 1 to 3% (fig. 1A). Based on number of insertions, LINEs were also the most abundant TEs in the horse genome, but SINEs and DNA transposons were the most abundant in the dog and the bat, respectively with over 1.65×10^6 insertions in all cases (fig. 1A). The bat genome stands out in this regard, as it has experienced a resurgence of DNA transposons when compared to most mammals, a characteristic shared with the closely related little brown bat, *Myotis lucifugus* (Pritham and Feschotte 2007; Ray et al. 2007, 2008; Mitra et al. 2013; Platt et al. 2014). I estimated that

approximately 11% of the bat genome derives from DNA transposon insertions, in contrast to ~1% in dog or horse (fig. 1A).

The historical patterns of TE accumulation also vary among these three species (fig. 1B). In the recent past, the dog genome has accumulated LINE and SINE insertions at a higher rate than either the horse or bat. Some LINEs have been deposited relatively recently in the horse genome, yet young LINE insertions are almost undetectable in the bat genome. Similarly, recent SINE insertions are very uncommon in the bat and horse genomes while in the dog, SINEs have accumulated at a relatively high rate. Recent DNA transposon insertions are uncommon in all three species. However, the bat differs from dog and horse in that there was a high rate of DNA transposon deposition in the recent past (fig. 1B). This is also seen in *Myotis lucifugus*, which diverged from the big brown bat lineage ~ 25 million years ago (Miller-Butterworth et al. 2007; Ray et al. 2007; Pagán et al. 2012; Platt et al. 2014). Despite the clear slowdown in DNA transposon accumulation in the genome of the big brown bat, these elements have remained the dominant TE type.

pilRNAs formed clusters, which were not enriched for TEs

We then moved on to characterize pilRNA diversity in these three species. The sequenced small RNAs were similar to previously characterized piRNAs extracted from other mammalian testes (Lau et al. 2006; Yan et al. 2011; Liu et al. 2012). Specifically, more than 75% of the sequenced small RNAs were between 24 and 32 nucleotides long and there was a strong uridine bias in the first base position (fig. 2A), consistent with previously described pilRNAs. Allowing one mismatched base between the pilRNA and genome alignment, between 51 and 72% of the unique pilRNA sequences mapped to

each genome, and the majority of these pilRNAs map to non-TE related genomic space (fig. 2B), which is characteristic of the pachytene piRNAs. Interestingly, the proportions of LINE, SINE, LTR, and DNA transposon derived pilRNAs were similar among the three species.

We then compared TE content within pilRNA clusters against genome wide patterns of TE accumulation. I restricted my analyses to clusters longer than 10Kb because these are more likely to contain full length TE insertions. The annotated clusters generally occupied unannotated space and generated ~50% of the unique pilRNAs. I annotated 290, 376, and 221 clusters in the dog, horse and bat genomes, respectively. By comparison, groups have annotated approximately 100 clusters in the mouse (Girard et al. 2006; Beyret et al.2012; Li et al. 2013). Although I annotated many more clusters in these genomes, cluster variation among species is typical. For example, Chirn et al. (2015) found that most piRNA clusters were species specific, few were conserved among species, and the number of piRNA clusters varied drastically.

Although piRNA clusters are derived from unannotated space, there were several TE insertions within any given cluster. For example, in a large cluster shared between the three species there were between 53 and 191 TE insertions, most of which were more than 20% diverged from the consensus sequence (fig. 2C). In fact, more than 98% of clusters included one or more TE insertions. However, I did not find that clusters were necessarily enriched for insertions, rather the number of genomic insertions from each family was tightly correlated with the total number of insertions among all clusters (fig. 2D), a relationship also observed in Hirano et al. (2014), suggesting that clusters do not preferentially accumulate TE insertions from particularly deleterious TE families.

Ping-pong response

The next step was to explore relationships between different TE family characteristics and ping-pong pilRNA expression (PPE) using bivariate and multivariate regression analyses. To estimate the expression of ping-pong pilRNAs, I mapped all pilRNA sequences to the TE consensus sequences and only calculated expression of pilRNAs that exhibited the signature of the ping-pong cycle, i.e. 10 bp overlap between pairs of pilRNAs where a uridine is in the first position of the primary pilRNA, and an adenine is in the 10^{th} position of the secondary pilRNA. As expected, only a small percentage of pilRNAs mapped to the consensus sequences (~3-6%), and half of these pilRNAs were found as ping-pong pairs. This was expected because pachytene pilRNAs are the most abundant in mature testes and are generated independently of the ping-pong cycle (Beyret et al. 2012).

To estimate the level of pilRNA response I initially discriminated between sense, anti-sense, and total PPE. However, because of the high correlation ($r^2 > 0.95$) observed among the three measurements, I only measured the impact of TE parameters on total PPE. Statistical analyses were performed for all TE families combined and separately for LINEs, SINEs, LTRs and DNA transposons among species. When I examined parameters individually, I found that the largest r^2 values were most often associated with estimates of TE family expression, especially in LINEs, SINEs, and DNA transposons in the bat ($r^2 = 0.47–0.81$, $p < 0.001$; fig. 3A), i.e. the most expressed families also generate the most ping-pong pilRNAs. In addition, I found that the estimated age of the TE family was also a strong predictor of ping-pong pairs among LINEs in all three species ($r^2 = 0.51–0.60$, $p < 0.001$; fig. 3A). I tested whether RSEM's mapping parameters potentially biased

RNASeq reads to younger elements by increasing the number of allowed mismatches in the seed region. I found that increasing mismatches did not change the overall pattern that younger elements had higher expression and only made inferences from the default parameters. This observation is not unexpected given that the youngest TE families are often the most expressed, aligns with the predictions of the ping-pong model, and Lukic and Chen (2010) and Mourier (2011) also reported this property. By contrast, variables related to the abundance of TEs in the genome and piRNA clusters, such as insertion number, total bases, and average length typically had much lower r^2 values suggesting that they are not as important with regard to the ping-pong response. The most obvious exceptions to this general pattern involved SINEs in the dog (see discussion below).

In order to explore relationships between TE metrics and PPE in a multivariate framework I combined all variables into a single model, and used stepwise regression to find a sub-model with an optimal r^2. Stepwise regression also selects optimal variables when one or more variables correlate, as is the case with genome insertions and cluster insertions (fig. 2D). Based on genomic insertions (fig. 1), LTR families generally appeared to not be transcribed or accumulating in all three species, and only estimates of subfamily expression had any meaningful and significant relationship with PPE for these TEs. Because of this, LTRs were excluded from the multivariate regression. For SINEs, LINEs, and DNA transposons, the parameters selected and their relative contribution to PPE varied among each element type and species. Between two and six independent variables were selected for each multivariate regression model (fig. 3B). With the exception of DNA transposons in dog and horse, both of which have very old families that are no longer transcribed, the models yielded high r^2 values (between 0.7 and 0.92),

and in most cases included TE expression as the most important variable in the model. A second common parameter selected among species and TE types was TE family age, which when selected, always had a negative relationship with PPE, i.e. younger families had higher PPE. When all TE families were combined, the number of cluster insertions was selected in all three species. However, when TEs were separated by type, cluster insertions were only meaningfully selected as part of the horse SINE and bat DNA transposon models (fig. 3B). The remaining piRNA cluster parameters, if selected by stepwise regression were typically not among the most influential parameters in the model, and had negative relationships with PPE.

TE expression is predicted to be the driver of the abundance of ping-pong piRNAs based on the ping-pong model. The bivariate and multivariate regression analyses generally confirmed this primary prediction yet there was variation among species and TE types. To illustrate the variation, I plotted TE expression and PPE for each family against family age (as proxied by median genetic distance) (fig. 4). The plot illustrates those TE families that are the most highly transcribed appear to elicit the strongest ping-pong response. This is particularly true for LINEs, but the relationships between expression and PPE appear to be more complex in SINEs. I found that the largest fraction of TE transcripts corresponded to SINEs in all species, ranging from ~50% in bat to 80% in dog. However, the piRNA response to SINE expression varies greatly among the three taxa. For example, in horse, SINEs are the most highly expressed TEs and also elicit the strongest piRNA response (fig. 4), whereas expression of ping-pong pilRNAs in the dog correlates more with the total number of SINE family insertions (fig. 3A) rather than with SINE expression. Moreover, in the dog SINE PPE is generally

much lower than SINE transcription levels (fig. 4). In the bat, the Ves SINE family is the only family significantly expressed, yet elicits a weak ping-pong response (fig. 4). Thus, while a pattern similar to that observed in mouse for LINEs is recapitulated in these three taxa, no such obvious pattern emerges for SINEs. Finally, there does appear to be a correlation between DNA transposon expression and PPE in the bat (fig. 3A). However, because recent DNA transposon activity is unique to the bat, a meaningful comparison with dog and horse is not possible.

Discussion

The relationship among TE expression, piRNAs, and TE accumulation is not entirely clear and it is difficult summarizing the outcome of these interactions for several reasons. For example, while some TE families have deleterious members capable of transcription, many families are completely transpositionally inert yet make up sizable proportions of the genome. This discrepancy makes teasing apart the relative contributions of self-transcribed TEs and those that are spuriously transcribed difficult. Among vertebrates, the mouse has been the primary model used to study the piRNA pathway. However, despite extensive investigation into the mechanisms of the ping-pong cycle, almost nothing has been described about the evolutionary dynamics between piRNAs and TEs. Indeed, the mouse genome exhibits only one of the many patterns of TE transcription and accumulation that exist among mammals and piRNA dynamics among taxonomically varied vertebrate genomes has not been investigated. Therefore I examined three laurasiatherian mammals with markedly different TE landscapes, patterns of TE expression and pilRNA repertoires to better understand the complex relationship between host defenses and TE accumulation over time, in an attempt to identify general

patterns. I tested parameters commonly associated with TE activity and in line with predictions from the mammalian ping-pong model, univariate and multivariate analyses identified the abundance of TE transcripts as a strong predictor of the abundance of ping-pong pilRNAs.

Ping-pong piRNAs target the most transcribed families

The primary prediction of the ping-pong model suggests that the TE families that are most transpositionally active, those that are the most deleterious, would be the most abundant representatives among piRNAs. I generally found that this prediction was satisfied (See fig. 4). In both the bivariate and multivariate tests, when all TE families were considered, the abundance of family transcripts was the best predictor of ping-pong piRNAs. When comparing within the different element types, the families that were the most transcriptionally active, usually elicited the strongest pilRNAs response. Similarly, Mourier (2011) found that among mouse TEs, the youngest families often had the most mapped piRNAs and there was some proportionality between cellular TE transcripts and piRNAs.

DNA transposons are the most abundant recently active elements in the big brown bat genome. Thomas et al. (2014) found evidence of low-level ongoing Helitron accumulation and Mitra et al. (2013) and Ray et al. (2008) suggested that piggyBac elements were still accumulating in the related bat, *M. lucifugus,* raising the possibility that these elements are still actively inserting in this genome but that such activity is simply too low to be detected. There was an abundance of DNA transposon transcripts in the testis transcriptome of the big brown bat and there was also a statistically significant response to these expressed elements (fig. 3A, 3B; fig. 4). This was an interesting and

unexpected response for several reasons. DNA transposons do not require an RNA intermediate for transposition, therefore it was unexpected to detect RNA associated with these elements, especially elements that are annotated as non-autonomous; which do not encode the proteins needed to mobilize. There are at least two plausible explanations as to why I observed expressed DNA transposons and a subsequent piRNA-response. First, there are large numbers of DNA transposon insertions in the bat genome. The sheer density of these insertions suggests that at least a subset will exist in close proximity to a promoter, leading to spurious transcription and incorporation into the ping-pong cycle. Second, several families of transposons harbor promoters that act to encourage transcription of their transposase and those transcripts could be targeted by piRNAs. Regardless of the mechanism, results indicate a statistically significant relationship between DNA transposons and ping-pong pilRNAs that may suggest a defensive response. The strong response to DNA transposons may suggest an adaptable defense to both Class I and Class II TEs. Because vesper bats are the only known vertebrate to harbor actively mobilizing DNA transposons, this relationship could be worthy of additional investigation.

piRNA cluster likely do not regulate TEs in mammals

In contrast to *Drosophila*, where piRNA clusters are thought to give rise to primary piRNAs (Kelleher and Barbash, 2013), TE transcripts are proposed as the substrates for primary piRNA processing in mammals (Aravin et al. 2008 and Girard et al. 2006). However, dissenting views exist. For example, Ha et al. (2014) and Hirano et al. (2014) both suggested pachytene piRNA clusters could be a source of anti-sense

piRNAs used in the ping-pong cycle. Here, I further concluded that piRNAs derived from these clusters are likely not involved in the TE silencing pathway.

We tested the role of piRNA clusters as determinants of the ping-pong response in a statistical framework and found that the abundance of insertions within clusters strongly correlated with the overall prevalence of genome insertions in all three species, but was not the most important factor with regard to PPE. Furthermore, I found total bases in clusters and the median age of cluster insertions did not correlate well with PPE. Interestingly, when all TE families were taken into consideration under the multivariate framework, cluster insertions were included in the final models, but TE expression was always the most important contributor to PPE. These results suggest there may be some role that piRNA clusters play in TE silencing, but it is likely minimal.

Unfortunately, this will likely continue to be a confounding factor. Because of the many TE insertions that exist in mammalian genomes, it is difficult to determine the ultimate source of any single TE-derived pilRNA, much less whether it arose from a mobilizing TE transcript or from an insertion that simply lies within a piRNA cluster. However, the families with the most cluster insertions were generally older, had few mapped ping-pong pilRNAs, and there was ultimately no relationship between the number of insertions in clusters and the abundance of ping-pong pilRNAs. However, because I cannot fully identify the source of TE derived piRNAs, I cannot fully rule out the notion that some TE silencing piRNAs are derived from piRNA clusters. Perhaps piRNAs that are processed from cluster insertions are incorporated into the ping-pong cycle, but unlike in *Drosophila* where clusters seem to act as TE “traps”, piRNA clusters do not appear function this way in mammals.

Complex relationship between TE accumulation and genome defense

These results have implications for understanding the relationship between TE transcription and accumulation. It is generally assume a dearth of recent deposition for a given TE is related to decreases in TE transcription in the recent past. Among LINEs in the three species this presumption appeared generally true. There is a large abundance of young LINE transcripts in the dog, and an abundance of young LINE insertions, by contrast, there is little expression of full length LINEs in the bat and very few recent insertions and the horse is intermediate between the two. Also, data suggest that the representation of LINE families in the piRNA pool among the three species fit the predictions of the ping-pong cycle.

SINEs presented a very different case. Initially, based on their genomic abundance, I expected that SINEs would only be highly expressed in the dog. However, after taking their sequence length into account, young SINE families were the most transcribed elements in all three species, but recent SINE accumulation is only seen in the dog genome. Fig. 4 suggests that the abundance of SINE insertions in the dog could be the result of a reduced pilRNA response. Although young SINE families have the highest abundance of ping-pong pilRNAs in the dog, the ping-pong response appears weak in relation to the level of SINE expression. The opposite is true in the horse genome, where SINEs are being expressed at comparable levels but the ping-pong piRNAs appear to offer a formidable response, potentially preventing the eventual reverse transcription and insertion. In bats, yet a third scenario appears to have played out, revealing the complexity of this system. There, the Ves SINE family is highly transcribed, the piRNA response appears limited, yet recent Ves insertions are not highly represented in the bat

genome. Unlike LINEs, which are completely autonomous, SINEs depend on the reverse transcriptase and endonuclease genes encoded by LINE elements to fully mobilize and this property may be responsible for the lack of a strong pattern similar to that among LINEs. This also raises the question of whether piRNA+ PIWIs complexes preferentially different target elements.

In summary, the data indicates that the level of piRNA response against a given TE subfamily is most strongly associated to the abundance of the corresponding transcripts, with other factors, such as the age of the subfamily playing a more modest role. These analyses suggest that piRNA responses are able to provide protection against TE invasion in mammalian genomes, but that TEs are still able to propagate even in the presence of a putatively robust piRNA response. Furthermore, it appears that the interplay between TEs and piRNAs is distinct among species and TE types. Expanding comparative studies of piRNAs and TEs to a broader array of mammals could help uncover a general model to account for the relationship between TE abundance at the genome and the piRNA response.

Acknowledgements

We acknowledge support from the National Science Foundation (EPS-0903787, DBI-1262901 DEB-1354147, MCB-0841821 and DEB-1020865). Additional support was provided by the College of Agriculture and Life Sciences at Mississippi State University and the College of Arts and Sciences at Texas Tech University. Tissues were provided by the College of Veterinary Medicine at Mississippi State University, Jeremiah Dumas, and the Natural Sciences Research Laboratory at the Museum of Texas Tech University. The Broad Institute Genomics Platform and Genome Sequencing and

Analysis Program, Federica Di Palma, and Kerstin Lindblad-Toh made the data for Eptesicus fuscus available. Raw small RNA and transcriptome sequences are available under the BioProject ID PRJNA290346.

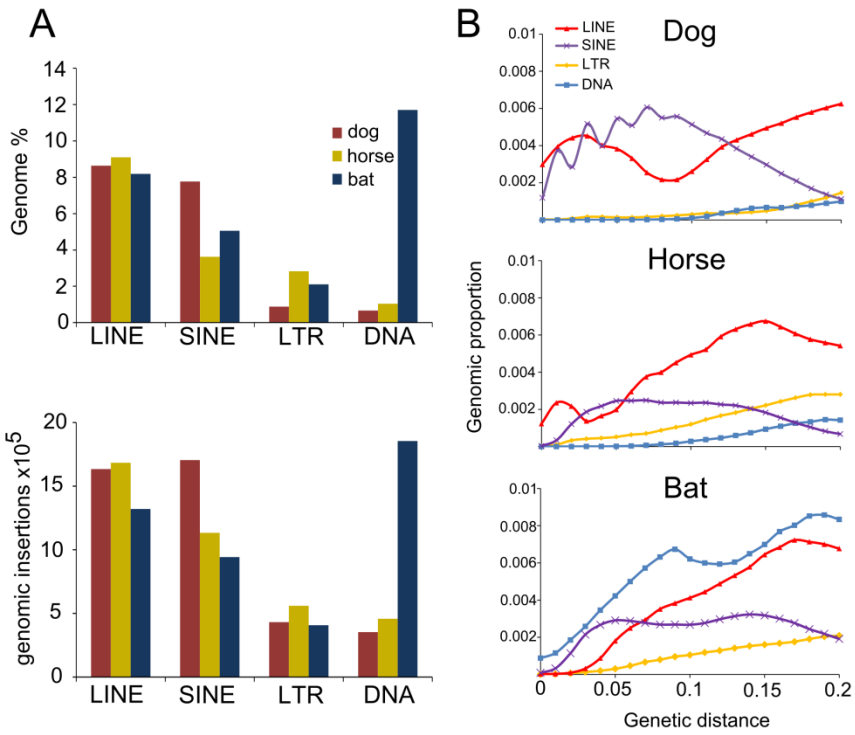

Figure 4.1 Genomic TE characteristics

A) The percentage of each genome composed of major TE types and the number of insertions for each type. Calculations were made from the insertions that were less than 0.2 divergent from the consensus sequence. B) The temporal contribution of major TE types in each genome. Insertions with lower genetic distances were deposited more recently.

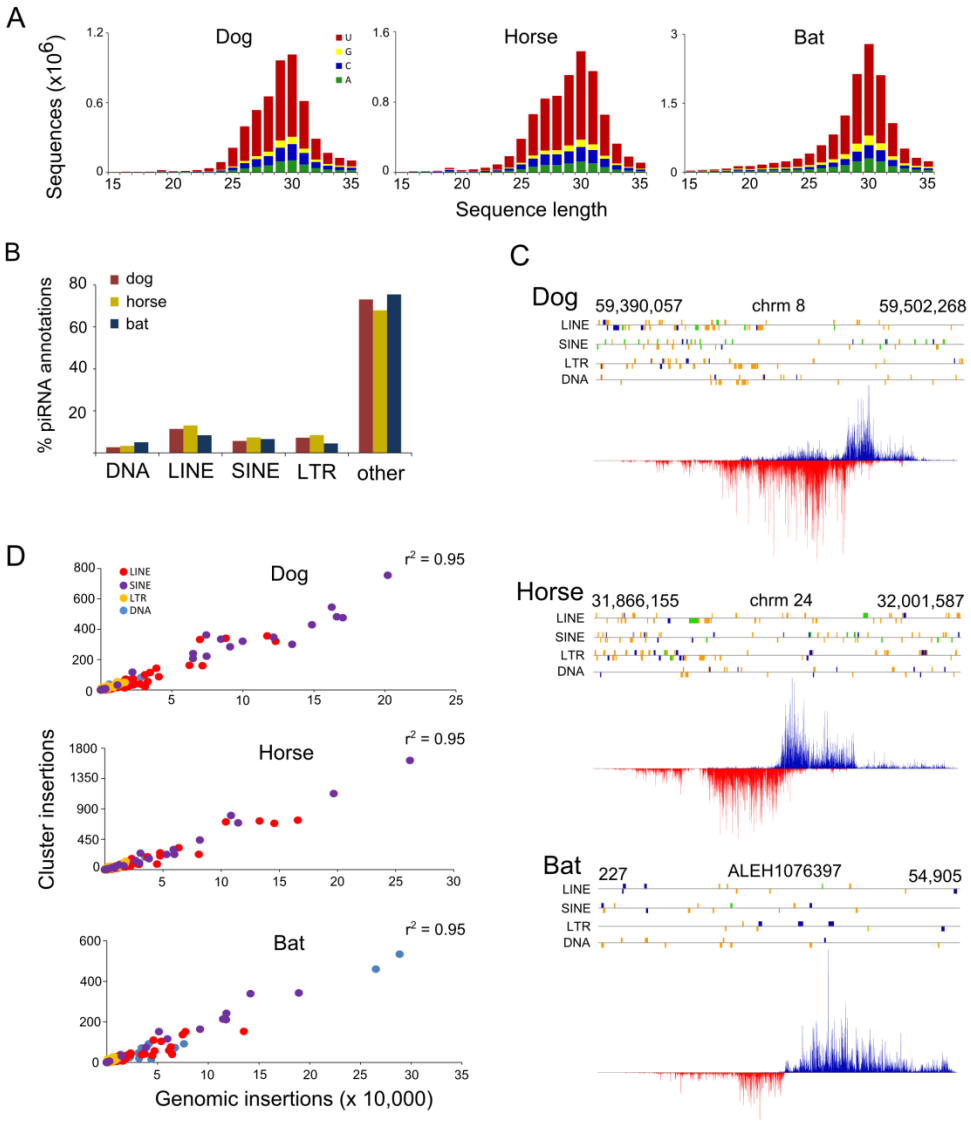

Figure 4.2 piRNA characteristics

A) The length and distribution of unique small RNA sequences presented with the frequency of the first nucleotide illustrating the 5' U bias. B) Proportion of singly mapping pilRNAs that mapped to TE and non-TE space. C) The TE content of one homologous cluster found in the dog, horse and bat. TE insertions with genetic distances less than 0.1 from the family consensus are colored green, between 0.1 and 0.2 divergent are blue, and greater than 0.2 are orange. pilRNAs that mapped anti-sense relative to the contig are red, and sense pilRNA are blue. D) The raw number of genomic insertions plotted against the total number of cluster insertions per TE family. r^2 values from simple linear regressions between the two variables are reported, $p < 0.001$ in all cases.

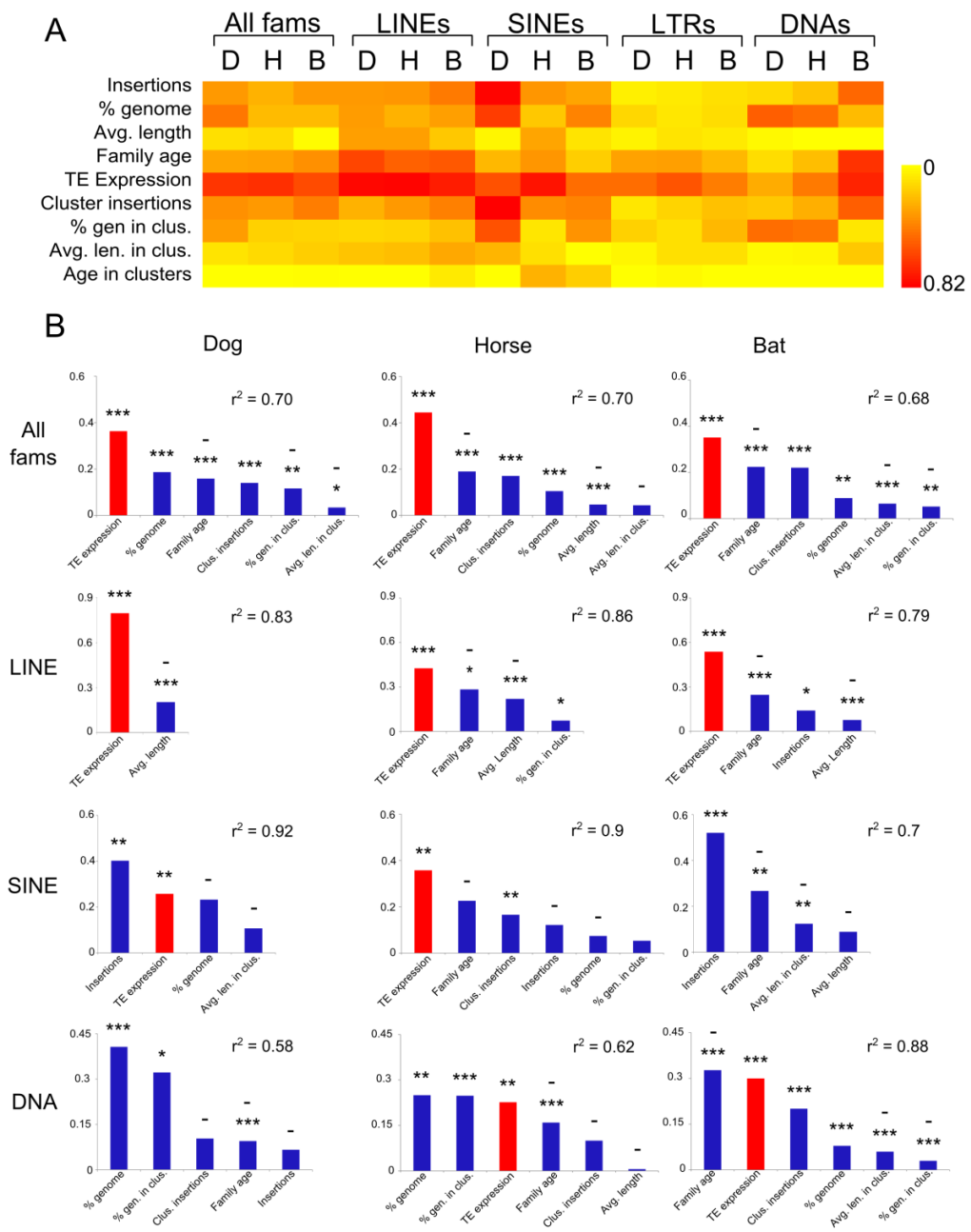

Figure 4.3 piRNA univariate and multivariate statistics

A) Heat map representing r2 values for independent linear regressions between PPE and each independent variable for dog (D), horse (H) and bat (B). B) The independent variables selected for each step-wise regression analysis and relative importance of each variable in the model. TE expression is colored red. Negative interactions are indicated by a "-" above the variable. For each full model I reported the corresponding r2 values. Abbreviations: avg: average, gen: genome, clus: cluster, len: length. * $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$

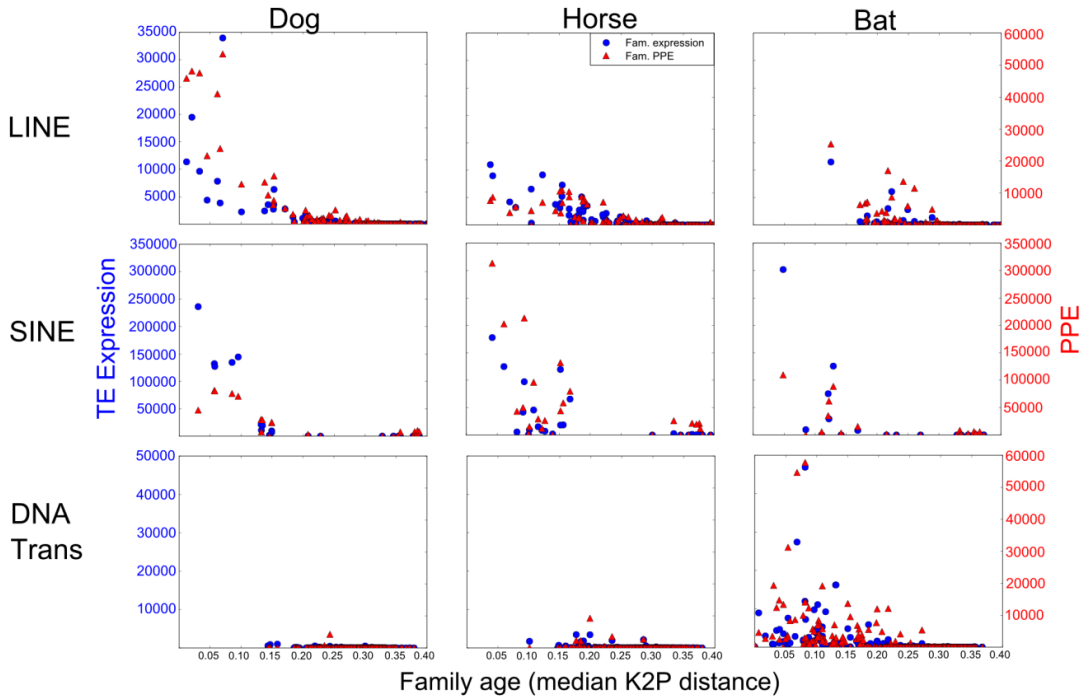

Figure 4.4 TE expression vs piRNAs

Separate dot plots illustrating the relationship between TE expression, TE family age, and PPE. For each TE family, I plotted the expression values (blue) and PPE (red) plotted against family age (median K2P) separately for LINEs, SINEs and DNA transposons in each species.

References

Aravin AA et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. Nature 442:203–207.

Aravin AA et al. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 31:785–799.

Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764.

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316:744–747.

Beyret E, Liu N, Lin H. 2012. piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell Res. 22:1429–1439.

Brennecke J et al. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1089–1103.

Brookfield JFY, Johnson LJ. 2006. The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? Genetics 173:1115–1123.

Callinan PA et al. 2005. Alu retrotransposition-mediated deletion. J Mol Biol. 348:791–800.

Chirn G-W et al. 2015. Conserved piRNA expression from a distinct set of piRNA cluster loci in eutherian mammals. PLoS Genet. 11:e1005652.

De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7:e1002384.

Gilbert N, Lutz-Prigge S, Moran J V. 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110:315–325.

Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442:199–202.

Gou L-T et al. 2014. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 24:680–700.

Ha H et al. 2014. A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. BMC Genomics 15:545.

Han K et al. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res. 33:4040–4052.

Hirano T et al. 2014. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA 20:1223–1237.

Höck J, Meister G. 2008. The Argonaute protein family. Genome Biol. 9:210.

Kapitonov V V, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. 23:521–9.

Kelleher ES, Barbash D a. 2013. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. Mol Biol Evol. 30:1816–29.

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

Langmead B, Trapnell C, Pop M, Slazberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Lau NC et al. 2006. Characterization of the piRNA complex from rat testes. Science 313:363–367.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Li XZ et al. 2013. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell. 50:67–81.

Liu G et al. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res. 13:358–368.

Liu G et al. 2012. Discovery of potential pirnas from next generation sequences of the sexually mature porcine testes. PLoS One 7:e34770.

Lukic S, Chen K. 2011. Human piRNAs are under selection in Africans and repress transposable elements. Mol Biol Evol. 28:3061–3067.

Meredith RW et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524.

Miller-Butterworth CM et al. 2007. A family matter: conclusive resolution of the taxonomic position of the long-fingered bats, miniopterus. Mol Biol Evol. 24:1553–1561.

Mitra R et al. 2013. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc Natl Acad Sci. USA. 110:234–239.

Mourier T. 2011. Retrotransposon-centered analysis of piRNA targeting shows a shift from active to passive retrotransposon transcription of developing mouse testes. BMC Genomics 12:440.

O'Donnell KA, Boeke JD. 2007. Mighty Piwis defend the germline against genome intruders. Cell 129:37–44.

Pagán HJT et al. 2012. Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats. Genome Biol Evol. 4:575–585.

Platt RN et al. 2014. Large numbers of novel miRNAs originate from DNA transposons and are coincident with a large species radiation in bats. Mol Biol Evol. 31:1536–1545.

Pritham EJ, Feschotte C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proc Natl Acad Sci USA. 104:1895–1900.

Ray DA et al. 2008. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. 18:717–728.

Ray DA, Pagan HJT, Thompson ML, Stevens RD. 2007. Bats with hATs: evidence for recent DNA transposon activity in genus Myotis. Mol Biol Evol. 24:632–639.

Reuter M et al. 2011. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature 480:264–267.

Saito K, Siomi MC. 2010. Small RNA-mediated quiescence of transposable elements in animals. Dev Cell. 19:687–697.

Sen SK et al. 2006. Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet. 79:41–53.

Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: the vanguard of genome defense. Nat Rev Mol Cell Biol. 12:246–258.

Thomas J, Phillips CD, Baker RJ, Pritham EJ. 2014. Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. Genome Biol Evol. 6:2595–610.

Yan Z et al. 2011. Widespread expression of piRNA-like molecules in somatic tissues. Nucleic Acids Res. 39:6596–6607.

Yohn CT et al. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. PLoS Biol. 3:e110.

CHAPTER V

CONCLUSIONS

This dissertation was used as an opportunity to study genomic components, primarily transposable elements (TEs), protein coding genes and small RNAs, to understand the complexity of vertebrate genome evolution. My objective was to study the genome evolution at three different scales; the first being the evolution of genes under sexual selection, the second being the evolution of a gene family responsible for relaying environmental information to the brain, and finally I studied how encoded proteins and small RNAs in the genome regulate and reduce the harm caused by TE mobilization.

In chapter II, I illustrated how a subunit (*Abpa*) of androgen binding protein (ABP) is a candidate protein for mate recognition in both New and Old World mouse-like rodents (Muroids). In both groups of Muroids, I found accelerated rates of nonsynonymous substitutions within and among genera, consistent with positive selection. More specifically, I found that five and 12 amino acid sites exhibit evidence of positive selection in the New and Old World rodent taxa evaluated. The amino acid sites that were under selection did not appear to alter the structure of the protein and most were in exposed regions of the protein, consistent with previous findings (Emes et al. 2004). These results along with the strong evidence that *Abpa27* plays an important role in speciation in *Mus* suggest changes among non-essential residues might play a significant role in the evolution of this protein and facilitate mate recognition. Additionally, despite

limited sample size, I discovered high levels of intraspecific variation among species within *Apodemus*, *Peromyscus* and *Reithrodontomys*. Therefore, the suggested role of the *Abpa* gene in New World muroids might be slightly different relative to the *M. musculus* complex where different alleles segregate with different subspecies (Bímová et al. 2005; Bímová et al. 2011).

In rodents, chemical cues are critical for mice evidenced by the approximately 1000 functional olfactory receptors and 212 vomeronasal receptors encoded in the mouse genome (Zhang and Firestein 2002; Shi and Zhang 2009). Currently, the patterns of ABP evolution are hypothesized to facilitate assortative mating, with evidence of coevolution between ABP pheromone and vomeronasal receptors of the V1R receptor family (Karn et al. 2010). Therefore, these two systems, ABP and V1R receptors, may facilitate the maintenance of species boundaries in Muroids.

In chapter III, I focused on the largest gene family in vertebrates. The olfactory receptor (OR) gene family directly communicates environmental information to the brain, and the composition of this gene family can provide clues as to how a species gathers information from environmental cues. I showed there was remarkable diversity of olfactory receptor (OR) repertoires among sauropsid (bird and reptile) lineages, and indicate that most sauropsids have diverse, relatively large, and highly lineage-specific repertoires. Genomic queries from representative species (anole, python, zebra finch, chicken, softshell turtle, painted turtle, crocodile, alligator, and gharial) determined that most of the OR subfamilies described in mammals are also present in sauropsids, which indicates these subfamilies were likely present in the common ancestor of all amniotes. I found that the number of intact ORs ranged from 100 in the green anole to 1100 in the

softshell turtle. Among vertebrates, turtles and crocodilian genomes are both notable for evolving very slowly (Shaffer et al. 2013; Green et al. 2014), yet the evolutionary history and diversity of their OR repertoires was remarkably different between the two lineages. The OR repertoires of turtles are very evolutionarily dynamic, and demonstrate multiple species-specific expansions. In contrast, crocodilian OR repertoires have experienced far less change in gene number and diversity among species. These observations collectively suggest that not only have crocodilians not experienced substantial change in the number and diversity of OR genes, but have also experienced fewer amino acid changes among orthologous OR genes since extant crocodilians diverged from the common ancestor. The largest difference in gene number was observed between snake and lizard (squamate) lineage. Both squamates have retained a very diverse ancestral OR repertoire, yet the python has at least five times as many putatively functional genes than the green anole, likely due to its reliance on chemosensory information. Finally, the most extreme OR repertoire was found in the two birds. Interestingly, the chicken and zebra finch are even remarkable among birds (Khan et al. 2015). These two birds have major expansions of subfamily 14 ORs and these expansions occurred independently in each species. The zebra finch has fewer intact ORs and a greater number of pseudogenes than the chicken.

Saurapids began radiating ~300 mya and there are more than 20,000 extant species that span an extensive array of natural histories, habitats, and ecological niches that may require specific olfaction and chemoreception specializations. My data leads to the expectation that OR repertoirs are remarkably diverse across all sauropsid lineages, and that sauropsids may be an excellent model system in which to study OR gene birth and death, and the forces that drive these patterns.

In the last major research chapter, chapter IV, I focused on the interaction between TEs and genomic defenses against TE proliferation. In humans, TEs make up between 45-70% of the genome (Lander et al 2001; de Koning et al. 2011), most of this content is "dead", i.e. the insertions are no longer capable of duplicating, but a fraction of these elements are capable of mobilization. I this chapter, I compared the PIWI interacting RNA (piRNA) and TE dynamics in dog, horse, and big brown bat, three species with markedly different TE landscapes and expression profiles to better understand the complex evolutionary relationship between host genome defenses and TEs.

I found that within major TE types, i.e. LINEs, SINEs, LTRs, and DNA transposons, the abundance of TE transcripts was the best predictor of the abundance of ping-pong pilRNAs, suggesting a direct response to TE expression, a property not found in *Drosophila* (Kelleher and Barbash 2013), but a property that is predicted by the mammalian ping-pong cycle. The second major conclusion I found was that piRNA clusters had little if any involvement in TE silencing in mammals, contradictory to what previously authors have proposed (Ha et al. 2014; Hirano et al. 2014).

It is generally assumed that a decline in TE deposition is related to a decline in transcription and this was generally true among LINEs. My data from LINEs indicate that the representation of LINE families among ping-pong piRNAs in the three species fits the predictions of the ping-pong cycle, i.e. the element families that are transcribed the most, are the most deleterious, and would be most targeted by piRNAs. However, SINEs deviated from this prediction. I observed that young SINE families were the most transcribed elements in all three species, but the piRNA response varied and recent SINE

accumulation is only notable in the dog genome. I speculate that the abundance of SINE insertions in the dog could be the result of a reduced pilRNA response. The piRNA response to SINEs in the dog appears weak in relation to the level of SINE expression. By contrast, SINEs are being expressed at comparable levels in the horse but piRNAs appear to offer a stronger response, preventing more insertions. However, unlike LINEs, SINEs are not fully autonomous and depend on LINE machinery to fully mobilize and it is possible that this property may also be responsible for the lack of a strong pattern.

Of the research topics presented here, both the ORs and piRNA pathway have been expounded on. For example, I have coauthored a publication at Genome Biology and Evolution describing the evolution of a transcription factor (A-MYB) responsible for the expression of pachytene piRNA precursors (Campanini et al. 2015). Li et al. (2013) found that A-MYB regulated the expression of some if not all pachytene piRNA precursor transcripts and pachytene piRNAs have only been described in vertebrates. The exact function of pachytene piRNAs is still under investigation, although Gou et al. (2014) suggested they remove gene transcripts from the cytoplasm. I and co-authors found drastically different structural differences and expression patterns between cyclostome and gnathostome A-MYB proteins and predicted that pachytene piRNAs are an innovation specific to gnathostomes. Furthermore, some of the work I conducted for the OR gene family also made appearances in two genome project papers, for the python (Castoe et al. 2013) and crocodilians (Green et al. 2014). Moreover, I have been asked by a collaborator to query the recently generated garter snake genome and perform analyses related to OR evolution in the garter snake relative to other squamates. As more squamate genomes become available, a second full paper will also be constructed.

The primary objectives of this dissertation were to study vertebrate genomes, more specifically the evolution of genes and the regulation of TEs. During this process, I identified a protein that appears to be involved with mate selection for all mouse-like rodents, not just *Mus*. The OR repertoires from the chicken and green anole were solely used to describe OR gene family evolution for all sauropsids and were thought to be relatively simple. Yet, after annotating a phylogenetically more diverse set of genomes, I found that ORs are incredibly diverse among taxa within this group. Lastly, it was assumed that the abundance of TE insertions could be related to TE expression, and those that are being the most expressed would be the most common in the genome. However, I have determined that in many cases, the abundance of TE transcripts is a poor predictor of insertion abundance. Although piRNAs and PIWI proteins act to silence TE transcripts, it is still unclear how their involvement impacts TE accumulation over time.

In conclusion, the evolution and genes and TEs involve complex processes which can become difficult to untangle. The work presented here addressed three questions in three disciplines of evolutionary biology. Although a high volume of data was processed, inferences for each chapter were derived from a restricted number of taxa. As the quality of sequencing improves, availability and diversity of genomes increases, and computation advances, more data points from a diverse array of taxon can be collected and used to resolve biological history.

References

Bímová B, Karn RC, Piálek J. 2005. The role of salivary androgen-binding protein in reproductive isolation between two subspecies of house mouse: Mus musculus musculus and Mus musculus domesticus. Biol. J. Linn. Soc. 84:349–361.

Bímová BV, MacHolán M, Baird SJE, Munclinger P, Dufková P, Laukaitis CM, Karn RC, Luzynski K, Tucker PK, Piálek J. 2011. Reinforcement selection acting on the European house mouse hybrid zone. Mol. Ecol. 20:2403–2424.

Campanini EB, Vandewege MW, Pillai NE, Tay B-H, Jones, JL, Venkatesh B, Hoffmann FG. Early evolution of vertebrate Mybs: an integrative perspective combining synteny phylogenetic, and gene expression analyses. Genome Biol Evol. 7:3009–2021.

Castoe TA et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc Natl Acad Sci. 110:20645–20650.

De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7:e1002384.

Emes RD, Riley MC, Laukaitis CM, Goodstadt L, Karn RC, Ponting CP 2004. Comparative evolutionary genomics of androgen-binding protein genes. Genome Res. 14:1516–1529.

Green RE, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346:1254449.

Gou L-T et al. 2014. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 24:680–700.

Karn RC, Young JM, Laukaitis CM. 2010. A candidate subspecies discrimination system involving a vomeronasal receptor gene with different alleles fixed in M. m. domesticus and M. m. musculus. PLoS One 5:1–11.

Kelleher ES, Barbash DA. 2013. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. Mol. Biol. Evol. 30:1816–1829.

Khan I, et al. 2015. Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. Mol Biol Evol. 32:2832–2843.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921

Li XZ et al. 2013. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell. 50:67–81.

Shaffer HB, et al. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14:R28.

Shi P, Zhang J. 2009. Extraordinary diversity of chemosensory receptor gene repertoires among vertebrates. Results Probl. Cell Differ. 47:1–23.

Zhang X, Firestein S. 2002. The olfactory receptor gene superfamily of the mouse. Nat. Neurosci. 5:124–133.