LEVERAGING THE DYNAMIC REPEAT ELEMENT LANDSCAPE OF SQUAMATE REPTILE GENOMES TO UNDERSTAND BROAD PATTERNS OF VERTEBRATE GENOME EVOLUTION AND TRANSPOSABLE ELEMENT BIOLOGY

by

GIULIA IRENE MARIA PASQUESI

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

DECEMBER 2019

Supervising Committee:
Todd A. Castoe, Supervising professor
Esther Betran
Jeff P. Demuth
Matt K. Fujita
Matt R. Walsh

Acknowledgements

The last five years of my life have often been challenging, and yet they represented a positive growth experience. For this, I am thankful to a number of people that welcomed me in their lives. First, thank you to the friends and lab mates in the Castoe Lab, older and newer – Daren Card, Drew Schield, Rich Adams, Blair Perry, Nicky Hales, Andrew Corbin, Aundrea Westfall, Ricky Orton, Zack Nikolakis, in no specific order because each and every one of you contributed to who I am today. You guys have been an inspiration, a joy to work with, and unvaluable companions in the field, at parties, and at conferences. Thank you to Drs. Robert Ruggiero, Steve Mackessy, and Mike Vandewege. Thank you to all of the friends and colleagues in the UTA biology department, especially to Danielle Rivera, Shannon Beston, Kathleen Curie, TJ Firneno, and Will Budnick. A particular mention to Dr. Corey Roelke, for the blunt honesty and reminding me that it is not worth to take myslef so seriously. Thank you to my committee members, Drs. Esther Betran, Jeff Demuth, Matt Fujita, and Matt Walsh, for your support, knowledge, and ideas for improving my research. And thank you to the peers in your labs for sharing ideas and providing unique perspectives. A huge thank you to the biology department staff – to Rachel Wostl, Linda Taylor, Ashley Priest, and Mal Roelke, thank you for putting up with me! I am especially grateful for the unconditionalu and unceasing support of my family (dad, mom, grandmas, cousins and aunties/uncles), Micol and her family (Amelia and Giu), Marta Gazzana, Elisabetta Villa, Sandro Pratesi and the Pratesi family, Alexandra Farber (and the TBT, for reminding me that there is still so much beauty), my dance family (without whom I could have not burned out the crazy), Mike (again), and Jill Castoe. Lastly, thank you to my advisor, Todd Castoe, for being an outstanding mentor, for the patience, and for teaching me to always check before sitting down – there is always the chance of a pillow case with a deadly snake inside.

November 22nd, 2019

Dedication

This dissertation is dedicated to all my loved ones. Especially to my parents, Cristina Vescia and Carlo Pasquesi. For showing me the beauty and the diversity of the World, allowing me to make my decisions, knowing that I would always have your back. And for supporting me, realistically and unconditionally. And to Micol. For understanding that I needed to go; there are no words to describe what we share, if not that you are family that I chose. And to Mike and Alexandra, without whom I do not know if I would have made it so far.

Abstract

LEVERAGING THE DYNAMIC REPEAT ELEMENT LANDSCAPE OF SQUAMATE REPTILE GENOMES TO UNDERSTAND BROAD PATTERNS OF VERTEBRATE GENOME EVOLUTION AND TRANSPOSABLE ELEMENT BIOLOGY

Giulia I. M. Pasquesi, PhD

The University of Texas at Arlington, 2019

Supervising Professor: Todd A. Castoe

Vertebrate genomes are mostly composed of transposable elements (TE), mobile DNA sequences that have shaped genome structure and evolution by promoting positive (e.g., regulatory network rewiring, embryo development) and negative (e.g., ectopic recombination, disease) genomic processes. Leveraging genomic and transcriptomic data from diverse vertebrate species, I present novel lines of evidence that underscore the unique value of squamate reptile genomes for investigating properties of TE landscape evolution. This dissertation demonstrates that squamate genomes defy paradigms of amniote repeat element evolution set by mammals and birds, in particular: that greater variability in TE content is found between major lineages, that genome size correlates to genomic TE content, and that effective population size relates to features of the TE landscape (i.e., full-length insertions and TE abundance). Squamates are also unique among amniotes for having a broad diversity of TE types and families that appear similarly prevalent in the genome and simultaneously active, whereas patterns of negative regulation of TEs in germline tissues are consistent with those of other vertebrate species (with the exception of mammals). The detailed investigation of the prairie rattlesnake genome further shows that TEs have been involved in sex chromosome evolution, gene duplication and isochore structure, demonstrating that the distinct evolutionary dynamics of squamate TE landscapes may be linked to more unique and variable aspects of squamate genome function and evolution compared to other amniote species.

Table of Contents

Chapter 1

Introduction

Most of the DNA in vertebrate genomes is composed of repetitive sequences, largely transposable elements (TEs) but also simple repeats (i.e., microsatellites). Most TEs within a genome represent relics of past activity, and only a considerably smaller fraction still retain the capabality to mobilize and spread into the host genome by intra-genomic copy number amplification and/or inter-genomic horizontal transfer. In the aftermath of their re-discovery in *Drosophila* hybrid-dysgenesis systems in the late '70s, TEs were considered little more than "selfish elements"; genomic translocation and accumulation of TEs can lead to an increase genome size, and negatively affect the fitness of the host directly through insertional mutagenesis in functional regions and indirectly by favoring ectopic recombination and genomic rearrangements. The advent of the age of genomics, however, forced researchers to re-evaluate the presence of TEs in a genome. The past two decades of TE research have brought additional support for the insertional mutagenic role played by TEs, and the formulation of a general model that translocation is recovered when mechanisms of negative TE regulation become ineffective, leading to both inheritable and non-inheritable diseases. However, TEs have also been re-appreciated as pivotal agents of narrow and large-scale genome evolution. By inserting in the proximity of a gene, TEs can modulate its expression, favoring the evolution of diverse cell types. Addiotionally, in virtue of their repetitive nature, TEs have been coopted as binding sites for transcription factors – allowing for the rewiring of regulatory networks and the emergence of novel phenotypes.

It is undeniable that thanks to massive technique and computational advancemens we are achieving a finer-scale knowledge of the complex relationships between host genomes and the TEs they harbor. Yet a noticeable trend in genomic research is an increase in focus, which brings remarkable sampling biases in terms of lineages, organisms, and systems (i.e., cell population or type of cancer) analyzed, while more integrated comparative perspectives are mostly left unevaluated. As a result, current paradigms of TE biology and its interplay with the host genome are mostly derived from in-depth studies on few model organisms: mammals in particular, and to some extent birds and fishes among vertebrates. Snakes and squamate reptiles have been greatly neglected in the genomic era. Yet from the little that was known, it emerged that squamate genomes represent a valuable system to study genome evolution, from sex chromosomes to the emergence of novel complex phenotypes (e.g., venom), isochore structure and recombination hotspots. As for repeat element landscapes, squamates appeared to challenge prevalent models set by mammals and birds, in particular the proposed paradigms of little within-clade variability in TE content and composition, and the correlation of genome size and TE content.

The main focus of my dissertation has been to evaluate whether widely accepted broad models of TE and host genome evolution set by intensive studies focused on mammals and birds still apply when a diverse group (squamates) is included. Specifically, I tested whether effective population size, which relates to how effective purifying selection is at removing mildly deleterious mutations, can explain the variance in features of the TE ladscape across species (Chapter 2), and analyzed whether differences exist in how TEs are regulated and expressed across somatic and germline vertebrate tissues, and if generalizable patterns can be detected (Chapter 4). By participating in the assembly and annotation of the first chromosome-level genome for a snake, the prairie rattlesnake, I was also able to evaluate chromosomal distribution patterns of repeat

elements, in particular on the Z sex chromosome, and corroborate previous hypotheses of isochore structure re-acquisition in snakes (Chapter 3).

Ultimately, this dissertation provides novel characterizations of the TE content and composition, TE regulatory mechanisms, and TE expression at broad phylogenetic scales, underscoring the potential shortcomings of broad assumptios that diverse vertebrate model systems share common biological features and evolutionary dynamics. Hopefully, it will also represent an additional reference in support of the tremendous potential that squamate reptiles, and snakes in particular, hold as a system for studying the impact of TEs in genome biology, function and structure.

Chapter 2

Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals

Giulia I.M. Pasquesi[1], Richard H. Adams[1], Daren C. Card[1], Drew R. Schield[1], Andrew B. Corbin[1], Blair W. Perry[1], Jacobo Reyes-Velasco[1,2], Robert P. Ruggiero[2], Michael W. Vandewege[3], Jonathan A. Shortt[4] and Todd A. Castoe[1, ‡]

[1]Department of Biology, 501 S. Nedderman Dr., University of Texas at Arlington, Arlington, TX 76019 USA

[2]Department of Biology, New York University Abu Dhabi, Saadiyat Island, United Arab Emirates

[3] Department of Biology, Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, 19122 USA

[4] Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045 USA

Abstract

Broad paradigms of vertebrate genomic repeat element evolution have been largely shaped by analyses of mammalian and avian genomes. Here, based on analyses of genomes sequenced from over 60 squamate reptiles (lizards and snakes), we show that patterns of genomic repeat landscape evolution in squamates challenge such paradigms. Despite low variance in genome size, squamate genomes exhibit surprisingly high variation among species in abundance (ca. 25% - 73% of the genome) and composition of identifiable repeat elements. We also demonstrate that snake genomes have experienced microsatellite seeding by transposable elements at a scale unparalleled among eukaryotes, leading to some snake genomes containing the highest microsatellite content of any known eukaryote. Our analyses of transposable element evolution across squamates also suggest that lineage-specific variation in mechanisms of transposable element activity and silencing, rather than variation in species-specific demography, may play a dominant role in driving variation in repeat element landscapes across squamate phylogeny.

Introduction

Transposable elements (TEs) and other repetitive sequences represent a major fraction of vertebrate genomes – in most mammals, repeat elements comprise 28-58% of the genome (Smit et al. 2015-2019; Platt et al. 2018), and may comprise more than two thirds of the human genome (de Koning et al. 2011). Several decades of genome research has led to the prevailing view that genome size and genome repeat content are tightly linked, such that shifts in genomic repeat content are expected to result in proportional shifts in vertebrate genome sizes (Chalopin et al. 2015; Elliott and Gregory 2015; Canapa et al. 2016). Recently, this correlation has come into question in favor of alternative hypotheses, such as the "accordion" model of co-variation between genomic DNA gained by repeat element expansion and genomic DNA lost through deletion (Kapusta et al. 2017). It has also been demonstrated that the relationship between genome size and repeat content may vary between vertebrate lineages (Agren and Wright 2011; Elliott and Gregory 2015; Canapa et al. 2016), with some lineages adhering more or less to a particular model or pattern (Blass et al. 2012; Chalopin et al. 2015; Elliott and Gregory 2015; Kapusta et al. 2017), underscoring the value of comparative analyses across diverse lineages.

Within vertebrates, our understanding of genome and repeat element evolution is largely biased towards mammals and archosaurian reptiles (mainly birds). The emerging pattern from studies of these groups is that large differences in the repeat element landscape exist among major amniote vertebrate lineages, yet fairly little variation in repeat content and diversity are observed within major amniote groups. For example, estimates based on *de novo* annotation of TEs in mammal and bird species suggest 1.7-fold and 2.2-fold variation in TE content across species for each group, respectively (Smit et al. 2015-2019; Kapusta et al. 2017). Although squamate reptiles

(lizards and snakes) represent a major portion of the amniote tree with over 10,000 species spanning more than 200 million years of evolution (Murphy et al. 2007), variation in genomic repeat content across squamate reptiles has remained poorly studied. From the few studies to date, genome size appears to be highly conserved in squamate reptiles (Gregory 2017), yet the little that we know about repeat element variation suggests that squamate reptile genomes vary greatly in repeat element content (Castoe et al. 2011; Castoe et al. 2013).

Motivated to assess whether squamate reptile genomic repeat element landscapes adhere to patterns observed in birds and mammals, we analyzed genomic repeat landscapes across 66 squamate species using low-coverage random whole genome shotgun sample sequencing data (Castoe et al. 2011; Castoe et al. 2013) and draft genome assemblies. We find that squamate reptile genomes indeed challenge the paradigm that genome size and repeat content are tightly linked, and the view that major differences in repeat element content occur only between lineages of amniotes. In addition to contributions from TEs, snake genome repeat content variation is further increased by the largest known instance of microsatellite seeding by long interspersed nuclear elements (LINEs) observed in any living organism. We also find evidence that multiple independent horizontal transfer events and highly idiosyncratic patterns of transposable element proliferation across squamates have further contributed to extreme variation in genome repeat content in this lineage. We further tested a demographic explanation for variation in repeat content, whereby fluctuations in the effective population size (N_e) of species impact the efficacy of selection against repetitive element insertion (Lynch and Conery 2003). We find no evidence that N_e explains the distribution and variation in characteristics of the repeat landscape in squamate reptiles, which indicates instead that variation in molecular mechanisms of TE proliferation, silencing, removal and truncation may underlie the extreme

repeat variation observed across squamates. Collectively, our findings challenge existing views related to repeat element and genome size co-evolution, and provide new evidence for unappreciated variation in genomic repeat content within and among major amniote lineages.

Results

Comparison of sampled and assembled genome data

Our analyses of genomic repeat content were based on the assemblies of 12 squamate genomes (including 1 new and 11 published assemblies), and low-coverage, unassembled genomic shotgun read datasets obtained from 54 squamate species (Supplementary Data 1; Castoe et al. 2013). Previous studies have shown that genomic repeat content estimated from unassembled shotgun genomic datasets are similar to estimates derived from assembled genomes (Castoe et al. 2011; Castoe et al. 2013). We confirmed this by comparing repeat annotations from assembled and unassembled genome data from the same species (Supplementary Fig. 1), and also confirmed that repeat estimates derived from unassembled genomic shotgun datasets are effectively independent of the amount of sequence data obtained (Supplementary Fig. 1).

Genome size and repeat content in major amniote groups

Squamate reptile genomes challenge the commonly accepted paradigm that genome size and repeat content are tightly linked (Chalopin et al. 2015; Elliott and Gregory 2015; Canapa et al. 2016), and also challenge the prevailing view that large variation in repeat content tends to be characteristic of major clades, rather than highly dynamic within clades (Smit et al. 2015-2019; Fig. 1). For example, mammalian genome sizes tend to be more highly variable (2.2-6.0 Gbp (Gregory 2017); Supplementary Data 2) in comparison with squamate and bird genomes, yet genomic TE estimates demonstrate only moderate levels of clade-specific variation (33.4-56.3%,

mean = 44.5%; Fig. 1a, Supplementary Data 3 and Supplementary Note 1). In contrast, birds have smaller genomes and higher conservation of genome sizes (1.0-2.1 Gbp (Gregory 2017); Supplementary Data 2), with relatively low levels of TE content (4.6-10.4%, mean = 7.8%, with the only notable exception being the downy woodpecker with an extremely high genomic TE content of 22.5%, which we excluded as an outlier from analyses here; Fig. 1b, Supplementary Data 3 and Supplementary Note 1).

With highly conserved genome sizes (1.3-2.8 Gbp) yet extensive variation in genomic content of readily detectable TEs (23.7-56.3%, mean = 41.8%; Fig. 1c), we find that squamate reptiles do not adhere to either of these trends. The relatively high degree of variation in genomic repeat content across remarkably short evolutionary time scales in squamates presents the greatest contrast with birds and mammals. Unlike the clade-specific pattern observed in mammals, the genomic repeat content variation of squamate reptiles exhibits a high degree of variation even between species within the same genus (e.g., within the genera *Ophisaurus* (44.8 – 48.9%), *Coniophanes* (59.4 – 73%), and *Crotalus* (35.3 – 47.3%); Fig. 1c, Supplementary Figs. 2 and 3, Supplementary Data 4). Across the 66 squamate species sampled, total genomic repeat element content varied from 24.4% to 73.0% (3-fold variation; Fig. 1c). Collectively, our analyses highlight the remarkable finding that the comparatively small genomes of squamates, similar to those of birds, can contain large and highly variable amounts of repeat elements, exceeding the range reported for mammals.

Genomic TE composition across squamate reptiles

The content and evolutionary dynamics of TEs in squamate genomes are unique in many ways when compared to that of mammals and birds, yet squamate genomes also share several key

features with both lineages. All three groups have TE landscapes largely dominated by non-long-terminal repeat (non-LTR) retrotransposons. However, unlike mammalian genomes in which L1 LINEs and associated short interspersed nuclear elements (SINEs) are the most dominant and active elements (de Koning et al. 2011; Huang et al. 2012), squamate genomes tend to contain three similarly-abundant and active LINE families (CR1, BovB, and L2 LINEs; Fig. 1, Supplementary Fig. 2, and Supplementary Data 4). While CR1 LINEs are ubiquitous across amniote genomes, CR1s are particularly abundant and recently active in squamate genomes (5.1%, compared to ~3.5% in birds and <1% in mammals; Smit et al. 2015-2019), as they tend to be in other non-avian reptiles (i.e., ~10% in crocodilians; Suh et al. 2015). In addition to non-LTR elements, DNA elements are also highly variable and particularly abundant in multiple divergent squamate lineages (Fig. 1). For example, Tc1-Mariner elements have experienced a 2.4 fold expansion in colubroid snakes compared to lizards (mean genomic abundance = 4.23% in colubroid snakes and 1.7% in lizards; Fig. 1, Supplementary Fig. 2 and Supplementary Data 4). The most striking contrast between squamate versus bird and mammal genomes is that squamate genomes contain an unusually broad diversity of types, subtypes, and families of TEs that appear simultaneously active (Alfoldi et al. 2011; Castoe et al. 2011; Tollis and Boissinot 2013; Suh et al. 2015; Yin et al. 2016; see also below, Fig. 4 and Supplementary Fig. 3), whereas genomes of mammals and birds tend to have a very small number of active elements (e.g., L1 LINEs and Alu SINEs in mammals, and endogenous retroviruses (ERVs) in birds; Brouha et al. 2003; Huang et al. 2012; Zhang et al. 2014; Chalopin et al. 2015).

GC content is known to play an important role in genome and repeat element evolution (Boissinot et al. 2001; Rizzon et al. 2002; Fryxell and Moon 2005; Hellen and Brookfield 2013; Georges et al. 2015). We found evidence of significant relationships between GC content and

total TE content, as well as GC and total microsatellite (or simple sequence repeat; SSR) content, in lizards and colubroid snakes (Supplementary Fig. 4). In contrast, we found no correlation between genomic GC content and any aspect of the genomic repeat element landscape in non-colubroid snake genomes (Supplementary Fig. 4). Consistent with previous studies (Castoe et al. 2013), our analyses highlight the surprisingly variable nature of GC content across squamate genomes, which tends to be higher in lizards than in snakes, yet highest in the colubroid snake *Coniophanes fissidens* (GC= 47.8%; Fig. 1c). These findings are also broadly consistent with previously reported shifts in GC isochore structure in squamate genomes (Alfoldi et al. 2011; Georges et al. 2015), including the absence of isochore structure in lizard species, and intermediate structure in snakes that appears to represent isochore reacquisition after isochore loss in a squamate ancestor (Castoe et al. 2013).

Unparalleled microsatellite abundance in squamate genomes

Our analyses revealed that some squamate genomes contain astonishingly high levels of SSRs, and that genomic SSR content in some snake species is the highest of any previously studied vertebrate (e.g., 14% according to RepeatMasker estimates in *Coniophanes fissidens*; Supplementary Data 4 and 5; Supplementary Fig. 5). While previous studies have suggested that the highest variation in SSR content tends to exist among major vertebrate lineages (Neff and Gross 2001), with fish, squamate reptiles, and mammalian genomes having similarly high genomic content (Alfoldi et al. 2011; Castoe et al. 2011; Castoe et al. 2013; Adams et al. 2016), our results provide new evidence that the highest variation known in genomic SSR content exists within lineages – squamates and snakes, specifically. We found up to 10.9-fold variation in the genomic density of SSR loci (262-2,845 loci/Mbp) and 16.6-fold variation in SSR-occupied bases per Mbp (4.08-67.94 Kbp/Mbp) among squamates overall, with non-colubroid snakes

tending to have the lowest genomic SSR abundance, and colubroid snakes having the highest (Supplementary Data 5; Fig. 2 and Supplementary Figs 5, 6). This extreme variation in the genomic SSR content of squamate reptiles exceeds the previous high benchmark set by fish genomes (8.2-fold loci/Mbp and 18.0-fold bp/Mbp variation), and dwarfs that of mammals (5.8-fold loci/Mbp and 5.4 bp/Mbp) and bird genomes (1.8-fold loci/Mbp and 2.8 bp/Mbp; Alfoldi et al. 2011; Castoe et al. 2011; Castoe et al. 2013; Adams et al. 2016).

Largest instance of microsatellite seeding among vertebrates

A peculiar feature of SSR evolutionary dynamics in squamate genomes is the significant shifts in 4mer and 5mer abundances across the squamate tree, including extreme expansion of specific 4mer and 5mer SSRs motifs in colubroid snake genomes (Kruskal-Wallis test p-value <0.001, Supplementary Fig. 6 and Supplementary Data 6). Two specific SSR sequence motifs, ATAG and AATAG, account for most of the microsatellite expansion in colubroid snakes, representing a 7.4-fold increase in ATAG (bp/Mbp) and an 87.7-fold increase in AATAG (bp/Mbp) compared to the averages of other squamate genomes (Supplementary Figs 7, 8). The extremely high genomic representation of these two similar SSR sequence motifs in snake genomes suggests a motif-specific mechanism has driven their expansion. Previous studies (Castoe et al. 2011; Castoe et al. 2013) have suggested that LINE retrotransposons that contain microsatellites on their 3' end in snakes might lead to SSR genomic expansion through a process called "microsatellite seeding".

To test the hypothesis that microsatellite seeding is responsible for the expansion of particular SSR sequence motifs, we surveyed the regions adjacent to the two most highly expanded SSR motifs (AATAG and ATAG) in 8 complete reptile genome assemblies. Consistent with the

expectations of microsatellite seeding, we found strong statistical support that CR1-L3 LINEs tend to be immediately adjacent to AATAG loci in colubroid genomes (Fisher's exact test p-value < $2.2e^{-16}$), as well as strong statistical enrichment of AATAG loci at the 3' end tail of Rex LINEs (p-value < $2.2e^{-16}$) in all squamate genomes sampled, suggesting that both CR1/CR1-L3 and Rex LINEs contribute to microsatellite seeding in squamate genomes (Fig.2b; Supplementary Data 7). In contrast to elements adjacent to AATAG repeats, we found no evidence of enrichment in adjacency for any particular TE for the second most expanded SSR motif (ATAG) compared to randomly sampled genomic regions; this suggests that the expansion of this motif is not directly driven by microsatellite seeding, although its similarity to AATAG suggests it might be indirectly related. To further identify the specific LINE element that is responsible for microsatellite seeding of AATAG SSR loci, we calculated the conditional probability of TE-SSR co-occurrence in a genome-wide context compared to the AATAG-adjacent context. Conditional probabilities of AATAG loci and CR1-like LINEs genomic co-occurrence are noticeably different only for CR1-L3 LINEs between colubroid snakes and other squamates (Fig. 2c), and are only barely detectable for Rex LINEs. Additionally, CR1 LINEs are a major contributor to the genomic TE landscape of squamates (particularly colubroid snakes), whereas Rex elements represent a very small fraction. Taken together, our data indicate that microsatellite seeding may be a common ancestral feature of multiple families of squamate LINEs, yet the high activity and expansion of CR1-L3 LINEs has driven associated AATAG loci to extremely high frequencies in colubroid snakes, leading to an astounding 74.73-fold genomic AATAG loci/Mbp increase in this lineage, and the highest levels of genomic SSR content among vertebrates. The ramifications of such extreme levels of homologous SSRs in colubroid snakes, in terms of genome function and evolution, remains uninvestigated. A potential role in mediating

increased ectopic recombination leading to gene duplication has been suggested by previous studies that have identified an enrichment of these repeats surrounding tandemly duplicated venom genes in snakes (Ikeda et al. 2010; Castoe et al. 2011; Dowell et al. 2016). Collectively, these findings imply the exciting possibility that LINE-SSR hybrid elements may have played key roles in the evolution of prominent phenotypes in snakes (i.e., venom evolution).

Multiple independent TE horizontal transfer events

Evidence for the horizontal transfer of BovB LINEs has been identified by previous studies (Kordis and Gubensek 1997, 1998; Kordis and Gubenšek 1998; Castoe et al. 2011; Walsh et al. 2013), and our analysis of squamate genomes provides new insight into the complexities of BovB horizontal transfer. Our phylogenetic reconstruction of BovB LINEs, including samples from our squamate genomes and other sequences from Genbank (Clark et al. 2016), highlights multiple horizontal transfer events, and supports ectoparasite-mediated transfers of BovB LINEs into and out of squamate reptile genomes (Fig. 3 and Supplementary Fig. 9a, Supplementary Data 9). We found BovB LINE sequences from squamate species clustering with other groups of metazoans in all branches of our phylogenetic tree, consistent with multiple horizontal transfer events of BovB from lizards to mammals and to other squamates, and from snakes to mammals and other squamates. Previous studies found support for virus-mediated transfer of TEs (Piskurek and Okada 2007), and suggested ectoparasites as potential transmission vectors (Silva et al. 2004; Gilbert et al. 2010; Novick et al. 2010; Walsh et al. 2013; Gilbert et al. 2014). Our analyses support the horizontal transfer of BovB from one reptile tick species (*Amblyomma limbatum*) to colubroid snakes (Supplementary Fig. 9a), and provide the first ever evidence for ectoparasite-mediated transfer from squamate genomes in the case of the reptile tick *Bothriocroton hydrosauri*. Samples containing BovB elements sequenced from this tick species

are deeply nested among lizard-derived BovB sequences, yet are unique in containing a large internal deletion (1691nt) relative to all other lizard-derived BovB sequences in this clade. Collectively, our analyses of BovB LINE evolution showcase a dynamic history of horizontal transfer that encompasses essentially all forms of the process of transfer into and out of squamate genomes, implicating the role of ectoparasites in both directions of the transfer process.

Testing explanations of variation in genomic TE abundance

Multiple studies have suggested that purifying selection acting against TE insertions may manifest in correlations between effective population size (N_e) and features of the genomic TE landscape. This prevailing demographic explanation for variation in repeat content has been invoked to describe patterns of genome complexity and evolution across the tree of life, and predicts that lineages with higher N_e should undergo more effective purifying selection and thus lower genomic accumulation of mutationally hazardous DNA (Lynch and Walsh 2007; Charlesworth 2009). Indeed, previous population (within-species) and phylogenetic (among species) studies have provided rationale and empirical evidence that transposable element insertion rates, fixation rates, and abundance may be correlated with effective population size (Lynch and Conery 2003; Petrov et al. 2003; Le Rouzic et al. 2007; Lynch and Walsh 2007; Blumenstiel et al. 2014). Relative insert length has also been linked to population size at the population-level by an ectopic recombination model in which element length is correlated with the strength of selection (Lynch and Conery 2003; Petrov et al. 2003; Song and Boissinot 2007; Petrov et al. 2011; Tollis and Boissinot 2013; Barron et al. 2014).

Using our phylogenetic-scale dataset, we tested if features of TE landscapes (i.e., genomic abundance, estimated age of activity, and degree truncation for BovB and CR1-L3 LINEs)

showed evidence of a correlation with estimates of effective population size consistent with a demographic model of TE landscape evolution. We first tested for a relationship between N_e and TE landscape characteristics using the median values of N_e estimates derived from PSMC analyses (Li and Durbin 2011) for 8 published squamate genomes (Fig. 4b-d, Supplementary Fig. 10). With this dataset, we found no evidence supporting a correlation between N_e and CR1-L3 and BovB length or genomic repeat element abundance (Fig. 4c-d, Supplementary Fig. 10c-e). Notably, we found that species with similar N_e estimates (Fig. 4b) showed different levels of truncation and of TE genomic abundance, and that even within a species TE truncation and abundance were poorly correlated (Fig. 4a, c-d; Supplementary Fig. 10 and 11). Second, to further test for correlations between N_e and element abundance or truncation using an approach that is independent of inferences of generation time and mutation rates, and independent of potential biases associated with coalescence-based estimates of N_e (i.e., population substructure, migration, selection; Nielsen and Beaumont 2009; Li and Durbin 2011; Mazet et al. 2016; Nadachowska-Brzyska et al. 2016; Orozco-TerWengel 2016; Schrider et al. 2016; Adams et al. 2018), we used adult body mass (Feldman et al. 2016) as a proxy for N_e for all species included in our study (as in Figuet et al. (2016); Supplementary Data 8). This approach has the added benefit of leveraging the much larger sample size of our entire dataset (compared to our PSMC analyses using 8 complete genomes). Similar to our PSMC-based analyses, we compared body mass to CR1-L3 and BovB genomic abundance, their degree of truncation, and total genomic repeat element and TE abundances. Consistent with our PSMC-based analyses, we failed to find a correlation between body mass and truncation (Fig. 4e and Supplementary Fig. 12b) that would support a demographic model of TE landscape evolution; the only correlative trend that we did find was a correlative trend that is opposite of that predicted by the demographic model between

N_e and genomic repeat element abundance instead (i.e., higher N_e was positively correlated with TE abundance; Supplementary Fig. 12d). Finally, to test more generally for evidence that selection acts on TE length at the phylogenetic scale, we tested for a link between TE truncation and TE age (Neafsey et al. 2004; Tollis and Boissinot 2013; Barron et al. 2014) using median pairwise divergence of TE copies from their subfamily consensus, π, as a proxy for age for CR1-L3 and BovB families, and found no correlation (Fig. 4f, Supplementary Fig. 13 and detailed in Supplementary Figs 14-16). While we acknowledge the complexity of testing links between two highly dynamic evolutionary processes (e.g., N_e and TE abundance), and the limitations of methods used to make inferences about these processes (i.e., N_e estimation), all of our analyses fail to provide support for N_e as a strong determinant of variation in the composition and characteristics of the repeat element landscape at the phylogenetic level across squamate reptiles. Although our analyses cannot fully reject a demographic hypothesis that a relationship between N_e and TE characteristics exists (i.e., we can only fail to reject a lack of relationship), the apparently poor explanatory power of the demographic hypothesis in predicting squamate TE activity and abundance suggests that perhaps other factors, such as variation in molecular mechanisms of TE proliferation, silencing, and removal, may better explain the majority of variation in TE abundance at the phylogenetic level in squamates.

Discussion

This broad glimpse into the diversity of repeat structure and composition of squamate reptile genomes suggests that this lineage possesses particularly distinct and often extreme repeat landscape characteristics compared to other amniotes. Our results provide evidence for surprisingly high variation in the content and composition of genomic repeat elements across

squamate lineages, including 3-fold variation in the identifiable genomic repeat element content. We also discovered that some snake genomes have experienced microsatellite expansion at unprecedented scales through the process of microsatellite-seeding by specific LINEs, leading to genomic microsatellite abundances that are the highest of any known vertebrate genome. Despite such extreme variation in genomic repeat element content, genome size across squamates is remarkably conserved (~0.2-fold variation), challenging the prevailing view that genomic repeat abundance and genome size tend to tightly co-evolve (Elliott and Gregory 2015). These findings provide some of the strongest evidence for a dynamic equilibrium or an "accordion" model, in which genomic DNA gain through TE expansion may be approximately balanced by genomic DNA loss through deletion (Petrov 2002; Neafsey et al. 2004; Kapusta et al. 2017). Overall, these results highlight extreme shifts in the structure of squamate reptile genomes, and further beg the question of whether particular aspects of squamate genome function and evolution are also more unique and variable compared to other vertebrates. These findings argue that squamates may represent a particularly powerful model system for testing hypotheses about genome structure, function, and evolution, and their interactions.

Many previous studies focused on population-level dynamics of TE evolution have shown that differences in N_e and the efficacy of purifying selection acting against TE proliferation has played a major role in structuring the repeat landscape of many eukaryote genomes (Charlesworth et al. 1994; Neafsey et al. 2004; Song and Boissinot 2007; Petrov et al. 2011; Blass et al. 2012; Le Rouzic et al. 2013; Tollis and Boissinot 2013; Barron et al. 2014; Ruggiero et al. 2017; Xue et al. 2018). Even in squamate species (e.g., *Anolis* lizards), variation in effective population sizes has been linked to TE insertion length and fixation probability (Tollis and Boissinot 2013; Ruggiero et al. 2017; Xue et al. 2018). Our phylogenetic-scale analyses

across squamate species, however, recovered no clear evidence linking genomic repeat abundance or activity with N_e estimates in squamates. Although coalescent-based estimates of N_e can be biased by a number of model violations (i.e., population substructure, selection), we also failed to find a significant relationship between genomic repeat characteristics and body mass – a known correlate of N_e. Population size is, however, likely to have influenced other aspects of genome evolution, such as fixation of deletions, that could contribute to the maintenance of nearly constant genome size in squamates.

Our results together with those from previous studies suggest that different evolutionary forces may dominate different evolutionary scales, and that while demographic processes (and purifying selection) may dominate population-level trends in TE evolution, phylogenetic-scale patterns in TE landscapes may be more strongly determined by other processes. Evidence for extreme variation in transcriptional levels of TE-derived transcripts across squamates (Castoe et al. 2011), together with evidence from this study of lineage-specific swings in repeat element proliferation, suggest that molecular mechanisms related to TE regulation may be particularly relevant at the phylogenetic-scale in squamates. Squamates may, therefore, represent a valuable system for studying the impacts of variation in molecular mechanisms of TE control, such PIWI-Interacting RNA (piRNA) dynamics and efficacy, epigenetic silencing of TEs, lineage-specific TE activity, DNA repair mechanisms, and post-insertion 5' removal of TEs. Further studies are needed to address the question of whether variation in molecular mechanisms of TE silencing and activity, as well as DNA repair, explain variation in squamate genomic TE content, and would provide fascinating insight into the factors that shape genomic repeat landscape variation.

Methods

Taxon sampling and library preparation

DNA extraction of 52 squamate samples (total = 45 species) was performed using a Phenol-Chloroform-Isoamyl alcohol (PCI) extraction protocol. Random shotgun genome libraries were prepared by fragmenting DNA samples to an average length of 300-600bp using a M220 Covaris Ultrasonicator. The NEBNext Illumina DNA Library Prep Kit (New England Biolabs) was used following the manufacturer's protocol to perform fragment-end repair, poly-A tailing, adapter ligation, and library amplification. After library preparation, fragments were size-selected using a BluePippin (Sage Science) for a length of 350-450bp. Pooled multiplexed libraries were sequenced on an Illumina MiSeq with 300bp paired-end reads. Paired-reads were merged based on sequence overlap and were adapter and quality trimmed using CLC genomics workbench v.9.0.1. 454 shotgun sequencing data of 9 snake species from previous studies (Castoe et al. 2011; Castoe et al. 2013) and draft genome assemblies of 12 additional squamate species (Supplementary Data 1) were also included. Our final sampling included a total of 66 different squamate species.

For each species, mitochondrial reads were filtered out in CLC genomics workbench 9.0.1 using the complete mitochondrial genome of the most closely related species available on Genbank (Clark et al. 2016). Reads that mapped to the reference were used to assemble species-specific mitochondrial genomes. Reads that did not map to the reference (i.e., nuclear reads) were used for downstream repeat element annotation and analyses.

Simple sequence repeat (SSR) identification and analysis

We used Pal_finder v0.02.03 (Castoe et al. 2010) (Palfinder hereafter) to identify microsatellites. Default Parfinder parameters were used to identify perfect dinucleotide (2mer), trinucleotide (3mer) and tetranucleotide (4mer) that were tandemly repeated for a total length of at least 12bp. Perfect pentanucleotide (5mer) and hexanucleotide (6mer) tandemly repeated motifs were annotated only if longer than 15bp. Loci/Mbp and bp/Mbp frequencies were calculated for all microsatellite motifs, length classes (2-6mers), and total content, and summarized per genome and major taxonomic group. Tests for multiple evolutionary rates of microsatellite abundance across lineages, ancestral state reconstruction of genomic microsatellite frequencies, and quantification of microsatellite landscape differentiation among species were performed using the R packages Phytools v.0.4-60 (Revell 2012) and APE v.3.3 (Paradis et al. 2004). For the multiple evolutionary rate analysis of microsatellite (and TE) abundance, we conducted censored rate tests using Phytools with 1000 simulations (to compute p-values) on 100 randomly sampled posterior trees using the restricted maximum likelihood technique (REML) to obtain unbiased estimates of the evolutionary rate parameter (σ) (Adams et al. 2016). We used the time-calibrated phylogeny and the *pic* function in R (provided by the APE package) to compute phylogenetic independent contrasts for tests of clade-specific differences in genomic microsatellite content. We performed the nonparametric Kruskal-Wallis H test in R after the data rejected normality (Shapiro-Wilks test; p-values < 0.05 before and after log transformation) and homogeneity of variances (Bartlett's test; p-values < 0.05 before and after log transformation). Between lineages variation was tested using a posthoc Dunn test for multiple comparisons using the Benjamini-Hochberg correction method in R (Supplementary Data 6).

Transposable element identification and analysis

Squamate genomic repeat elements were annotated according to homology-based and *de novo* identification approaches. Because repeat element annotation can be highly dependent on the repeat library used, we built large multi-species (clade-specific) repeat libraries that we used to annotate repeats for all members of a clade. To build these clade-specific libraries, we first performed *de novo* repeat element annotation on each species (except where already published) using RepeatModeler v.1.0.9 (Smit and Hubley 2008-2017), followed by further repeat classification in CENSOR (Kohany et al. 2006). Second, we built clade-specific *de novo* repeat element libraries, one for all lizard species (33 species *de novo* reference library) and one for all snake species (*de novo* transposable element libraries for 21 species were combined, and merged with the reference library generated by Castoe et al. (2013). Each clade-specific library was then filtered to avoid redundancy of highly similar elements. We tested whether using a single squamate-specific library for all species would change the inferred relative TE content and overall amount of repeat identified; we found no detectable difference between the results of the two masking protocols (Supplementary Fig. 17), and therefore decided to use the two clade-specific libraries in order to reduce masking time by reducing the overall library size. Additional classification of unknown (unclassified) elements was achieved by comparing these unclassified elements to all elements that were classified using BLAST (Johnson et al. 2008). Additionally, we generated squamate-specific BovB and CR1-L3 LINEs reference sequence libraries for all 66 species included (additional information regarding library generation are provided in the following paragraph).

Repeat element analyses were performed in RepeatMasker v.4.0.6 (Smit et al. 2015-2019) with default parameter settings. To maximize element identification, we used a custom bash script to

specify the order of the four libraries used as references for the masking process: (i) BovB-L3 LINEs library, (ii) Tetrapoda *RepBase* library (version 20.11, 07 August 2015; Bao et al. 2015), (iii) classified elements from the clade specific library for either snakes or lizards, and (iv) unknown elements from the clade specific library. We used the BovB-L3 LINEs library first to control for limited sampling and low quality reference sequences of squamate reptile BovB and L3 LINEs in the tetrapoda library. RepeatMasker output files were post-processed using a custom-modified implementation of the *ProcessRepeat* script included in the RepeatMasker package. Specifically, we modified the output to include additional summary information in the *.tab* output file for TE subfamilies that are important and/or frequent in squamate reptiles (e.g., CR1-L3, L2 and Rex). Also, because the provided *ProcessRepeat* script still reflects old and outdated classification schemes of TEs (e.g., Penelope elements are inappropriately classified as LINEs), we made other modifications to the *ProcessRepeat* script to correct for such errors according to the classification reported in Chalopin et al. (2015).

Comparing sampled and assembled genomes

We tested whether genomic repeat content estimated from unassembled shotgun genomic datasets were similar to estimates derived from fully-assembled genomes. We compared RepeatMasker estimates of total TE genomic abundance between assembled genomes and unassembled shotgun genomic datasets for the same species (*Python molurus*, *Boa constrictor*, *Thamnophis sirtalis*, and *Deinagkistrodon acutus*) or for two closely related species belonging to the same genus (*Gekko gecko* vs. *Gekko japonicus* and *Ophisaurus attenuatus* vs. *Ophisaurus gracilis*). We also tested for potential biases due to unequal genomic sampling in the shotgun datasets. We extracted at random subsamples of 3Mbp, 5Mbp, 8Mbp, 10Mbp, 30Mbp, 50Mbp, 100Mbp, 250Mbp from unassembled genomic shotgun datasets of 4 species (*Python molurus*,

Gekko gecko, *Ophisaurus attenuatus*, and *Pantherophis emoryi*), and compared RepeatMasker estimates of total TE genomic abundance for each. Read extraction was performed using the *subsample_fasta.py script* part of the *QIIME pipeline* (Caporaso et al. 2010). Finally, we compared RepeatMasker estimates of total TE genomic abundance in relation to the amount of sequence data obtained for all Illumina and 454 genomic shotgun datasets to test for biases related to sequencing technology, and for biases related to the amount of sequence data collected per individual, versus estimates of total TE genomic abundance.

CR1 and BovB LINEs phylogenetic and evolutionary analyses

Species-specific consensus sequences for both CR1-L3 and BovB LINE retrotransposons were generated in CLC genomic workbench 9.0.1 using default parameters, a linear gap cost, and the global alignment setting. Nuclear reads for each species were mapped to the consensus sequence of the LINE consensus sequence from the most closely related species available, which was used as initial reference (e.g., both CR1-L3 and BovB reference sequences for the Burmese python were generated by Castoe et al. (2013), and used as reference for building the consensus for the Mexican burrowing python). The first consensus generated was then used as a new reference for further rounds of re-mapping of nuclear reads until no additional mapping reads were recovered. Consensus sequences were determined by simple majority rule consensus, removing regions with coverage <10x after the second mapping iteration, and <20x in the final mapping. Consensus sequences were aligned in ClustalW (Larkin et al. 2007) with a gap open penalty of 50, and alignments were manually adjusted prior to downstream analyses (Supplementary Data 10, online). To the CR1 consensus sequences generated from our 66 squamate species, we added CR1-L3 and CR1-L2 vertebrate consensus sequences available in *RepBase*, for a total of 155 sequences (Supplementary Data 10, online). Squamate BovB consensus sequences we generated

from our 66 squamates were combined with other metazoan consensus sequences available in *RepBase*, for a total of 87 sequences (Supplementary Data 9, online). Bayesian phylogenetic tree reconstruction analyses of squamate CR1 and BovB LINEs were performed in BEAST2 (Bouckaert et al. 2014). Two independent analyses were run for 200 million generations each, following the Yule model of speciation and a relaxed log-normal clock model; MCMC chains were sampled every 1000 generations. The program Tracer v1.6 (Rambaut and Drummond 2007) was used to confirm that the MCMC chains had reached convergence. We conservatively discarded the first 25% of collected MCMC generations as burn-in, based on evidence that the likelihood and parameter values reached stationarity after approximately 15% of the sampling process.

CR1 and BovB LINEs coverage and age analyses

For each species, the species-specific CR1-L3 and BovB consensus sequence was used as a reference to estimate read coverage using the BWA *mem* alignment tool (Li and Durbin 2009), and the BEDTools2 (version 2.26.0) *coverage* tool (Quinlan and Hall 2010). Coverage counts were normalized by the total number of reads aligned to the full-length reference sequence. Read coverage was estimated for: (i) each 10bp sliding window, (ii) for the first and second half of the reference sequence, and (iii) for each third of the reference.

We used pairwise sequence divergence from the consensus (pairwise π) as a proxy to infer age and relative element level of activity through time. Pairwise distances values for each element and species were estimated following a custom pipeline starting from BWA alignments. An R (R Core Team) custom script built on the pegas (Paradis 2010) and stringr packages was used to calculate pairwise π estimates using multi-fasta pairwise alignments of reads to the reference.

Because we expected multiple TE subfamilies to exist, sequence divergence was estimated by excluding sites that define different CR1 and BovB subfamilies. For each species, we calculated the relative nucleotide frequency for each position in the multiple sequence alignment, and then calculated the mode of the frequency distribution (bins of 0.01) of the most frequent nucleotide at each position. Sites for which the most frequent nucleotide was in a bin more than 3 bins away from the mode were discarded as defining a separate subfamily.

Time calibrated phylogeny of 66 squamate reptiles

We estimated a time-calibrated phylogeny for the 66 squamate species in our study and an additional 8 outgroup vertebrates for comparative analyses of genomic repeat content. We downloaded and parsed 12 mitochondrial-encoded protein-coding genes for each species with a mitochondrial genome sequence available on Genbank. The same genes were parsed from *de novo* assembled mitochondrial genomes after genome annotation in MITOS (Bernt et al. 2013). We aligned the 12 protein coding genes encoded on the mitochondrial heavy strand using MUSCLE v.3.8.21 (Edgar 2004) and concatenated the sequences into a supermatrix alignment to be used for divergence dating (10,479 bp). Prior to divergence dating, we estimated the best-fit partitioning scheme and associated models of nucleotide substitution using Bayesian Information Criterion and the heuristic search algorithm provided in PartitionFinder v. 1.1.1 (Lanfear et al. 2012). We provided a starting partitioning scheme that defined 36 partitions (splitting codon positions for each of the 12 genes). PartitionFinder identified the best-fit partitioning scheme comprising a single partition for each codon position (3 total) and a GTR + I + G model for each partition. We estimated divergence times using BEAST v.2.3.4 (Drummond and Rambaut 2007) with a calibrated Yule model of speciation and a log-normal relaxed clock model. We constrained the topology to that provided from previous studies of the squamate phylogeny and

diversification (Benton and Donoghue 2007; Pyron et al. 2013). We constrained a total of 7 nodes using fossil calibrations also provided in previous studies. Calibration points and associated prior distributions are given in Supplementary Data 11. Two independent MCMC runs were conducted for 100 million generations each, with MCMC chain sampling every 10,000 generations. We assessed convergence to the posterior based on likelihood and parameter stationarity (ESS > 200 for all parameters) using the program Tracer. We discarded the first 10% of generations as burn-in, based on the likelihood and parameter values exhibiting stationarity at around 10% of sampling.

AATAG microsatellite seeding by TE analyses

We performed adjacency analyses of AATAG and ATAG SSR loci on high-quality assembled genomes for seven snake species, and used the green anole lizard as an outgroup. To increase specificity, genomes were first masked only for simple repeats. We extracted coordinates of annotated AATAG and ATAG SSR loci from the *.out* RepeatMasker output files, and used these coordinates to extract target regions 400 bp upstream and downstream of each microsatellite locus. We then performed a second run of RepeatMasker to mask only TEs located in the extracted target regions that flank AATAG and ATAG loci. Following this strategy, we were able to annotate TEs located in close proximity to SSR loci, and to differentiate TEs that harbor microsatellite-like regions in their reference sequences. The composition of TEs physically associated with SSR loci regions was then compared to the average of five independent randomly generated genomic backgrounds matching in sample size the corresponding microsatellite landscape. For each species, genomic background reads were generated by using the *random* tool in the BEDTools2 v.2.26.0 package, in which we specified the number of sequences to be extracted and that their length was to match the SSR-adjacent genomic

subsample. The generation of random bed files was performed independently five times per species, the TE composition was averaged across these five genomic backgrounds, and then compared to SSR loci adjacent regions. Fisher's one-tailed exact tests were performed to evaluate the enrichment of TE families in SSR loci regions (at $\alpha = 0.01$). Finally, to identify the specific element types involved in microsatellite seeding, we estimated genomic and SSR-adjacent conditional probabilities of TE-SSR co-occurrences. We estimated the conditional probability of sampling an AATAG SSR with an adjacent CR1 LINE present within 400 bp, and compared this to the estimated joint probability of sampling an AATAG SSR locus and a CR1 LINE using the genome-wide frequencies. We also calculated the conditional and joint probabilities for Rex LINEs, and compared those to the conditional and joint probabilities of CR1 LINEs, respectively.

Effective population size (N_e) estimation

Whole genomic Illumina pair-end reads for 8 squamate reptiles species were first preprocessed for quality using Trimmomatic (Bolger et al. 2014). Clean paired and unpaired reads were aligned to the respective reference genomes using BWA v.0.7.12, and SNPs were called with SAMtools (v.0.1.18) *mpileup* (Li et al. 2009). We applied the pairwise sequential Markovian coalescent model (PSMC; Li and Durbin 2011) using a generation time of 3 years across all 8 species (which represents the average of generation time approximations available from the literature; Supplementary Data 12) after verifying that the application of a single generation time yielded results consistent with estimates of average N_e produced by the application of generation times within the range reported in the literature. Multiple studies provided evidence of relatively similar mutation rates across lineages of squamates (Castoe et al. 2013; Green et al. 2014). Therefore, in our PSMC analyses we used the generalized squamate mutation rate reported in

Green et al. (2014) of 2.4 x 10^9 /year/site (as estimated from 4fold degenerate sites between anole and python). To test the robustness of the returned population estimates, we conducted 100 bootstrap replicate analyses by splitting the scaffolds into smaller segments and randomly sampling the segments with replacement. Default outputs of the *psmc_plot.pl* script were used to graphically summarize N_e changes over time estimations per each bootstrapped sample (Supplementary Fig. 10b).

Coalescent approaches for estimating N_e and changes in N_e over time (like PSMC) have several intrinsic limitations. Importantly, they rely on explicit assumptions of a single population coalescent model (without subdivision, gene flow, or selection) to estimate the time since the most recent common ancestor of alleles at each locus, as well as an assumed generation time and substitution rate. Population structure has been identified as one major factor that can bias PSMC-based estimates of N_e (Nielsen and Beaumont 2009; Mazet et al. 2015; Boitard et al. 2016; Nadachowska-Brzyska et al. 2016). For example, the inferred trend in N_e variation of a structured population can portrait either a bottleneck or an expansion in population size whether the alleles were sampled from the same subpopulation or from different subpopulations, respectively (Mazet et al. 2016). Episodes of natural selection can also bias estimates of *Ne* obtained using PSMC, as selection can manipulate the rate of coalescence at specific loci that are directly or indirectly linked to targets of selection (Schrider et al. 2016; Adams et al. 2018). Given the nature of our data, we are not able to assess the presence and extent of population substructure or selection, and therefore cannot exclude that our PSMC estimates are immune to such biases. Additionally, PSMC has low power at recovering rapid changes in N_e, which may be incorrectly estimated to have occurred over a longer period of time, and cannot recover recent nor very ancient changes in N_e (e.g., younger than ~10kyBP and older than ~3myBP for humans;

Li and Durbin 2011; Mazet et al. 2016). Thus, we suggest caution when interpreting our PSMC estimates of N_e and N_e changes through time. However, we found low variance across bootstrapped N_e estimates once the most recent and most ancient time points are removed, and patterns of expansion and contraction of N_e are consistent with alternations of glacial and interglacial periods during the middle Miocene climate transition, the Pliocene and the Pleistocene (Zachos et al. 2001).

In an attempt to cope with biases associated with PSMC estimates of recent and ancient changes in N_e, median N_e values were calculated after removing the first and the last time points from each sample. We replicated each analysis (see below) after applying different filtering schemes to the standard PSMC outputs (e.g., removal of 10% and 25% of time point data, and inclusion of only time points between 20kyBP and 10myBP). Since all tests provided the same conclusions, we report only analyses performed using median N_e values that were calculated according to the original filtering scheme. Additionally, we replicated all of our analyses using adult body mass as a proxy for effective population size (Figuet et al. 2016) to avoid potential biases associated with our coalescence-based methods of N_e estimation (i.e., Fig. 4e). For each of the 66 squamate species, we obtained adult body mass measurements from the literature (Feldman et al. 2016) which were used to further test for a demographic explanation for variation in repeat content alongside coalescent-based estimates of N_e.

Testing demographic explanations of repeat content variation

We performed linear regression analyses to test for correlations between N_e and truncation, N_e and genomic abundance of BovB and CR1-L3 LINEs, truncation and genomic abundance of repeats, and between truncation and estimates of ages of repeat element activity. We used the *pic* function in APE and the time-calibrated phylogeny to compute phylogenetic independent

contrasts (PICs) to be used for all linear regressions. These analyses were conducted for both the coalescent-based estimates of N_e and adult body mass as a proxy for N_e. Since truncation values violated both assumption of normality and homogeneity of variance (Shapiro-Wilks test; p-values < 0.05 and Bartlett's test; p-values < 0.05), we performed statistical analyses on log-transformed values (Shapiro-Wilks test; p-values > 0.05 and Bartlett's test; p-values >0.05).

Data availability

New raw, unassembled shotgun sequencing data and new assembled genome data has been deposited at NCBI under the following accessions: PRJNA413172 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA413172) and PRJNA413201 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA413201). The authors declare that all data and scripts used in this study are available via public databases or available from the corresponding author upon request.

Acknowledgment

Support for this work was provided from startup funds from the University of Texas at Arlington to TAC. We acknowledge the Texas Advanced Computing Center (TACC) for providing access to computational resources.

Figures

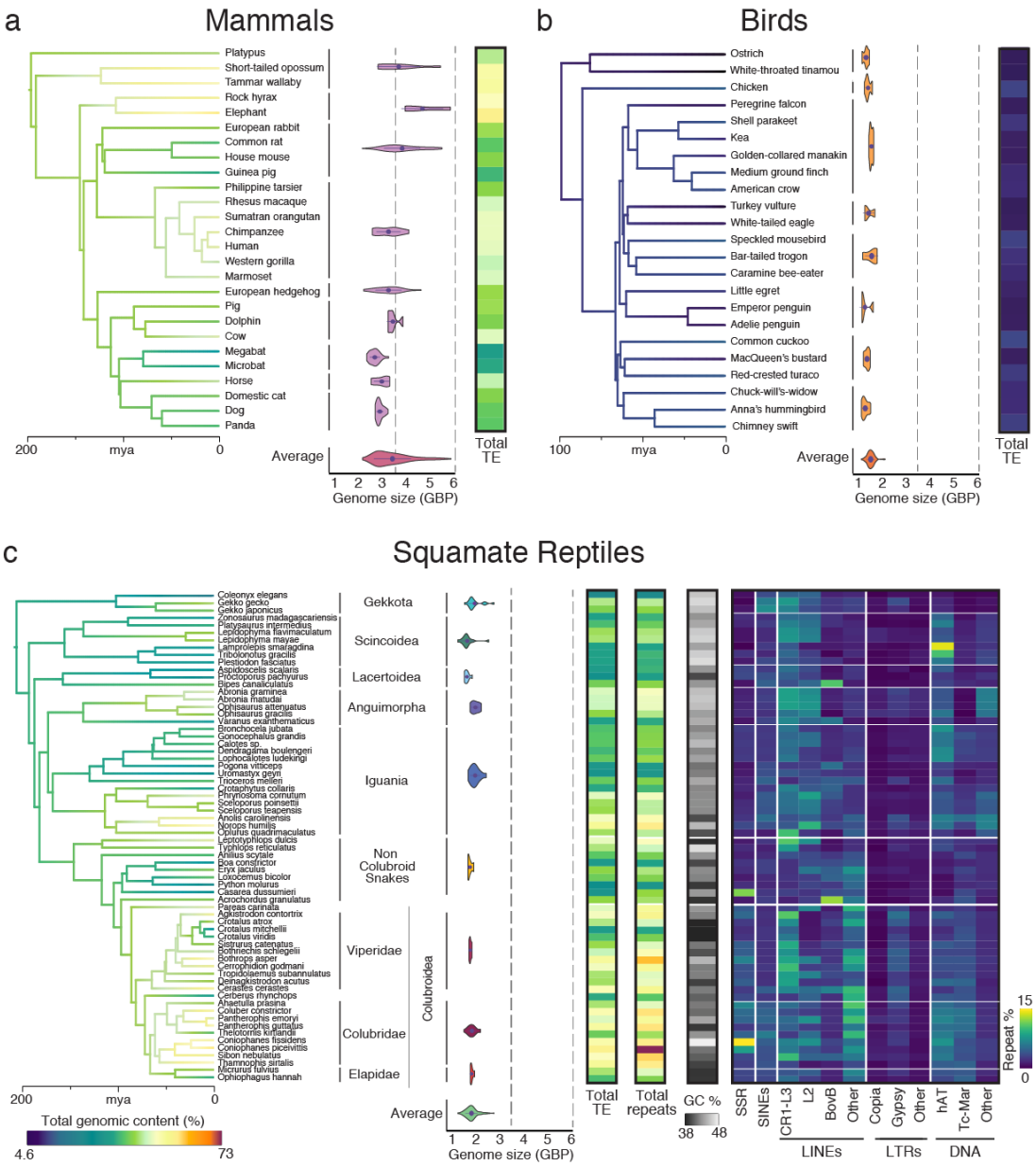


Figure 1. Genomic transposable element (TE) abundance and genome size variation in mammals, birds, and squamate reptiles. Branches on the time-calibrated consensus phylogeny are colored according to the estimated rate of genomic TE evolution. Violin plots show distributions of flow cytometry-based genome size estimates for major groups of **a** mammals, **b** birds, and **c** squamate reptiles, and the associated heat maps reflect the total genomic TE content (%) for each taxon. For squamate reptiles, additional heat maps show percent genomic repeat element content, percent genomic GC content, and percentages of major components contributing to the overall repeat element landscape.

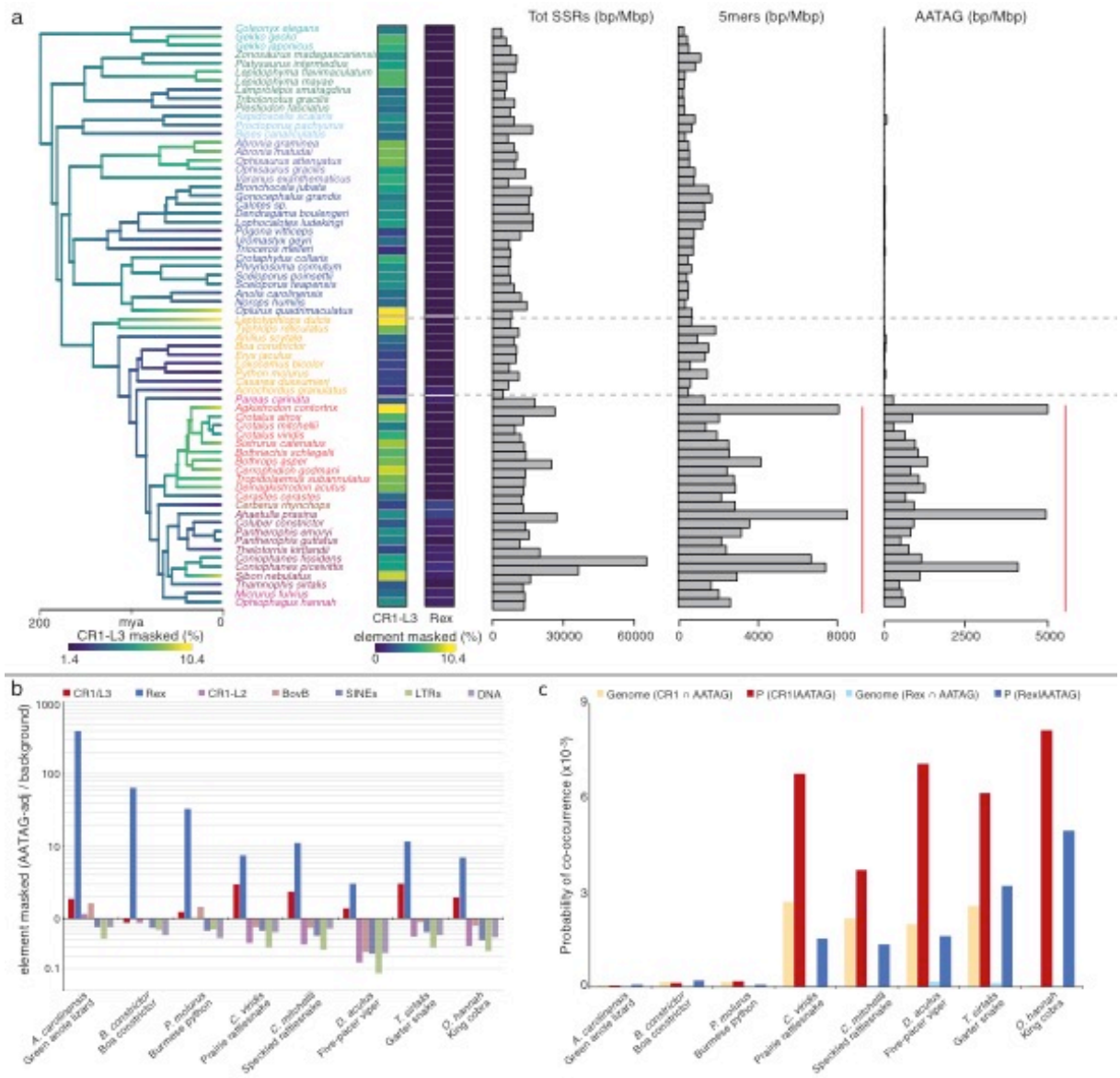


Figure 2. Microsatellite seeding by transposable elements (TEs) in squamate reptiles. **a** Branches on the time-calibrated consensus phylogeny are colored according to estimated rates of genomic CR1-L3 LINE evolution. Heat maps show the total genomic content (%) of LINE retrotransposon types involved in microsatellite seeding. Associated bar plots represent the total (left), 5mer (middle), and AATAG (right) microsatellite bp/Mbp density frequencies for each genome sampled. Red lines to the right of the bar plots highlight pronounced seeding of 5mer and AATAG microsatellites in colubroid snakes. **b** The ratio between TE mapping at the 5′ tail of AATAG microsatellite loci (AATAG-adjacent) and TE content averaged over five independent, randomly simulated genomic backgrounds for each class of TEs (SINEs; CR1-L3, Rex, CR1-L2 and BovB LINEs; LTRs; and DNA transposons). Ratios are plotted on a log scale to highlight enriched elements flanking AATAG loci (ratio >1) in contrast to elements more abundant in the genomic background (ratio <1). **c** Histogram shows joint and conditional probabilities of associations between AATAG loci and CR1-L3 and Rex. Genomic joint probabilities are shown in orange and light blue for CR1-L3 and Rex, respectively. AATAG-adjacent conditional probabilities are shown in red and dark blue for CR1-L3 and Rex, respectively.

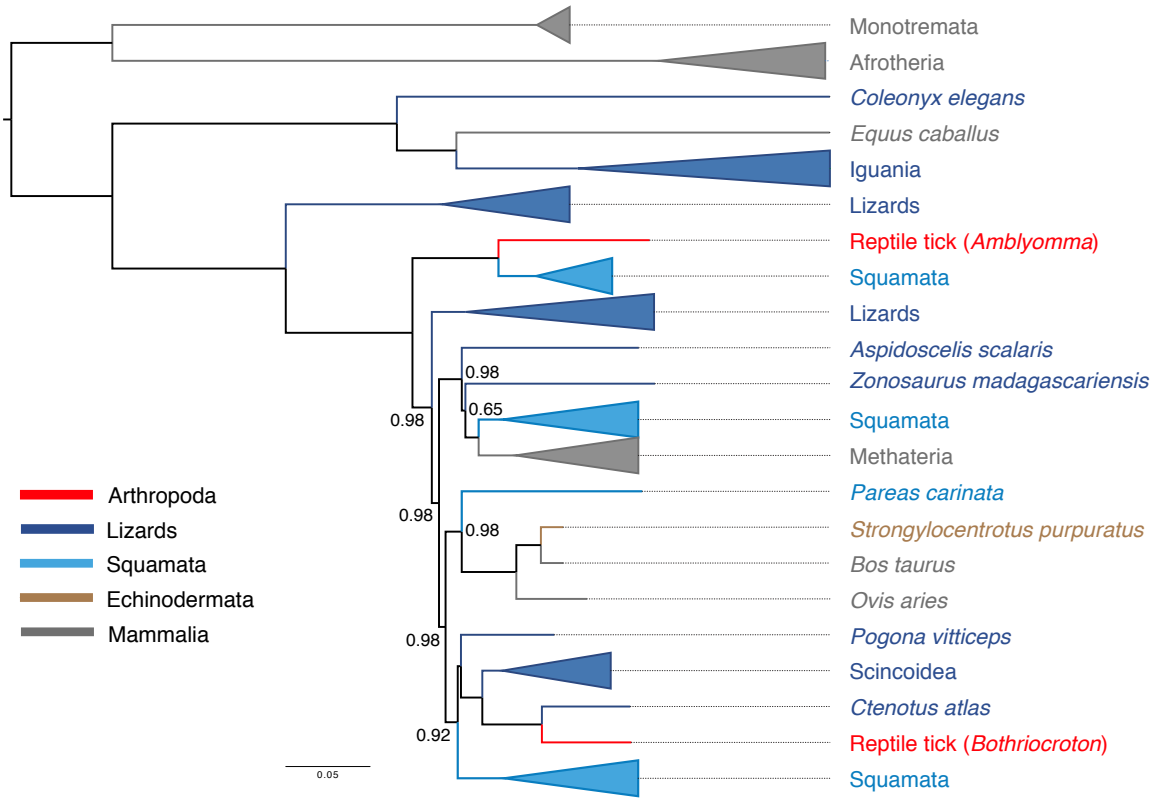


Figure 3. Evidence for ectoparasite-mediated horizontal transfer of BovB LINEs in squamate reptile genomes. A summarized Bayesian phylogenetic tree of full-length BovB LINE sequences for 87 metazoan species, including two reptile ticks. Branches have been collapsed and colored to represent major clades. Posterior probabilities are shown only at nodes that had posterior support <0.99

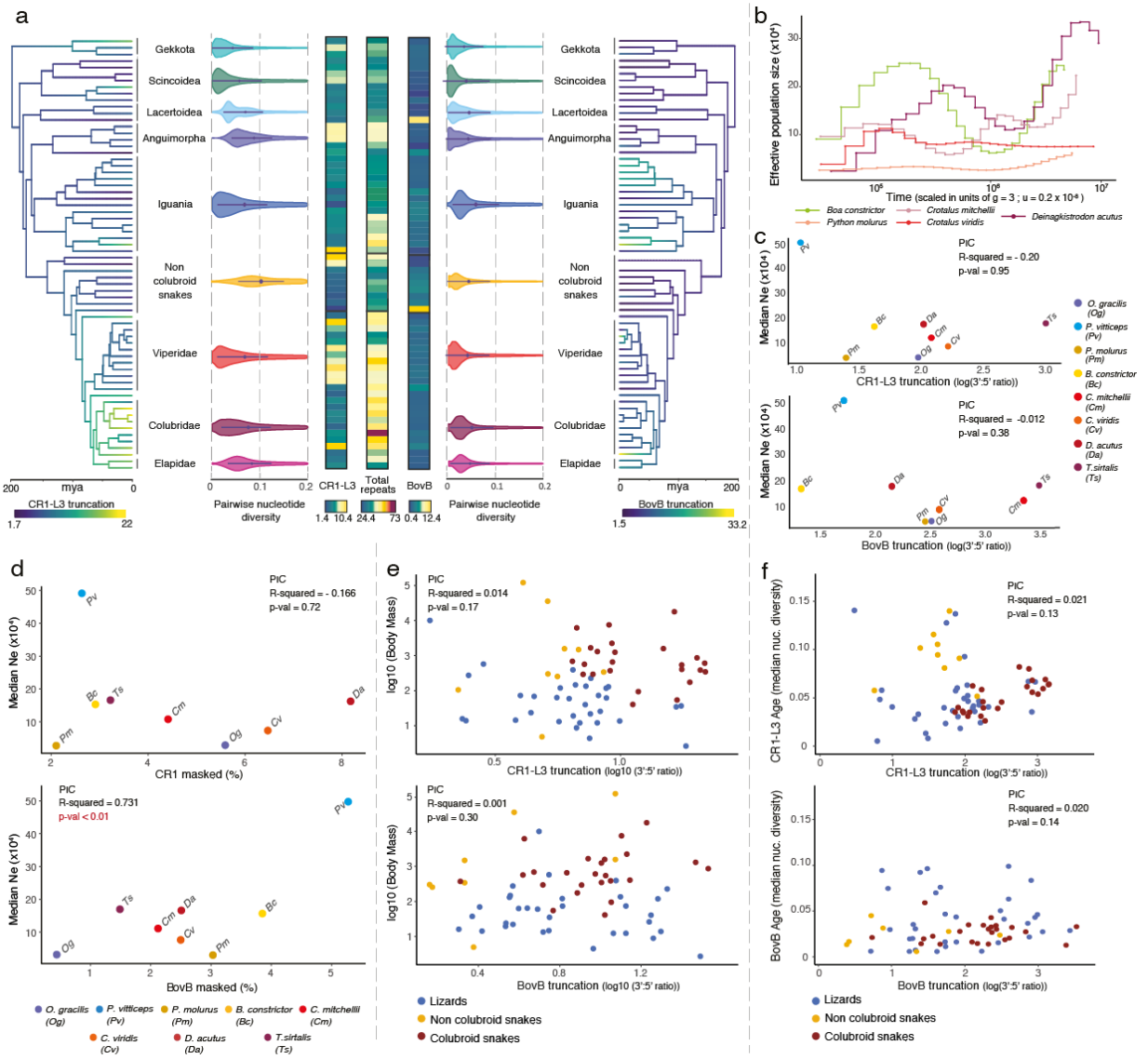
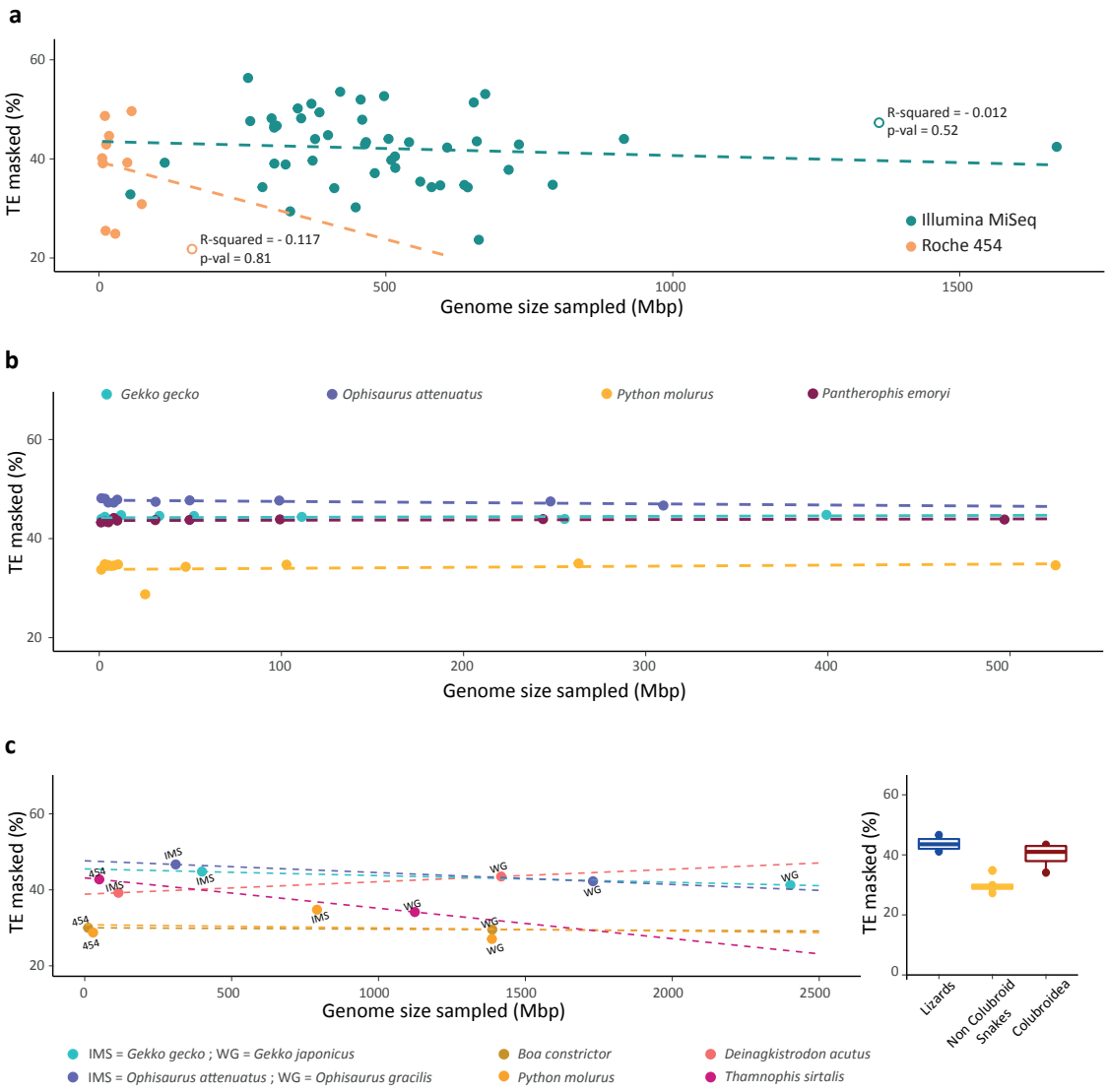


Figure 4. Relationships between truncation, effective population size, body mass, and divergence estimates for CR1-L3 and BovB LINE retrotransposons among squamates. **a** Branches on the time-calibrated consensus phylogenies are colored according to the calculated 3′:5′ read depth coverage ratio for CR1-L3 (left) and for BovB (right) LINEs. Heat maps show the genomic content of CR1-L3 LINEs, total repeats, and BovB LINE retrotransposons represented as percentages of the total genome. For each major clade, violin plots show the density distributions of divergence estimates (pairwise π) for all CR1-L3 and BovB elements compared to the species-specific consensus sequence. **b** Variation in effective population size (N_e) over time for five snake species scaled by generation time and mutation rate ("g" and "u" on the x-axis). **c** Relationship between N_e and truncation of CR1-L3 (top) and of BovB (bottom) LINEs for eight squamate species. **d** Relationship between total genomic abundance of CR1-L3 (top) and BovB (bottom) LINEs and N_e. **e** Relationship between adult body mass and degree of truncation across 66 squamate species for CR1-L3 (top) and BovB (bottom) LINEs. **f** Relationship between age (median pairwise π) and truncation for CR1-L3 (top) and BovB (bottom). Summary statistics from phylogenetically independent contrasts (PIC) are shown as insets for each plot in **c**–**f**. Statistical analyses were performed after log transformation of truncation values in plots **e** and **f**.

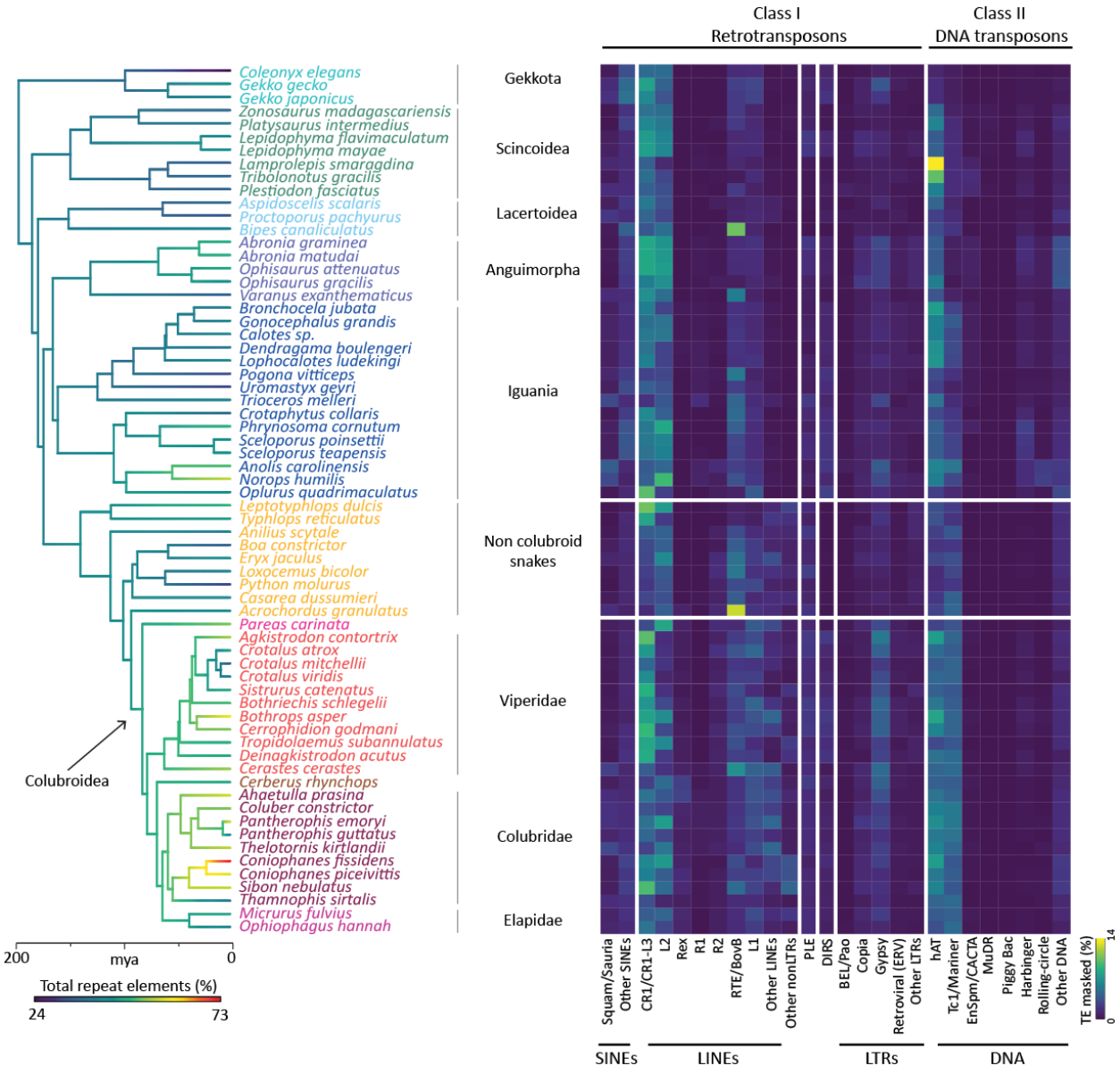
Supplementary Note 1

Mammal and bird genome size and transposable elements analyses. Genome size estimates based on flow cytometry analyses were retrieved for all mammal, bird and squamate reptile species available on the Genome Size database (Gregory 2017 - last accessed on 05 August 2017; Supplementary Data 2) These estimates were used to calculate ranges of genome size for each lineage and for each major clade of mammals, birds and squamate reptiles (Fig. 1). For bird and mammal species, we reported estimates of the genomic TE content when *de novo* repeat annotation had previously been performed for each individual species if available. For mammal species, we report estimates available on the RepeatMasker online database (Smit et al. 2015-2019), and we used data available in Kapusta et al. (2017) for bird species (Supplementary Data 1).

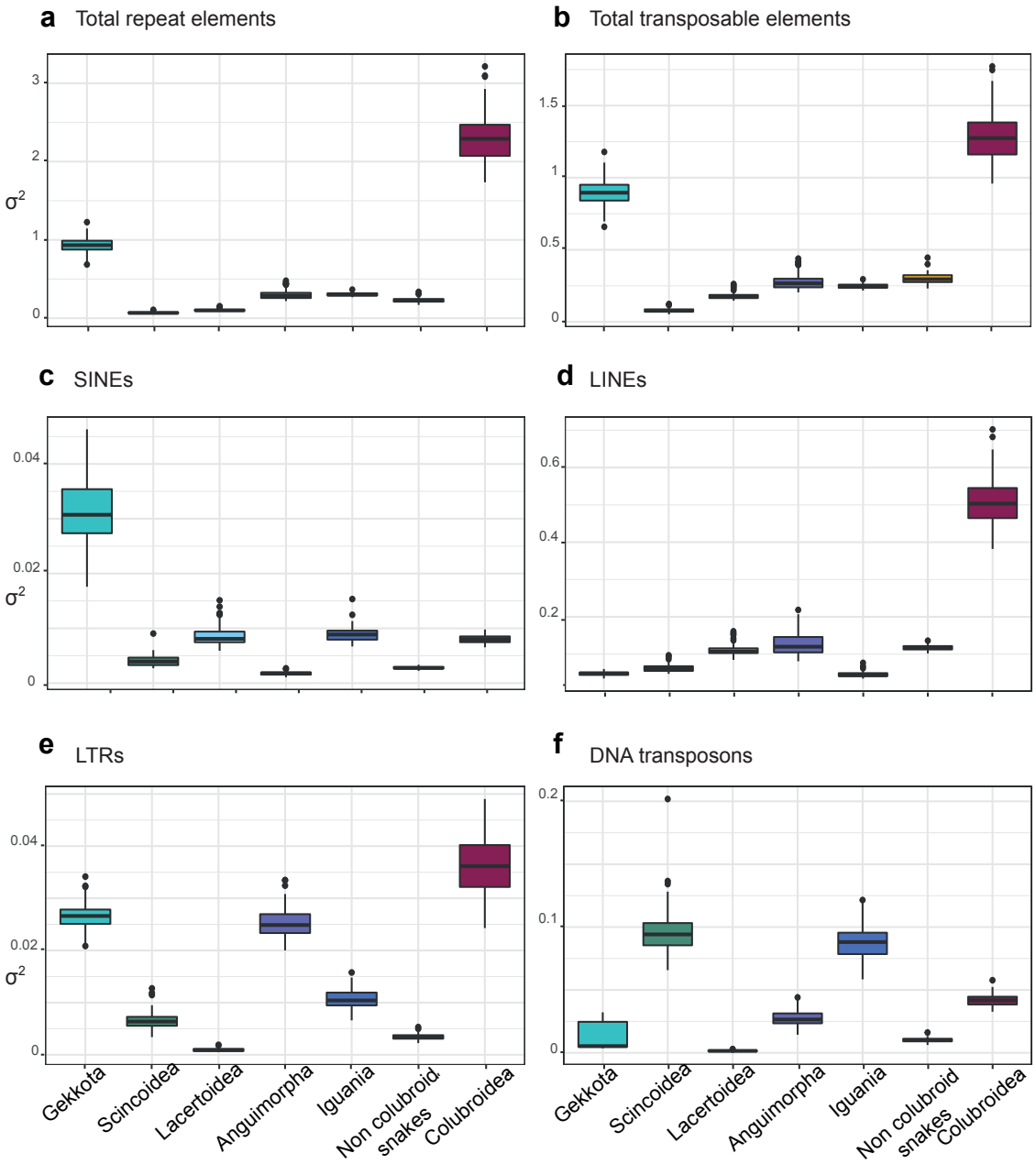
Supplementary Figures



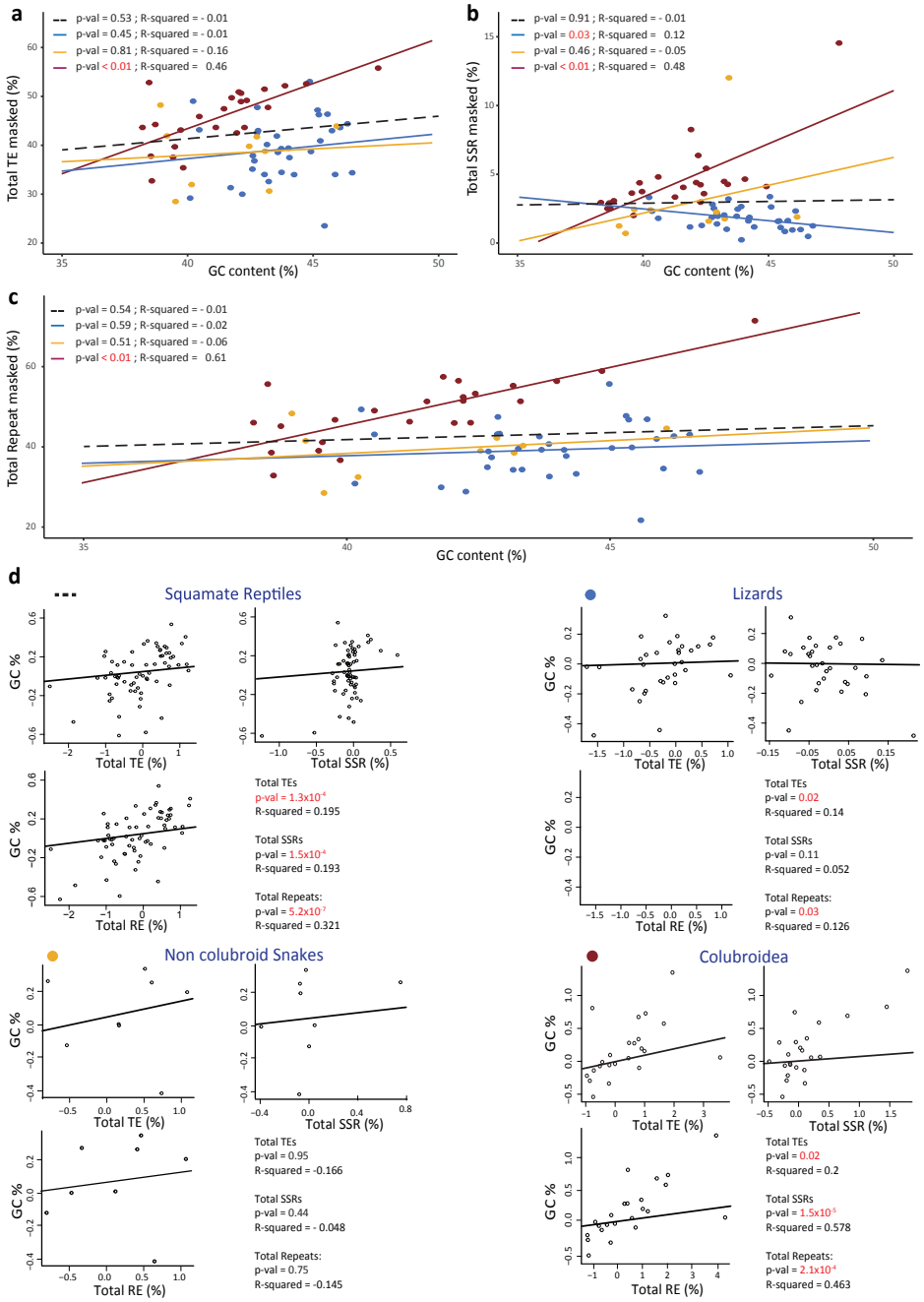
Supplementary Figure 1. Consistency of transposable element genomic content estimates across sequencing techniques, data assembly methods and proportion of genome sampled. **a**) Scatterplot shows estimates of genomic TE content in relation to the amount of genome sampled from unassembled shotgun sequencing data using Roche 454 (454) and Illumina MiSeq (IMS) sequencing technologies. **b**) Comparison of genomic TE content estimates across subsamples of the total amount of sequence data obtained. **c**) Comparison of TE estimates between unassembled genomic shotgun reads and assembled whole genomes for the same species or for closely related species belonging to the same genus (left); boxplot shows the distribution of genomic TE estimates for the same species, clustered according to major squamate reptile clades (right).



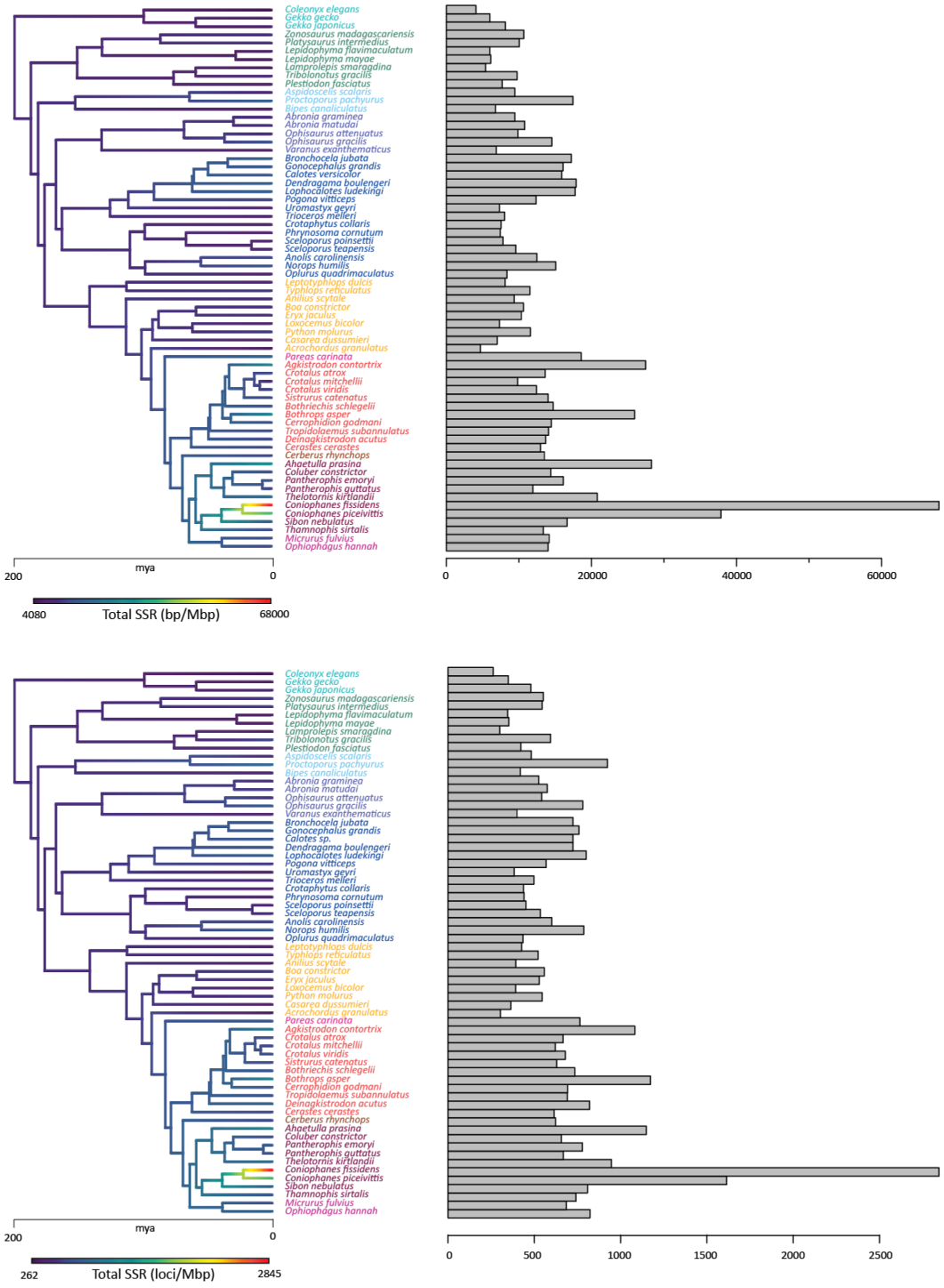
Supplementary Figure 2. Genomic repeat element landscape for 66 squamate reptile species. Branches on the time-calibrated consensus phylogeny (left) are colored according to the estimate rates of total genomic repeat elements masked (%), allowing for an intuitive visualization of the extent of the variation in repeat element content across squamates (dark blue = lower values, red = higher values). Heat map (right) reflects variation in the relative abundance of repeat elements across 66 squamate species, and highlights both between and within clade significant differences. For example, Gypsy LTR and Tc1 DNA transposons are more abundant in the genomes of colubroid snakes than in other squamate genomes. Cells in the heat map are colored according to the color gradient: dark blue= low; yellow= high. From left to right: Short Interspersed Nuclear Elements (SINEs); Long Interspersed Nuclear Elements (LINEs); Penelope-Like Elements (PLEs); DIRS; Long Terminal Repeat (LTR) retrotransposons, and DNA transposons.



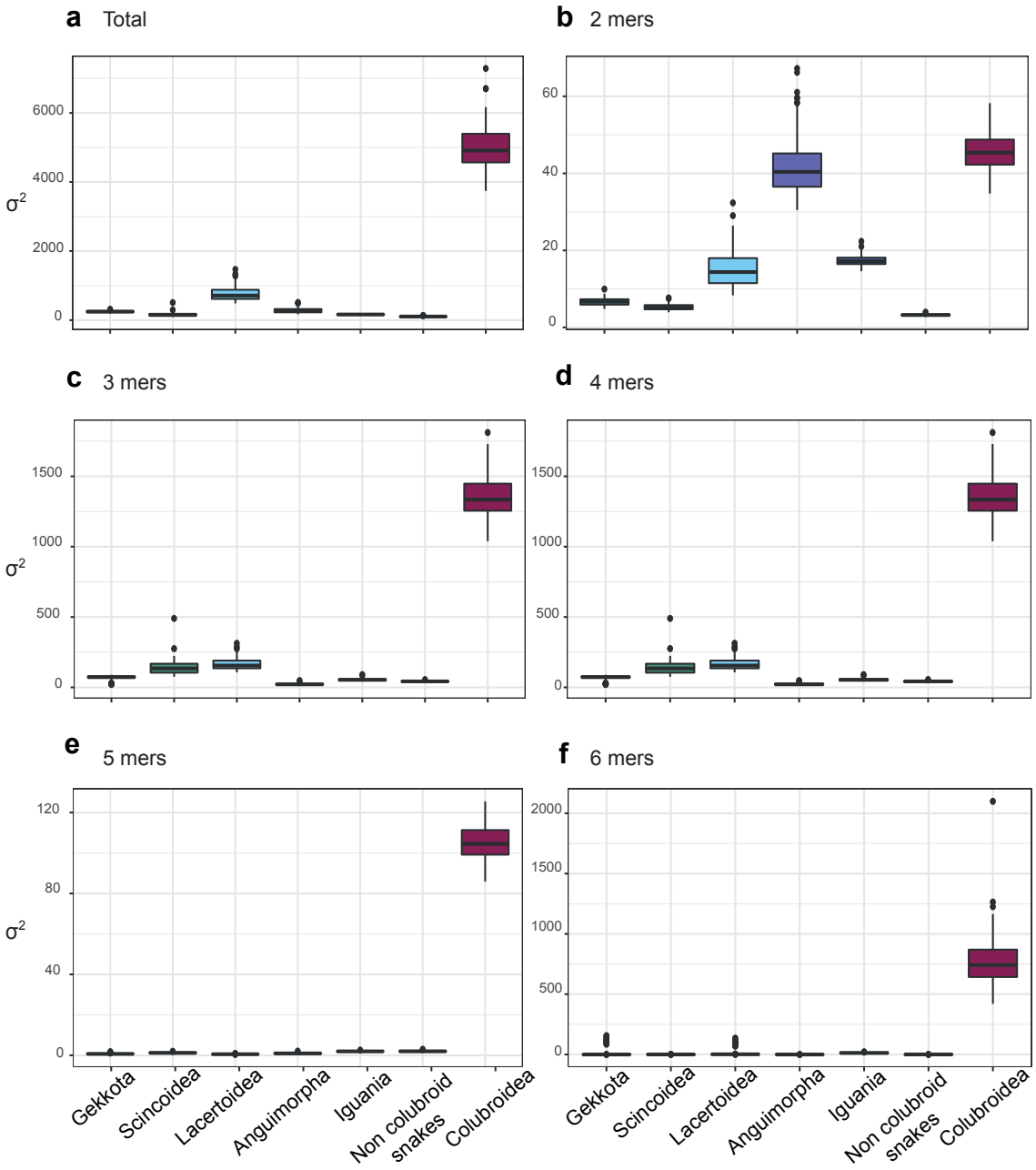
Supplementary Figure 3. Censored rate test results for lineage-specific rates of repeat elements evolution across seven major squamate clades. Box plots represent the rate parameter (σ2) estimates obtained across 100 trees sampled from the posterior distribution inferred from *BEAST* for the 7 major clades: gekkota(3), scincoidea(7), lacertoidea(3), anguimorpha (5), iguania (15), non colubroid snakes (9), and colubroidea (24). The null hypothesis of a single rate of evolution for all branches was rejected for all 600 censored rate tests (100 tree for all major families of TEs and for the total repeat element and TE content; for all tests, p-values < 0.01). Results confirm that, during squamate evolution, different lineages experienced differential rates of repeat element genomic accumulation (e.g., SINEs and LTRs in colubroidea and gekkota, or LINEs in colubroidea specifically). Results are shown for (**a**) total repeat element content, (**b**) total TE content, (**c**) Short INterspersed Elements (SINEs), (**d**) Long INterspersed Elements (LINEs), (**e**) Long Terminal Repeats (LTRs), and (**f**) DNA transposons genomic percentages.



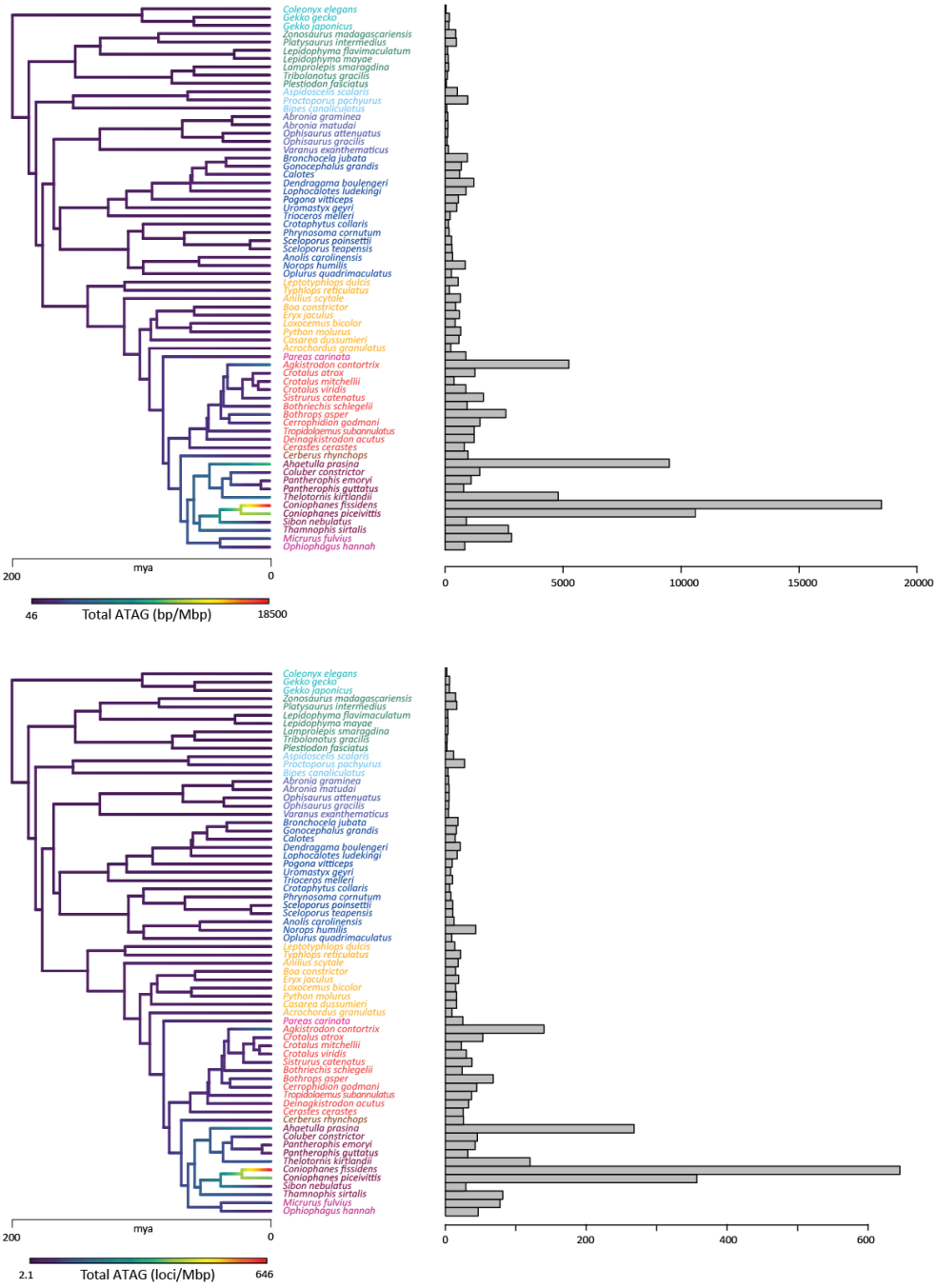
Supplementary Figure 4. Relationship between genomic GC content (%) and genomic TE, SSR and total repeat content (%). Scatter plots reflect the relationship between average genomic GC content and genomic estimates of the major components of the repeat element landscape (%) for each clade of squamate reptiles (lizards = blue; non colubroid snakes = yellow; colubroidea = dark red). **a**) Analysis of the the genomic GC content and total TE content (%). **b**) Analysis of the the genomic GC content and microsatellite (SSR) estimates performed in RepeatMasker. **c**) Analysis of the genomic GC content and total repeat element estimates. **d**) Phylogenetically independent contrasts (PICs) between genomic GC content, TE %, SSR % and total repeats (RE) %, for all squamates (top left) and for individual squamate lineages.



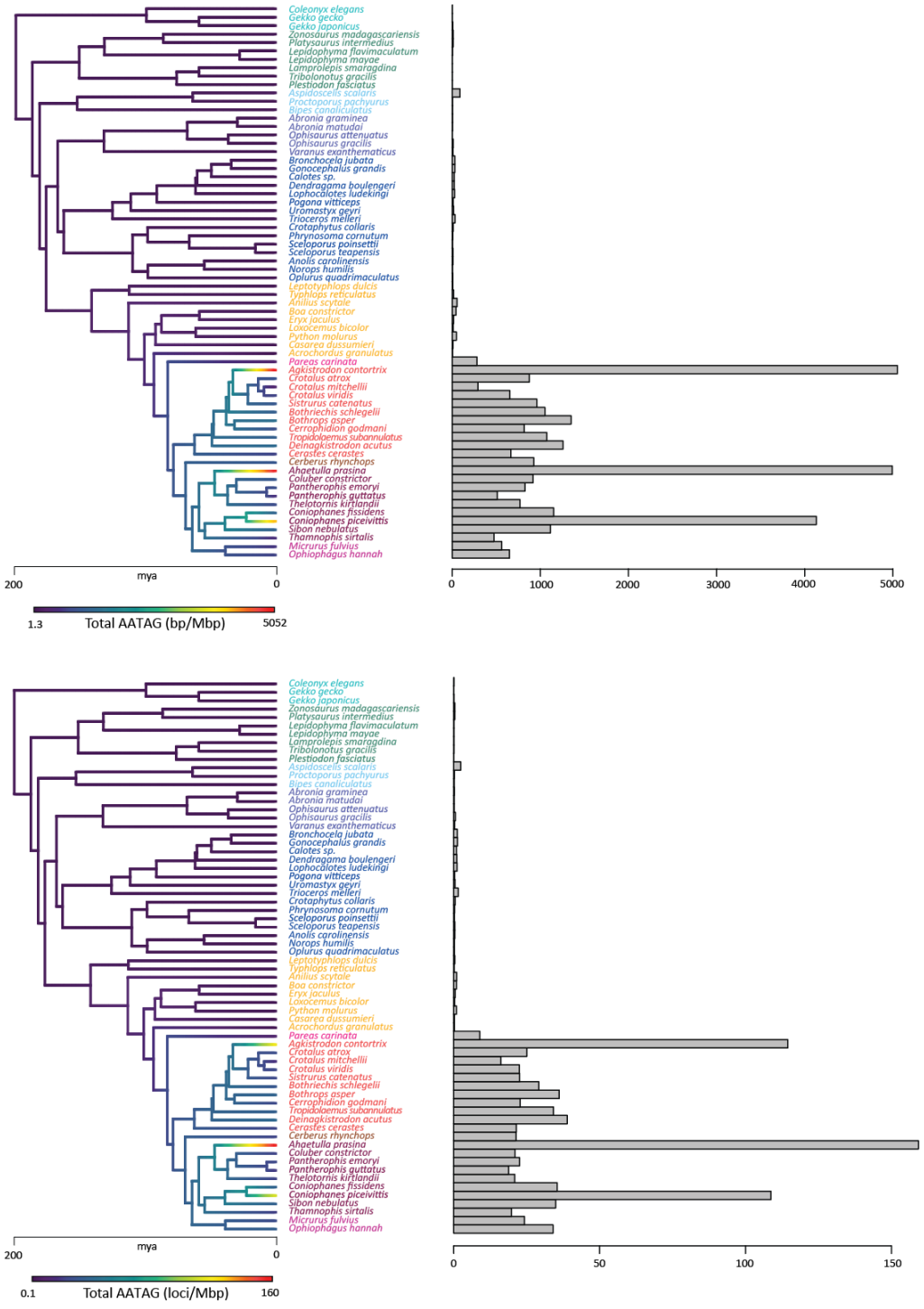
Supplementary Figure 5. Observed total microsatellite frequencies and their lineage-specific evolutionary rates across 66 squamate species. Horizontal bar plots represent the observed total microsatellite bp/Mbp (top) and loci/Mbp (bottom) density frequencies for each squamate genome sample. Branches on the time-calibrated consensus phylogeny are colored according to the estimated rates of microsatellite evolution.



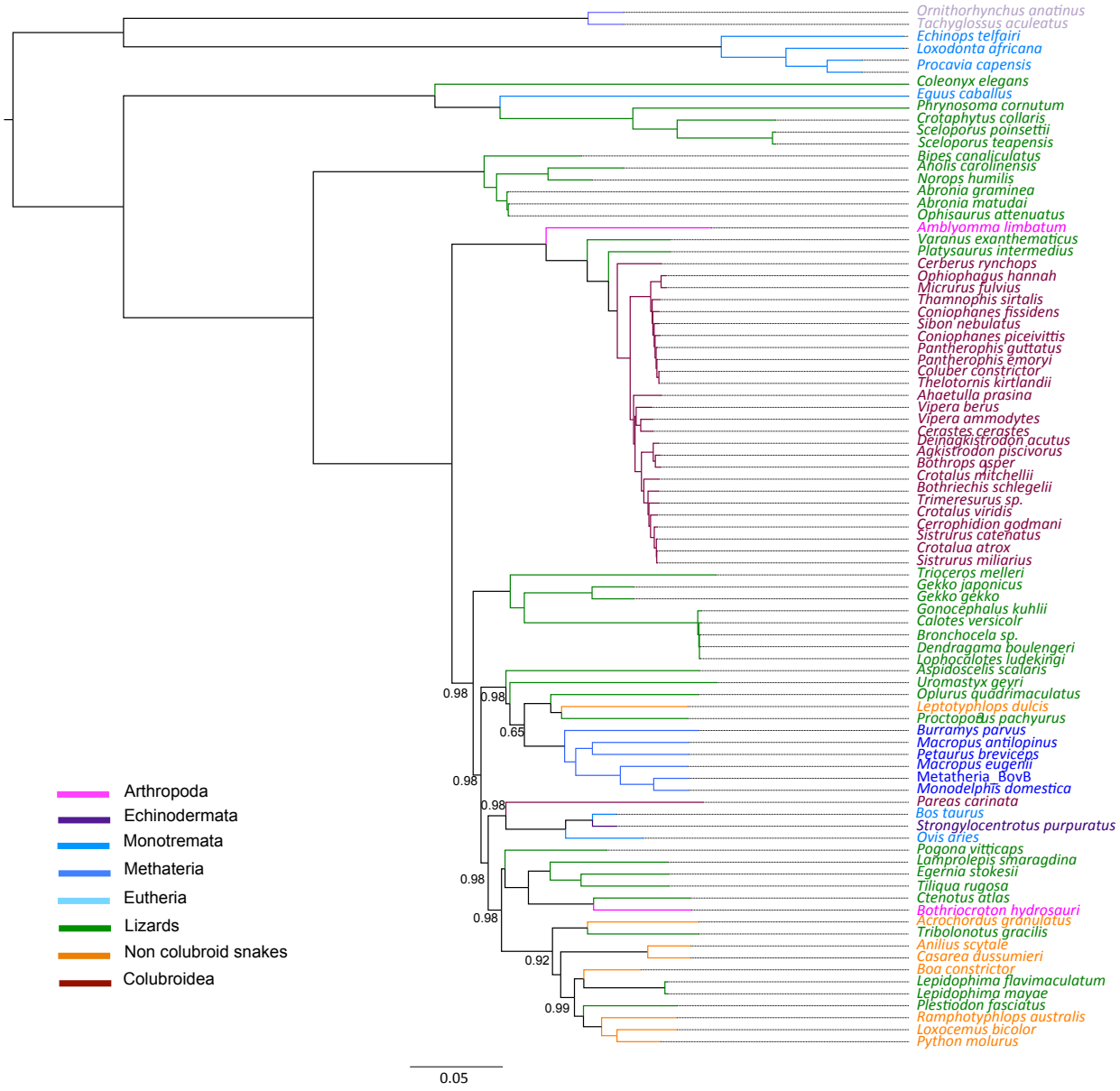
Supplementary Figure 6. Censored rate test results for lineage-specific rates of microsatellite evolution across seven major squamate clades. Box plots represent the rate parameter (σ^2) estimates obtained across 100 trees sampled from the posterior distribution inferred from *BEAST* for the 7 major clades: gekkota(3), scincoidea(7), lacertoidea (3), anguimorpha (5), iguania (15), non colubroid snakes (9), and colubroid snakes (24). The null hypothesis of a single rate of evolution for all branches was rejected for all 600 censored rate tests (100 tree for total loci/Mbp microsatellite density estimates and for 2-6mer SSR loci/Mbp density estimates; for all tests, p-values < 0.01). Results confirm that, during squamate evolution, there has been a significant expansion of all microsatellite types among the colubroidea branch specifically compared to all other squamates. Results are shown for (**a**) total microsatellite content, (**b**) 2mer, (**c**) 3mer, (**d**) 4mer, (**e**) 5mer, and (**f**) 6mer loci/Mbp density frequencies.



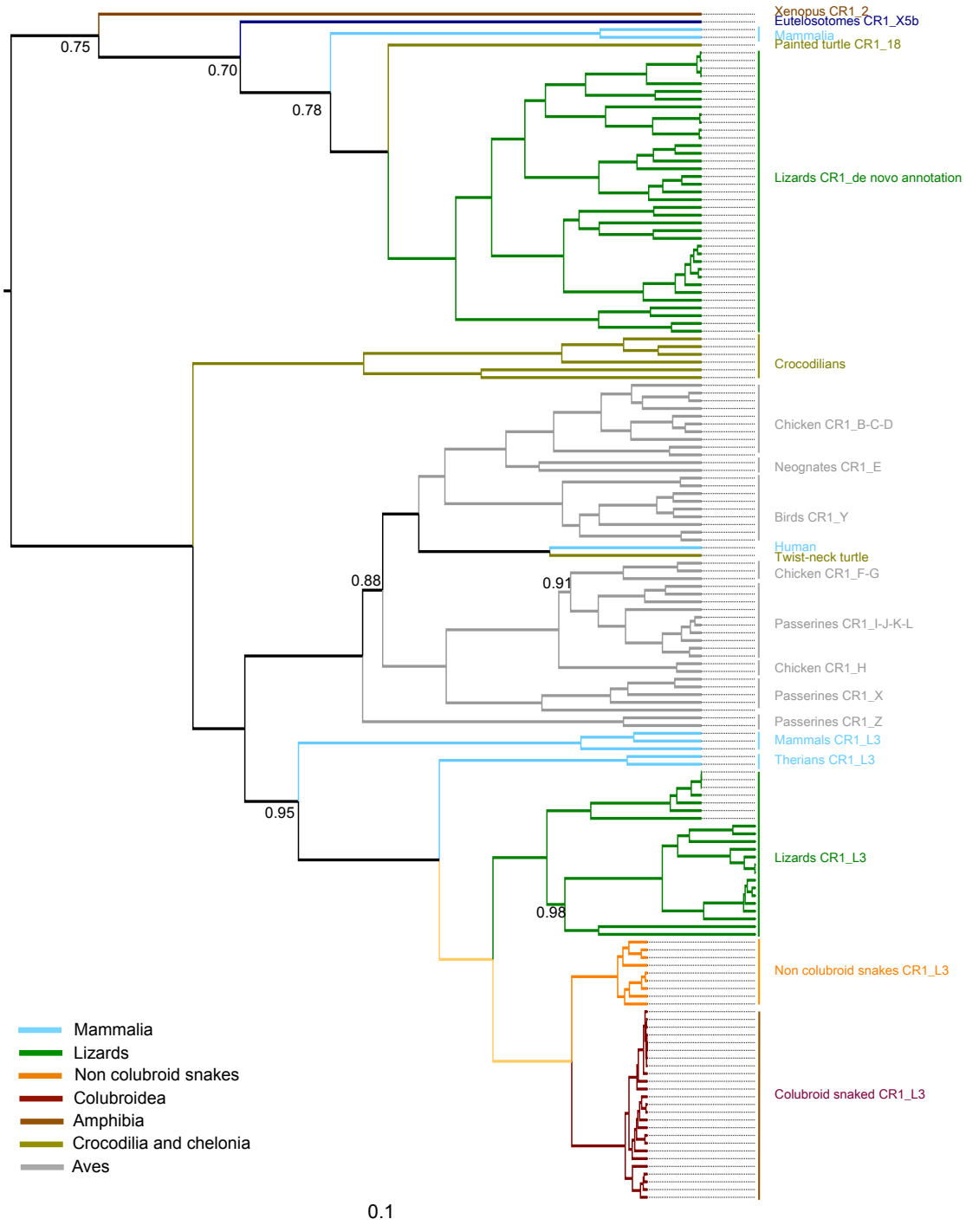
Supplementary Figure 7. Observed ATAG microsatellite loci frequencies and their lineage-specific evolutionary rates across 66 squamate species. Horizontal bar plots represent the observed ATAG 4mer microsatellite bp/Mbp (top) and loci/Mbp (bottom) density frequencies for each squamate genome sampled. Branches on the time-calibrated consensus phylogeny are colored according to the estimated rates of microsatellite evolution.



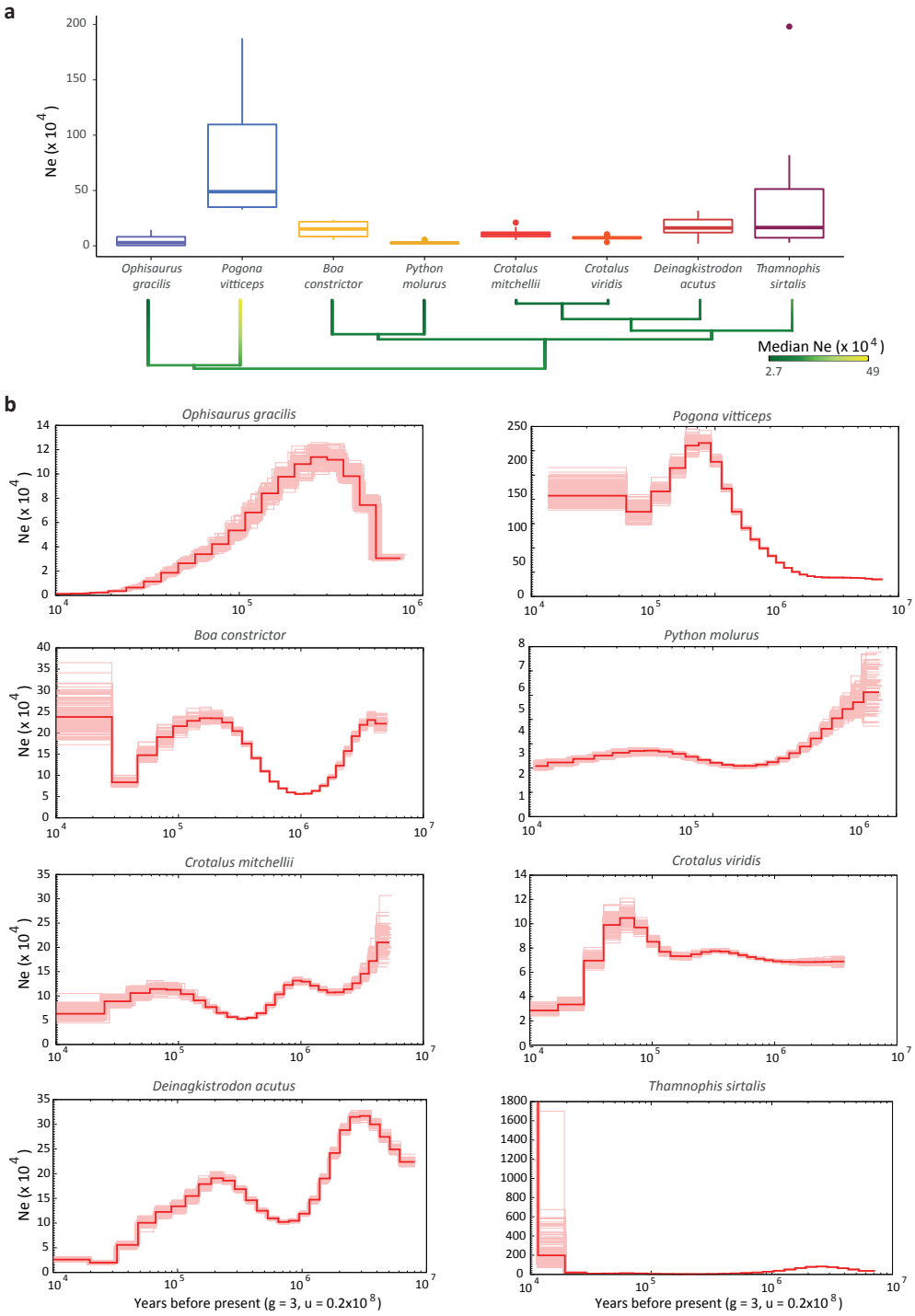
Supplementary Figure 8. Observed AATAG microsatellite loci frequencies and their lineage-specific evolutionary rates across 66 squamate species. Horizontal bar plots represent the observed AATAG 5mer microsatellite bp/Mbp (top) and loci/Mbp (bottom) density frequencies for each squamate genome sampled. Branches on the time-calibrated consensus phylogeny are colored according to the estimated rates of microsatellite evolution.



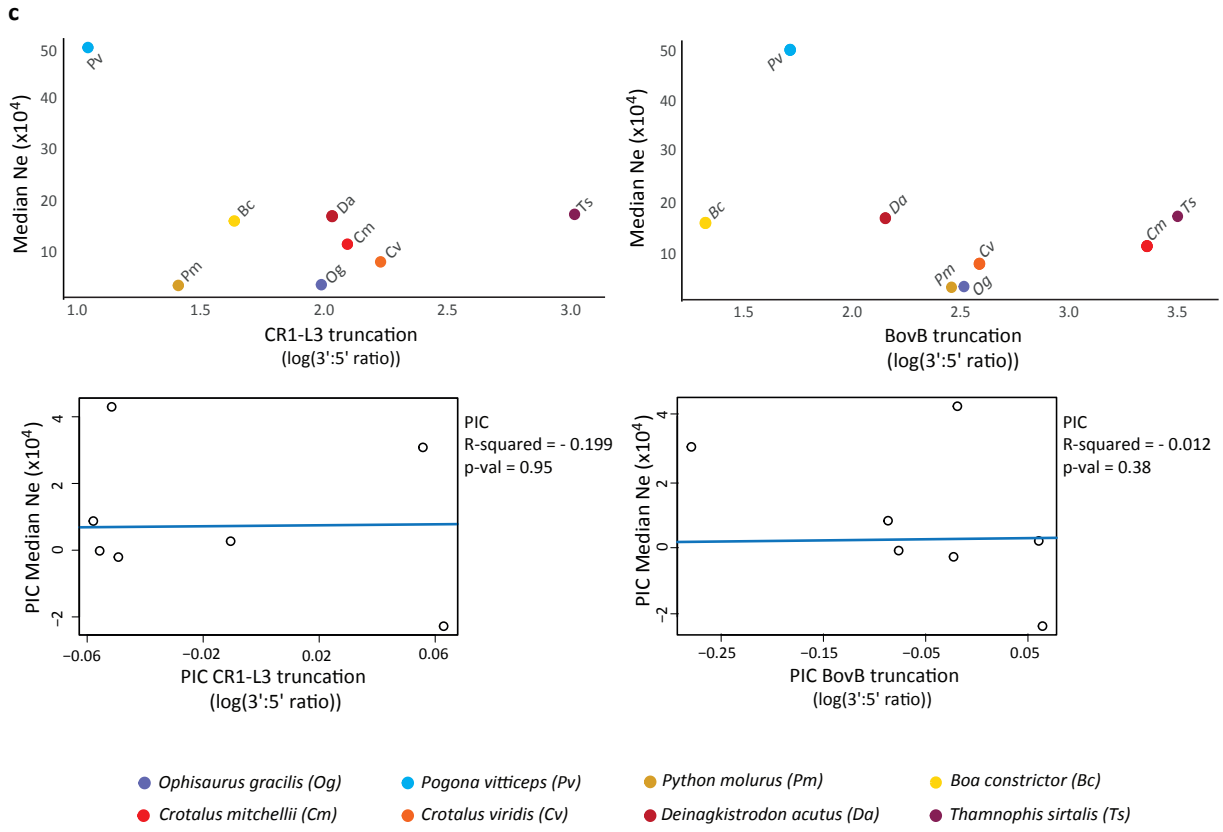
Supplementary Figure 9a. Phylogenetic tree reconstruction of 87 metazoan BovB sequences. Bayesian phylogenetic tree was built using *BEAST2*. 141 metazoan sequences were initially aligned in Clustal W, then manually edited and curated (final alignment length of 3134bp). For displaying purposes, we pruned the RTE-2 sequences of monotremata used as outgroup to root the tree. Posterior values are reported only for nodes with posterior support < 0.99.



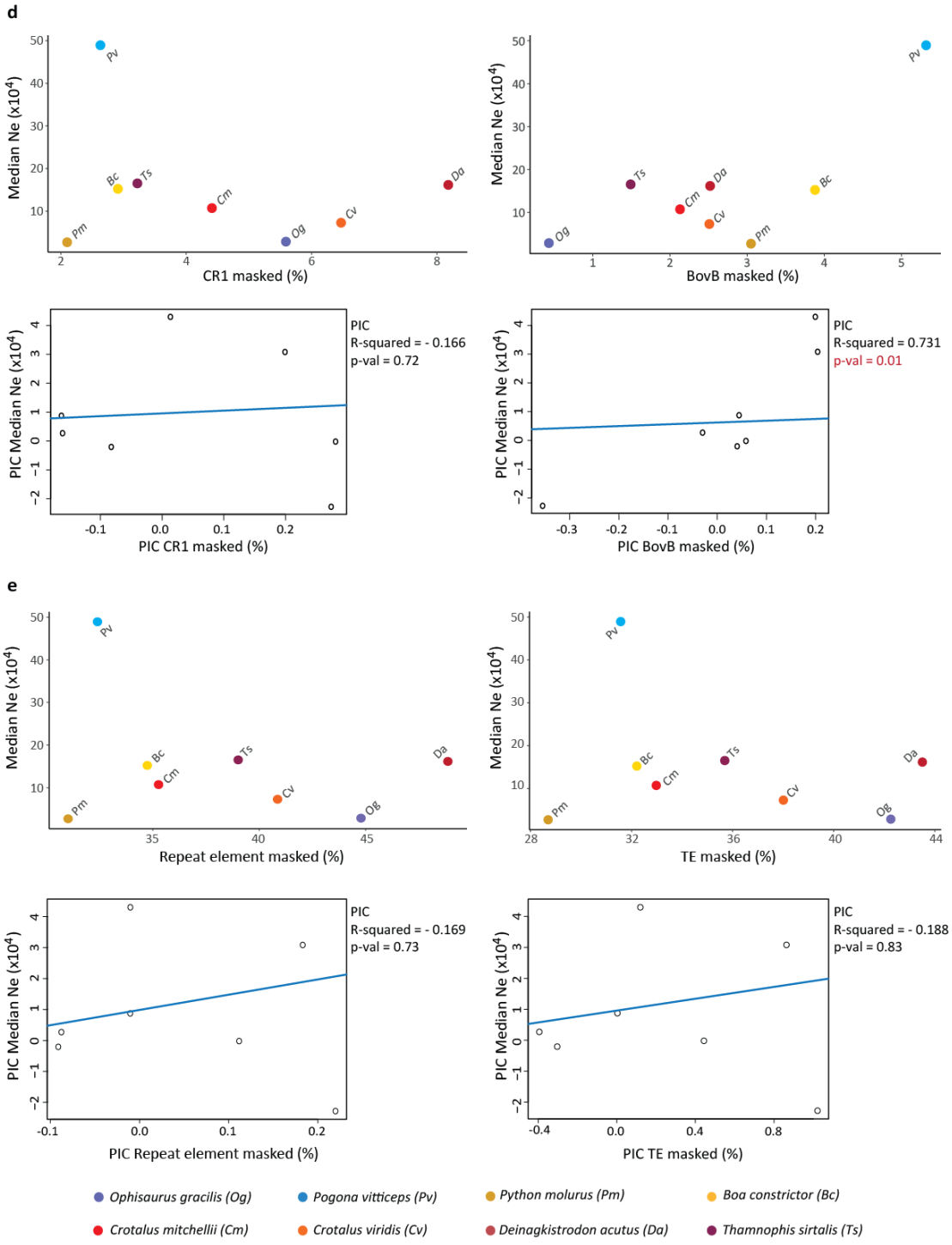
Supplementary Figure 9b. Phylogenetic tree reconstruction of 161 metazoan CR1-L3 LINE sequences. Bayesian phylogenetic tree of the full length CR1-L3 ORF was built using BEAST2. 161 vertebrate sequences were initially aligned in ClustalW, then manually edited and curated (final alignment length of 2967bp). Posterior values are reported only for nodes with posterior support < 0.99. For displaying purposes, we pruned L2 LINEs that were used as outgroup to root the tree and extremely divergent CR1 sequences of *Xenopus tropicalis* (Xt 1a and 1b), *Danio rerio* (Dr 29) and of *Latimeria chalumnae* (Lme 1).



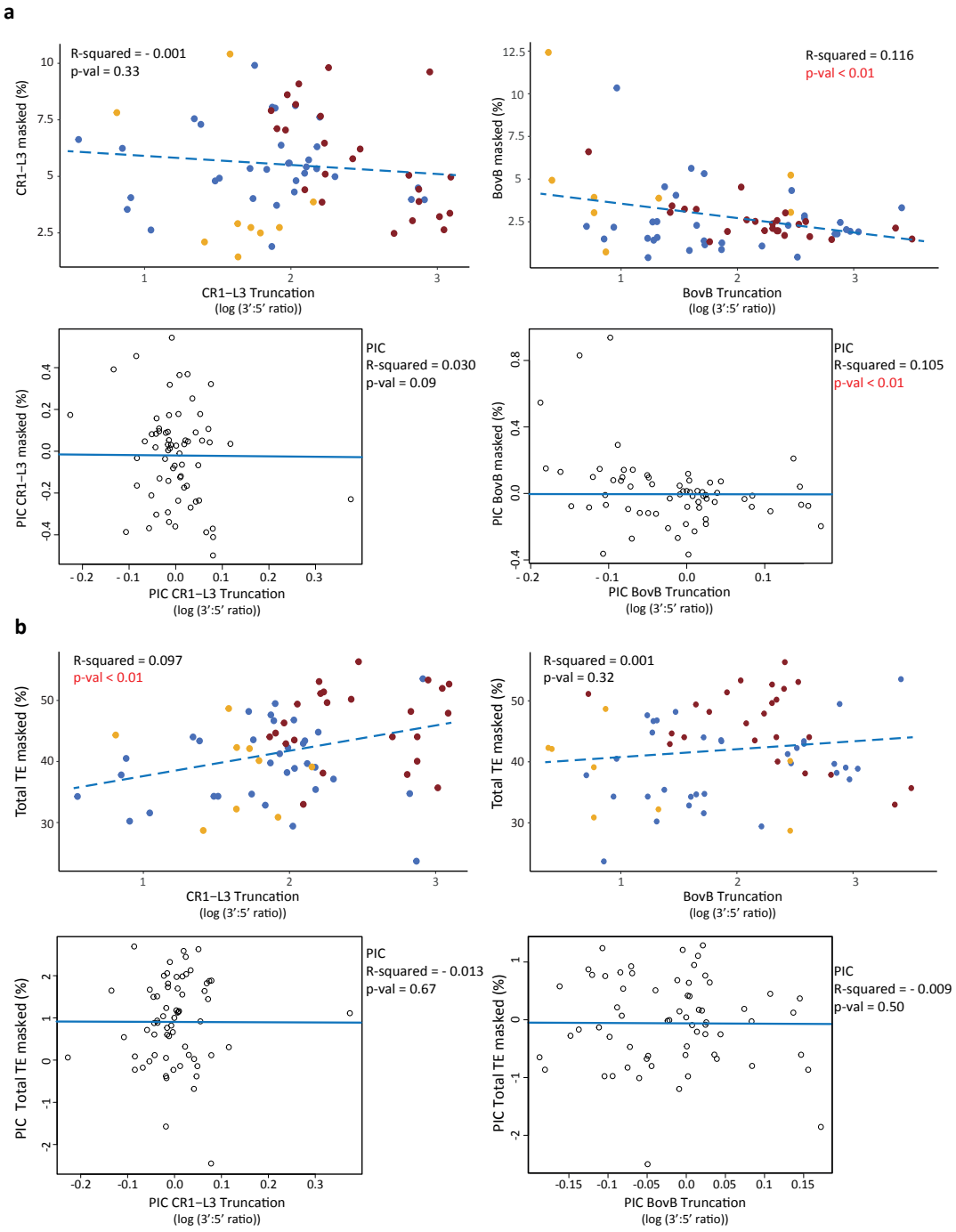
Supplementary Figure 10a and b. Effective population size (N_e) changes over time. **a**) Box plot shows the distribution of effective population size over time inferred from each genome. The first and the last time points were excluded from the input dataset. **b**) PSMC estimates of the changes in N_e over time inferred from each genome applying a generation time of 3 years and a mutation rate of 0.2×10^{-8}. Plots were generated using all time points per bootstrapped sample using the *psmc_plot.pl* script.



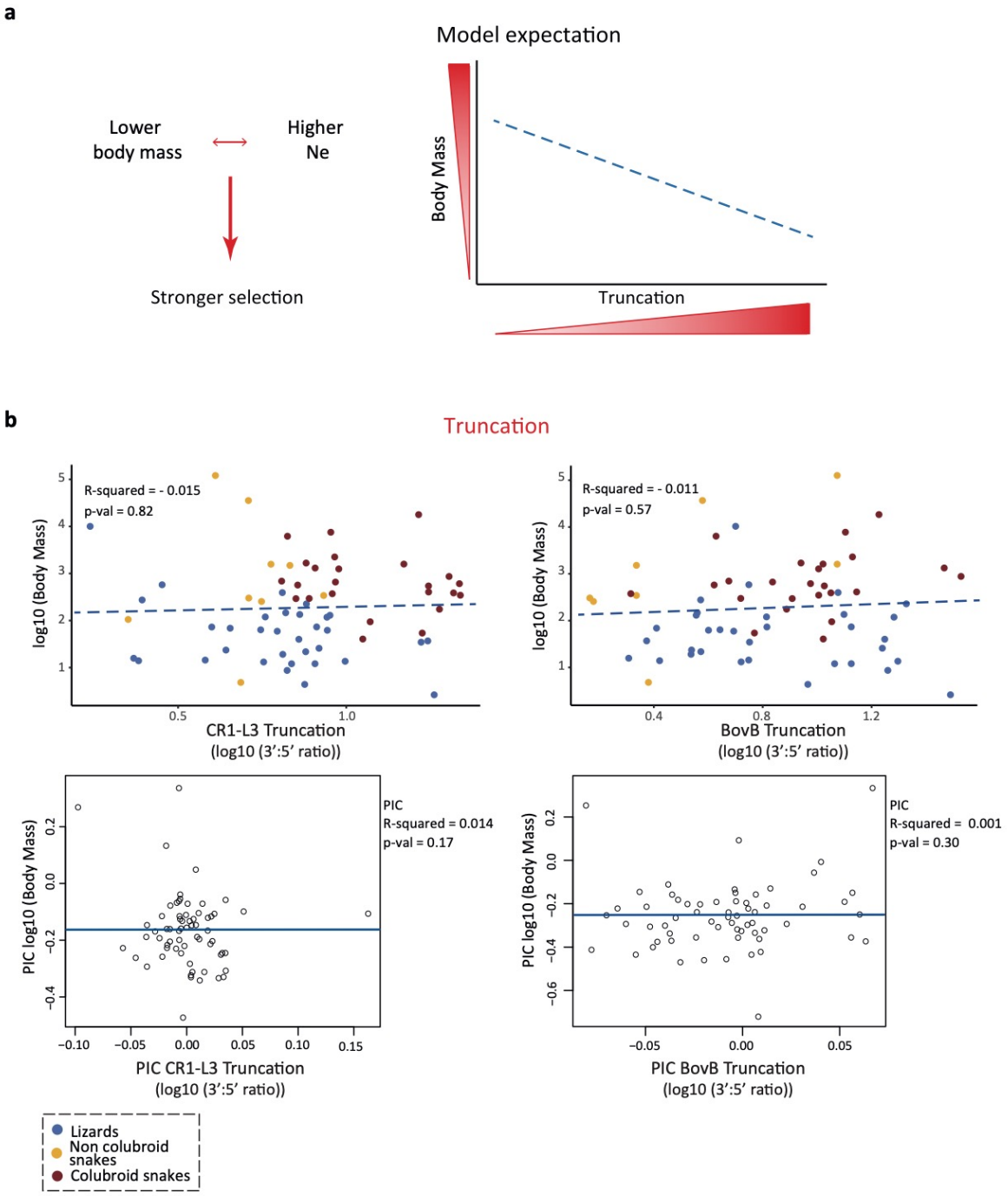
Supplementary Figure 10c. Relationship between effective population size (N_e) and truncation. Scatter plots and associated phylogenetically independent contrasts (PICs) show a lack of a correlation between median PSMC estimates of N_e and truncation of CR1-L3 LINEs (left) and BovB LINEs (right). Adjusted R-squared values and p-values were calculated using raw, untransformed data.



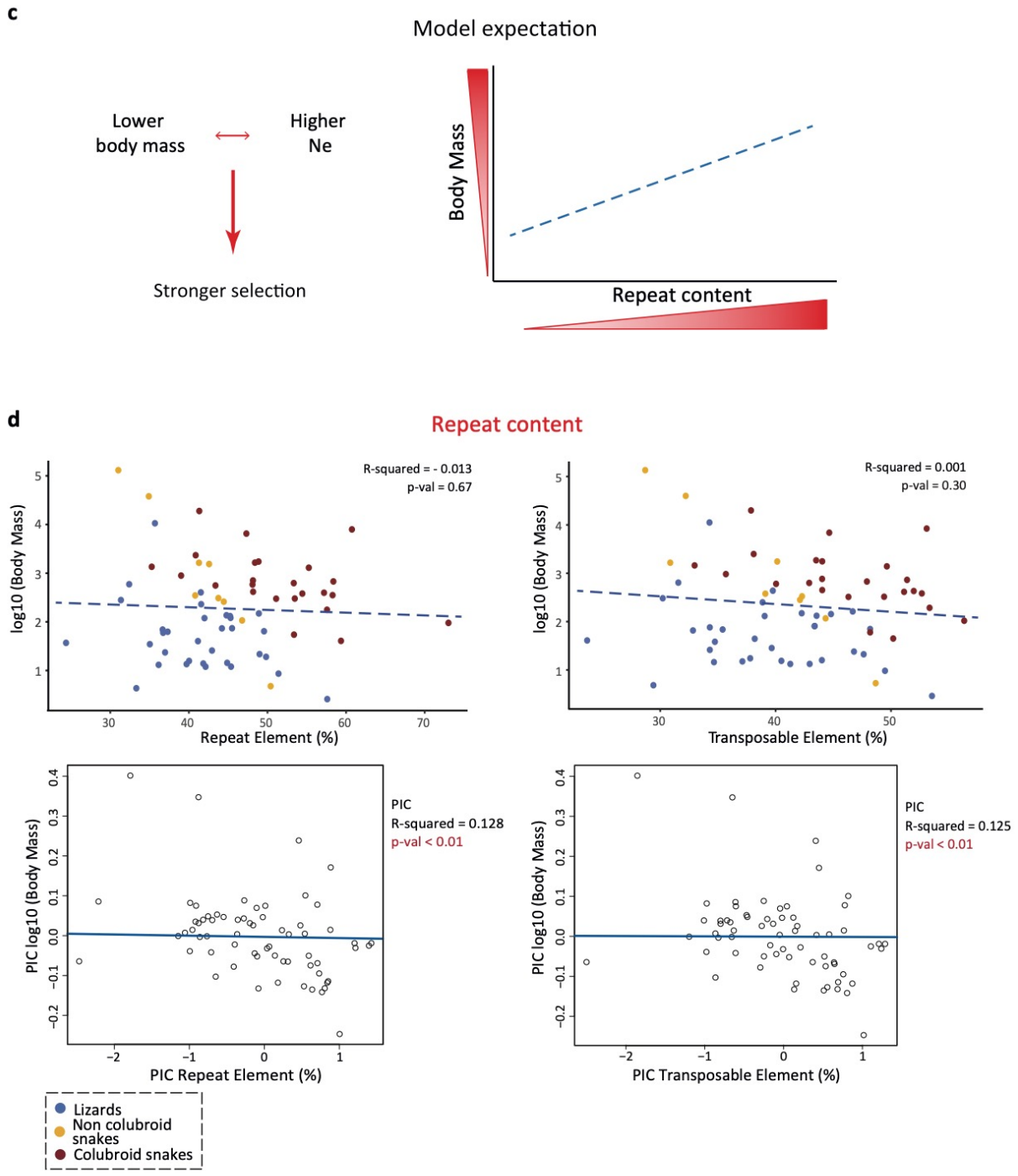
Supplementary Figure 10d and e. Relationship between effective population size (N_e) and genomic repeat content. **d**) Scatter plots and associated PICs reflect a lack of correlation between PSMC median estimates of N_e and total genomic frequency of CR1-L3 LINEs (left), and a significant positive correlation opposite to what would be expected between median N_e and BovB genomic content (right). **e**) Scatter plots and associated PICs reflect a lack of correlation between PSMC median estimates of N_e and total repeat element genomic content (left), and total TE content (right).



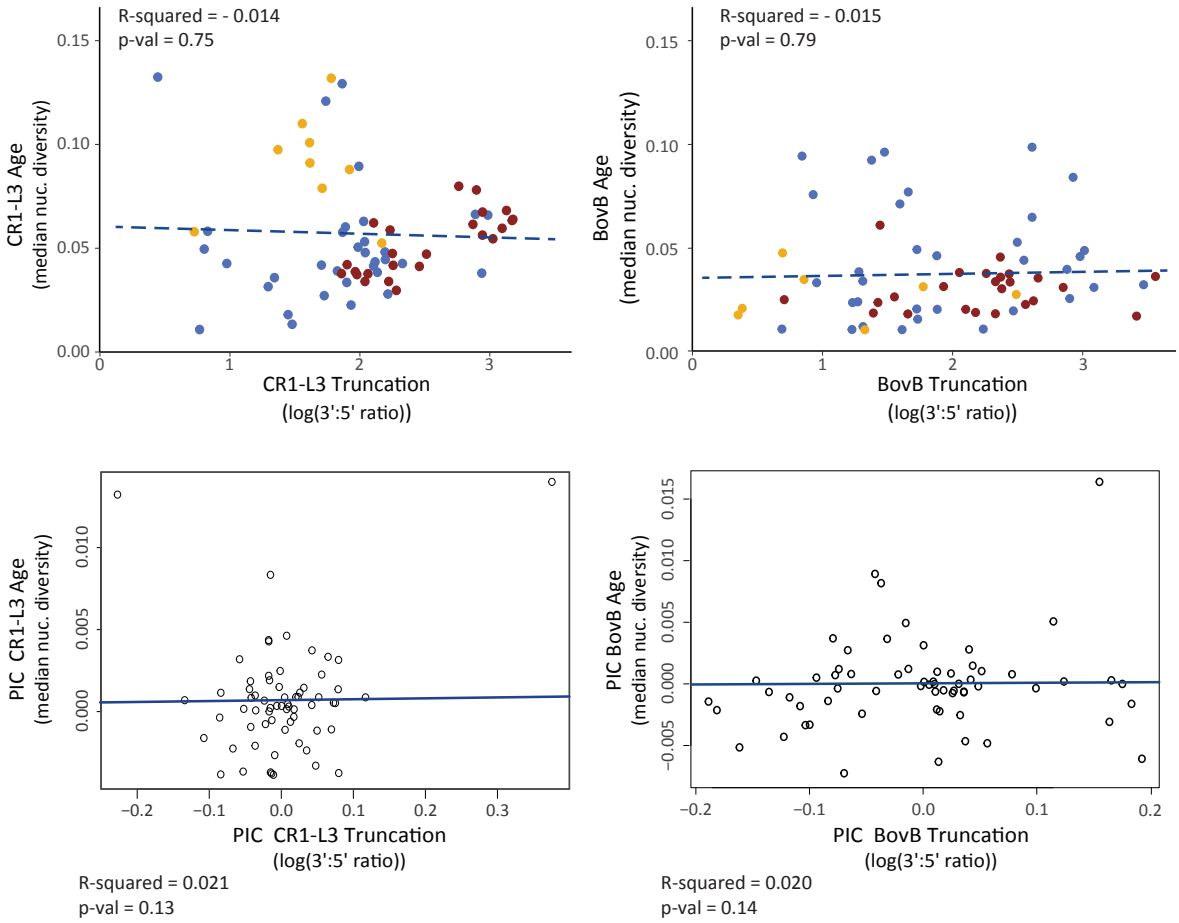
Supplementary Figure 11. TE truncation and genomic content. Scatter plots and associated phylogenetically independent contrasts (PICs) reflect a lack of correlation between degree of truncation and total genomic content of CR1-L3 LINEs (**a**, left), and a weak, although significant, negative correlation between degree of truncation and total genomic content of BovB LINEs (**a**, right). Additionally, we report a lack of correlation between total genomic TE content and the truncation level of BovB (**b**, right) across squamate reptile species. For CR1-L3, we report a positive correlation that contrast model expectations. Truncation values were log-transformed prior to perform statistical tests to meet the assumption of normality.



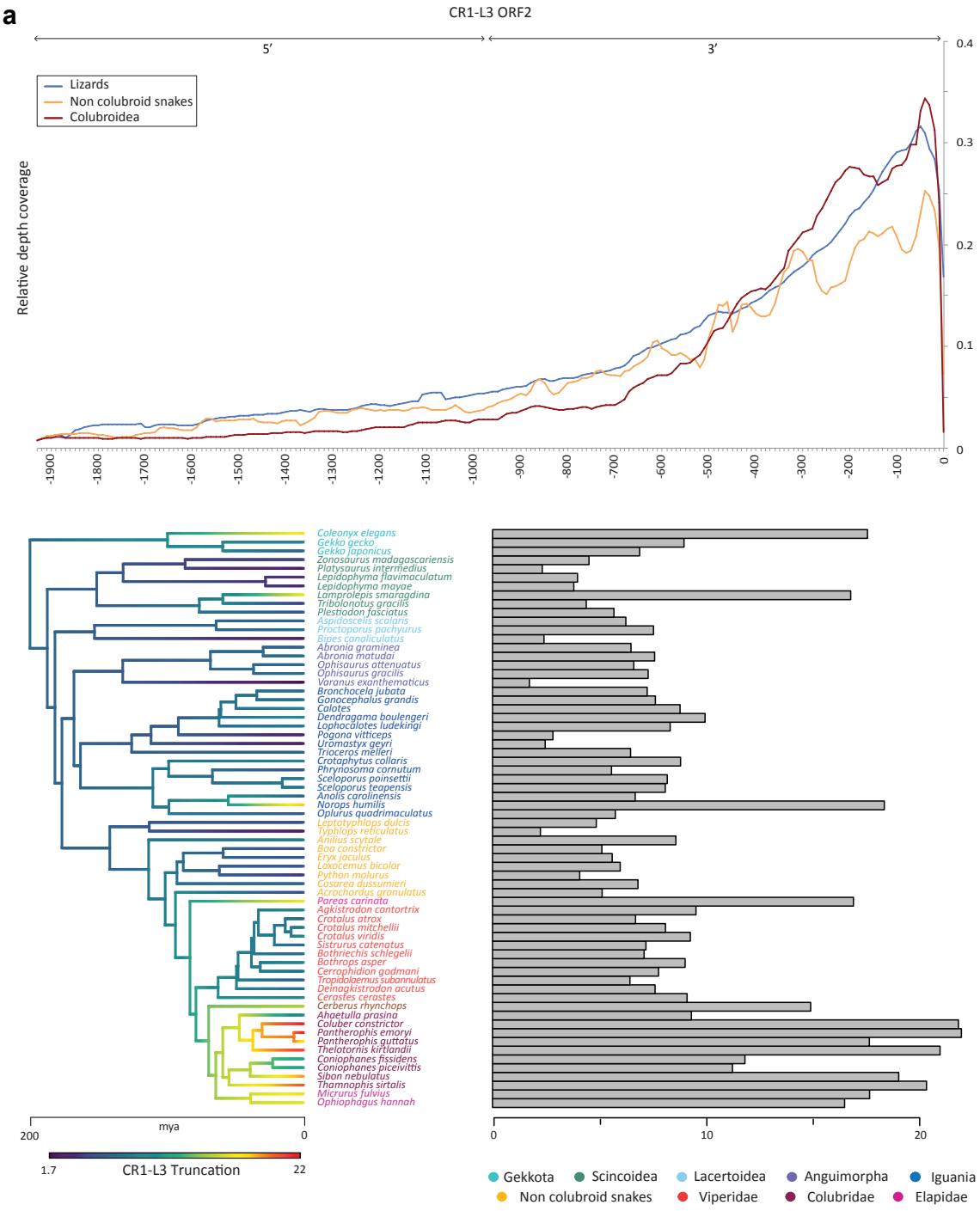
Supplementary Figure 12a and b. Relationship between adult body mass and truncation across 66 squamate species. **a**) Expected trends linking adult body mass (here used as proxy for effective population size) under a demographic explanation of the relationship between repeat element genomic abundance and TE copy length (truncation). **b**) Scatter plots and associated phylogenetically independent contrasts (PICs) show the lack of significant relationships between adult body mass and CR1-L3 (left) and BovB (right) truncation, which contrast with what would be expected under the demographic model. Truncation values were log-transformed prior to statistical tests to meet the assumption of normality.



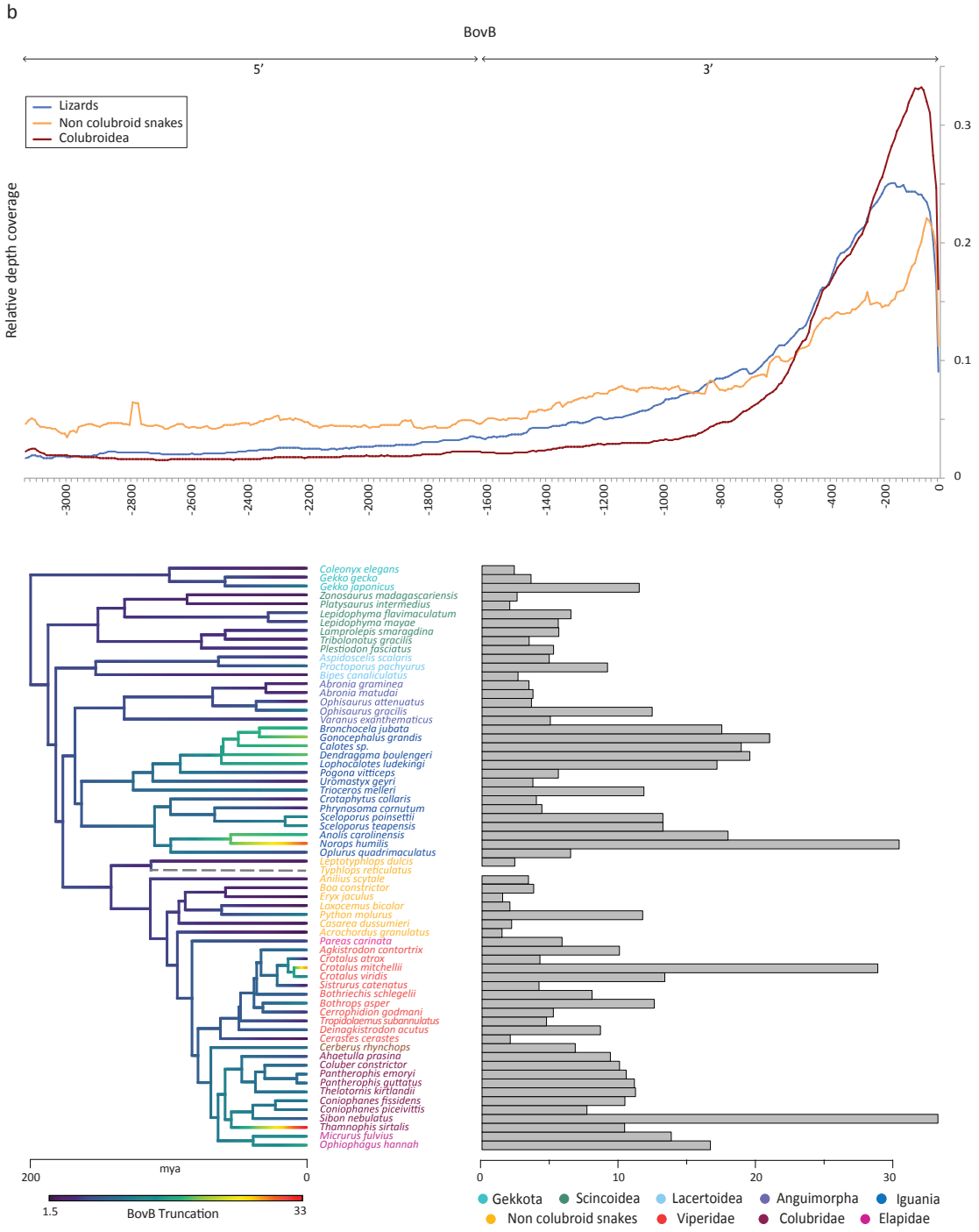
Supplementary Figure 12c and d. Relationship between adult body mass, repeat and TE content across 66 squamate species. **c**) Expected trends linking adult body mass (here used as proxy for effective population size) according to a demographic explanation of repeat element genomic abundance. **d**) Scatter plots and associated phylogenetically independent contrasts (PICs) show a significant relationship between adult body mass, total genomic repeat element content (left) and TE content (right), a trend opposite to what would be expected according to the model.



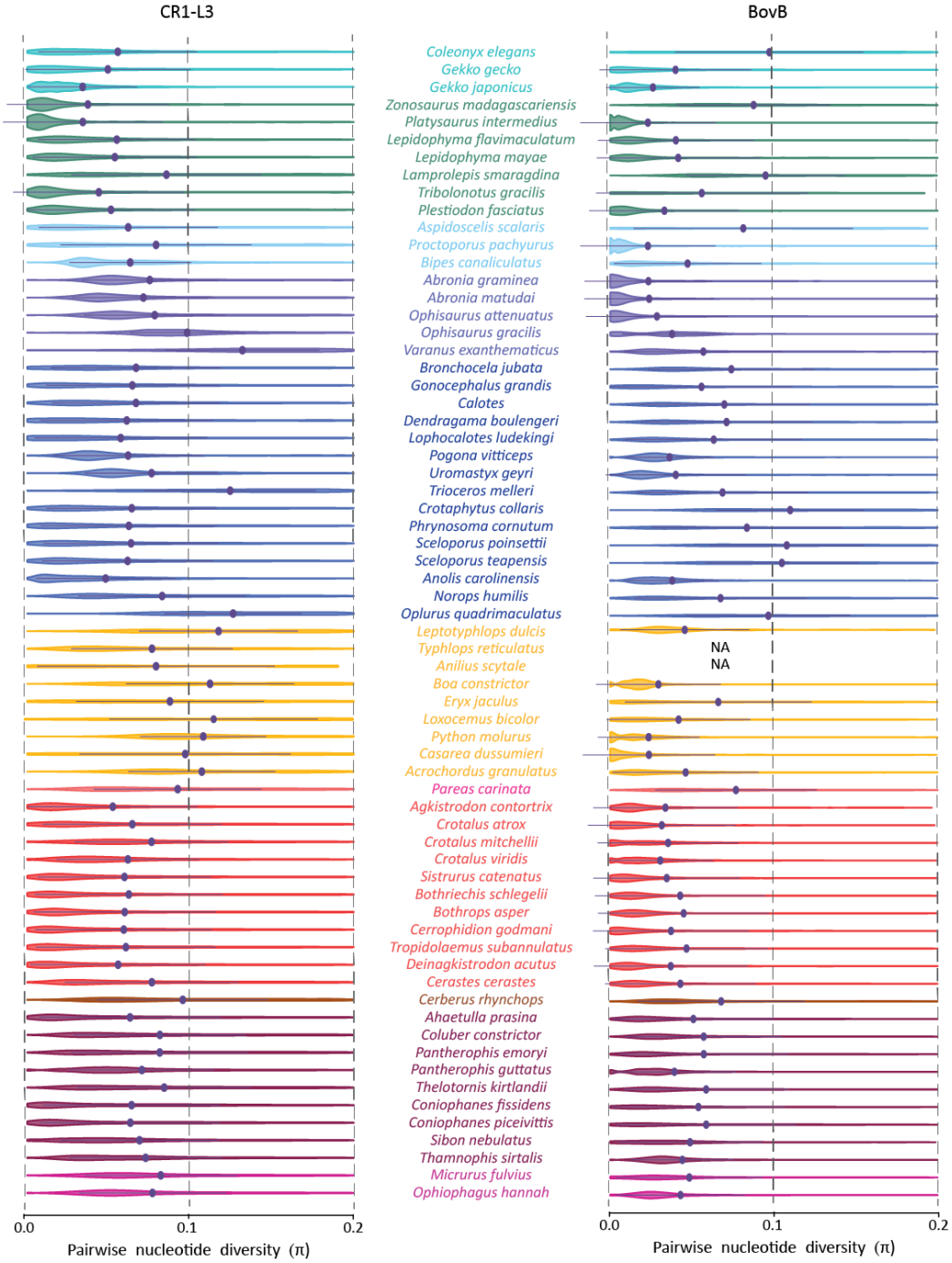
Supplementary Figure 13. Relationship between median sequence divergence of CR1-L3 and BovB LINEs and their genomic abundance. Species-specific consensus sequences were used as reference to calculate estimates of nucleotide divergence (pairwise π) for all alignable CR1-L3 (left) and BovB (right) sequences. We excluded sites that appeared to define subfamilies prior to pairwise π estimation, and calculated relative frequency including only sequences with $\pi < 0.2$. Scatter plots (top) and associated phylogenetically independent contrasts (PICs, bottom) show no correlation between the median pairwise nucleotide diversity (used as a proxy for element age) and genomic abundance of CR1-L3 and BovB LINEs. Truncation values were log-transformed prior to performing statistical analyses to meet the assumption of normality.
Typhlops reticulatus and *Anilius scytale* were not included in analyses of BovB LINEs (age data not retrievable).



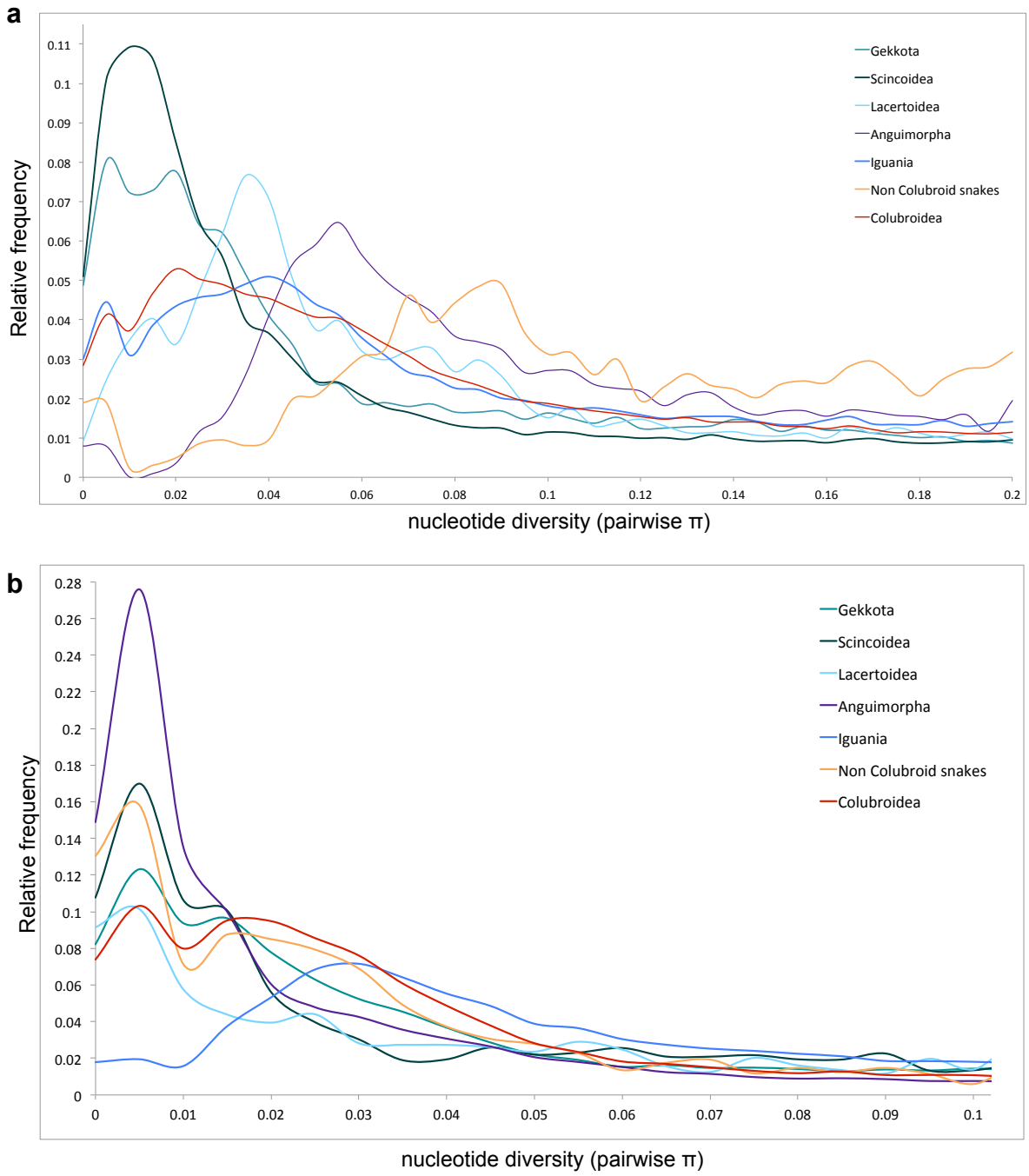
Supplementary Figure 14a. Observed truncation patterns of CR1-L3 LINEs across 66 squamate species. Top: line graph shows the average relative coverage of mapped reads along the ORF2 of CR1-L3 LINE consensus sequences for each major squamate group. Position is reported relative to the 3'end of the reference. Bottom: horizontal bar plot reports the extent of truncation for each squamate species sampled, calculated as the ratio of read depth coverage of reads mapping to the second half (3') and to the first half (5') of the consensus. Branches on the time-calibrated phylogeny are colored accordingly the degree of truncation.



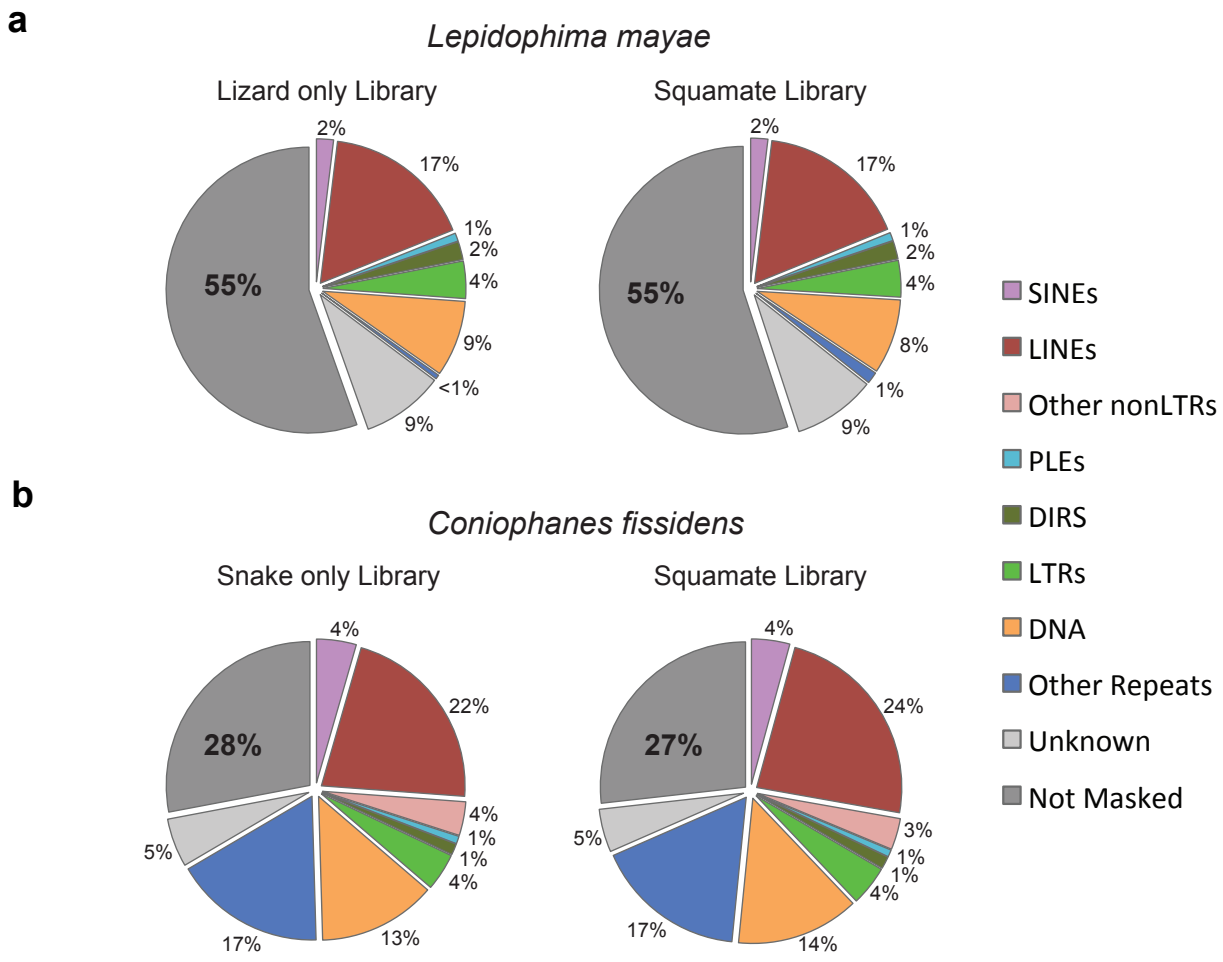
Supplementary Figure 14b. Observed truncation patterns of BovB LINEs and lineage-specific evolutionary rates across 66 squamate species. Top: line graph shows the average relative coverage of mapped reads along BovB consensus sequences for each major squamate group. Position is reported relative to the 3'end of the reference. Bottom: horizontal bar plot reports the extent of truncation for each squamate species sampled, calculated as the ratio of read depth coverage of reads mapping to the second half (3') and to the first half (5') of the consensus. Branches on the time-calibrated phylogeny are colored accordingly the degree of truncation.



Supplementary Figure 15. Estimated sequence divergence of CR1-L3 and BovB LINEs. Species-specific consensus sequences were used as reference to calculate estimates of divergence levels (pairwise π) for all alignable CR1-L3 (left) and BovB (right) sequences. We excluded sites that appeared to define subfamilies prior to pairwise π estimation, and calculated relative frequency including only sequences with $\pi < 0.2$. Violin plots show the empirical distribution of sequence divergence from the consensus for each species according to a kernel density estimation (KDE). Species with higher frequency of copies on the leftmost side of the violin plot are characterized by more recent amplification of the TE subfamily. Data that satisfies the filtering parameters are missing for *Typhlops reticulatus* and *Anilius scytale* BovBs.



Supplementary Figure 16. Estimated sequence divergence of CR1-L3 and BovB LINEs. Species-specific consensus sequences were used as reference to calculate estimates of divergence levels (pairwise π) for all alignable CR1 (**a**) and BovB (**b**) sequences. We excluded sites that appeared to define subfamilies prior to pairwise π estimation, and calculated relative frequency including only sequences with $\pi < 0.2$. Line graphs show the average sequence frequency distribution for each major squamate clade with detailed taxonomic organization of lizard species. Clades with higher relative frequency of copies on the leftmost side of the plot are characterized by a more recent amplification of the TE subfamily (e.g., Gekkota and Scincoidea for CR1-L3, and Anguimorpha for BovB).



Supplementary Figure 17. Masking strategy comparison for two squamate species. Pie charts report the repeat element genomic composition for (**a**) one lizard -*Lepidophyma mayae*- and (**b**) one snake -*Coniophanes fissidens*- species (randomly chosen for the comparative analysis). Charts on the left report masking results performed in RepeatMasker using a lineage specific repeat element reference library, whereas charts on the right reflect data generated using a single squamate repeat element reference library. Given that the two strategies brought to no difference in the masking quality, we consistently used the first approach (lineage-specific library) for within lineage consistency and optimization of RepeatMasker running time.

Chapter 3

The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes

Drew R. Schield[1], Daren C. Card[1], Nicole R. Hales[1], Blair W. Perry[1], Giulia M. Pasquesi[1], Heath Blackmon[2], Richard H. Adams[1], Andrew B. Corbin[1], Cara F. Smith[3], Balan Ramesh[1], Jeffery P. Demuth[1], Esther Betrán[1], Marc Tollis[4], Jesse M. Meik[5], Stephen P. Mackessy[3], and Todd A. Castoe[1, §]

1. Department of Biology, University of Texas at Arlington, Arlington, TX, USA

2. Department of Biology, Texas A&M University, College Station, TX, USA

3. School of Biological Sciences, University of Northern Colorado, Greeley, CO, USA

4. School of Life Sciences, Arizona State University, Tempe, AZ, USA

5. Department of Biological Sciences, Tarleton State University, Stephenville, TX, 76402 USA

Abstract

Here we use a chromosome-level genome assembly of a prairie rattlesnake (*Crotalus viridis*), together with Hi-C, RNA-seq, and whole genome resequencing data, to study key features of genome biology and evolution in reptiles. We identify the rattlesnake Z chromosome, including the recombining pseudoautosomal region, and find evidence for partial dosage compensation driven by an evolutionary accumulation of a female-biased upregulation mechanism. Comparative analyses with other amniotes provides new insight into the origins, structure, and function of reptile microchromosomes, which we demonstrate have markedly different structure and function compared to macrochromosomes. Snake microchromosomes are also enriched for venom genes, which we show have evolved through multiple tandem duplication events in multiple gene families. By overlaying chromatin structure information and gene expression data we find evidence for venom gene-specific chromatin contact domains, and identify how chromatin structure guides precise expression of multiple venom gene families. Further, we find evidence for venom gland-specific transcription factor activity, and characterize a complement of mechanisms underlying venom production and regulation. Our findings reveal novel and fundamental features of reptile genome biology, provide insight into the regulation of snake venom, and broadly highlight the biological insight enabled by chromosome-level genome assemblies.

Introduction

Squamate reptiles have become important models for a broad range of research, including studies on genome structure (Alfoldi et al. 2011), coevolution (Geffeney et al. 2002), development (Cohn and Tickle 1999), and regenerative biology (Secor and Diamond 1998). Among squamates, snakes represent an enriched system for studying a number of extreme or unique biological features. For example, snakes are an emerging model system for studying sex chromosome evolution, given their lack of apparent global dosage compensation (Vicoso et al. 2013), independent origins of ZW and XY sex determination systems (Gamble et al. 2017), and wide range of differentiation between sex chromosomes among lineages (Matsubara et al. 2006). Snakes also possess microchromosomes, which have been shown in birds to have intriguing and unique genome biology (Hillier et al. 2004; Backstrom et al. 2010), but are virtually uncharacterized in reptiles. Snake venom systems are the most intensely studied feature of snake biology due to their medical relevance (Mackessy 2010), and also because they provide a unique opportunity to study the evolution of a complex phenotype that required gene duplication, shifts in gene function and regulation, and numerous structural and physiological adaptations for venom storage and delivery. Although numerous studies have characterized the composition and activity of snake venoms, progress in understanding the genomic context for venom evolution and precise cellular and regulatory mechanisms underlying venom expression has been severely limited by the fragmentary nature of existing snake genome assemblies (Bradnam et al. 2013; Castoe et al. 2013; Vonk et al. 2013; Yin et al. 2016).

Here we leverage a chromosome-level assembly of the genome of the prairie rattlesnake (*Crotalus viridis*), assembled using a combination of second-generation sequencing and Hi-C scaffolding (Lieberman-Aiden et al. 2009), to study key questions about reptile and snake genome biology that have been previously difficult to address due to the fragmentary genome assemblies available for reptile species. We trace patterns of chromosome-level synteny and composition across amniotes, specifically exploring synteny between reptile and avian genomes and testing hypotheses about the evolution of GC-isochore structure in reptiles. We further characterize genome-wide chromatin contacts using Hi-C data to demonstrate differences between classes of chromosomes, and distinctions from patterns observed in mammalian datasets. Rattlesnakes have highly-differentiated ZW sex chromosomes (Matsubara et al. 2006), and we use our genome and additional resequenced genomes to identify the Z chromosome, the pseudoautosomal region of Z and W chromosomes, and an evolutionary stratum in the process of degeneration. We further studied patterns of partial dosage compensation and used inferred ancestral genome-wide expression levels to characterize the evolution of dosage compensation in snakes. Lastly, we use a combination of Hi-C chromatin contact data from the rattlesnake venom gland, RNA-seq data from diverse tissues, and the chromosomal locations of snake venom gene families to identify mechanisms of venom gene regulation in the venom gland.

Results

Genome assembly and annotation

We sequenced and assembled a rattlesnake reference genome from a male prairie rattlesnake (*Crotalus viridis viridis*) that was sequenced at 1,658-fold physical coverage using multiple approaches including the Dovetail Genomics HiRise sequencing and assembly method (Putnam

et al. 2016) that combines Chicago (Putnam et al. 2016; Rice et al. 2017) and Hi-C (Lieberman-Aiden et al. 2009) data, yielding a final scaffold length of 1.34 Gbp (Supplemental Fig. S1, Supplemental Tables S1, S2). Our annotation, which incorporated data from 24 RNA-seq libraries (Supplemental Table S3), included 17,352 protein-coding genes and an annotated repeat element content of 39.49% (Supplemental Tables S4, S5). Macrochromosomes were matched to scaffolds based on scaffold size and known chromosome-specific markers (Matsubara et al. 2006; Supplemental Table S6). Of six chromosomal markers from (Matsubara et al. 2006) that did not map to predicted chromosomes in our rattlesnake assembly, we were able to corroborate the accuracy of our assembly for five using multiple lines of evidence, including cross-species synteny with *Anolis* and local Hi-C contact frequencies (Supplemental Methods, Supplemental Table S7, Supplemental Fig. S2). We also identified the rattlesnake Z Chromosome using multiple lines of evidence, which we discuss below. In our preliminary assembly, microchromosomes were over-assembled into a single large scaffold, which we manually split based on multiple lines of evidence (see below and Supplementary Methods). The refined assembly had microchromosome scaffolds with lengths matching the predicted sizes of rattlesnake chromosomes (Baker et al. 1972). Our chromosome-level scaffolds include assembled telomeric and centromeric regions, with centromeres containing an abundant 164 bp monomer (Supplemental Fig. S3).

Synteny and chromosomal composition

The rattlesnake microchromosomes contain higher and more variable GC content than do the macrochromosomes, and have particularly high gene density (Welch's two-sample *t*-test on 100 kb windows, p-value < 0.00001) and reduced repeat element content compared to

macrochromosomes (Welch's two-sample *t*-test, $p < 0.00001$; Fig. 1A). These patterns are similar to those in the chicken (Supplemental Fig. S4). Rattlesnake chromosomes are highly syntenic with those from *Anolis*, except for fusion/separation of *Anolis* Chromosome 3 into rattlesnake Chromosomes 4 and 5 (Fig. 1B, Supplemental Fig. S5). Our synteny inferences also confirm that *Anolis* Chromosome 6 is homologous to the sex chromosomes of rattlesnakes (Srikulnath et al. 2009). Despite conservation of squamate microchromosome homology, patterns of chicken-squamate homology suggest that there were major shifts between macro- and microchromosome locations for large syntenic regions early in amniote evolution. We find evidence for multiple macrochromosomal shifts in synteny between the chicken and squamate reptiles, some of which appear quite complex. For example, the chicken Chromosomes 1 and 2 show synteny patterns that are scattered across multiple squamate macrochromosomes, including the rattlesnake Z Chromosome. Furthermore, only about half of chicken microchromosomes are syntenic with squamate microchromosomes (Fig. 1B), while the rest of chicken microchromosomes share synteny with squamate macrochromosomes. Despite independent origins of some avian and squamate microchromosomes, there are broad similarities among squamate and avian microchromosomes (e.g., GC content variation, gene density; Supplemental Fig. S4). Further, the presence of microchromosomes in most extant diapsids (Olmo 2005; Organ et al. 2008), the ancestral diapsid genome (O'Connor et al. 2018), amphibians (Voss et al. 2011), and fish (Braasch et al. 2016) broadly suggest that the majority of vertebrate evolution has been shaped by the distinctive, but poorly understood biology of microchromosomes.

GC-isochore and repeat element evolution

Squamate reptiles have become important models for studying the evolution of genomic GC content and isochore structure, due to the loss of GC isochores in *Anolis* yet the apparent re-emergence of isochore structure in snakes (Fujita et al. 2011; Castoe et al. 2013). Comparisons of orthologous genomic regions across 12 squamates demonstrates that there have been two major transitions in genomic GC content, including a reduction in GC content from lizards to snakes, and a further reduction in GC content within the colubroid snake lineage that includes the rattlesnake and cobra (Fig. 1C). This suggests that higher genome-wide GC content was likely the ancestral squamate condition, and that snakes have evolved increased nucleotide composition variation through an increase in genomic AT content, rather than a buildup of GC-rich isochores. Based on studies of mammals (Duret and Galtier 2009), GC isochore structure is thought to be driven mainly by GC-biased gene conversion that results in GC-biased allele substitution in some genomic regions. The negative relationship between genomic GC content (Fig. 1C) and GC isochore structure (Fig. 1D; Supplemental Table S8) across squamate evolution indicates that this explanation may not apply to the apparent trends in snakes. Instead, GC content variation in snakes appears to be driven by AT-biased processes, including AT-biased substitution that was suggested by previous comparisons of lizard and snake genomes (Castoe et al. 2013). Similar to the patterns observed in GC content variation, genomic repeat element content has also undergone a major shift in colubroid snakes, which show substantial increases in transposable elements overall, and specific increases in hAT and Tc1 DNA elements, CR1-L3 LINES, and simple sequence repeats (SSRs; Fig. 1E, Supplemental Fig. S6). It remains an open question, however, if shifts in GC content and genomic repeat landscapes are related in colubroid snakes (see also Pasquesi et al. 2018).

Sex chromosome evolution

Snake sex chromosomes have evolved multiple times, apparently from different autosomal chromosomes (Gamble et al. 2017), and colubroid Z/W Chromosomes are homologous with *Anolis* Chromosome 6 (Srikulnath et al. 2009; Vicoso et al. 2013). We identified a single 114 Mb scaffold as the rattlesnake Z Chromosome that contains known Z-linked genes (Matsubara et al. 2006; Supplemental Table S6), and demonstrates roughly half female (ZW) versus male (ZZ) mapped genomic read coverage based on additional male and female samples we sequenced (Fig. 2A; Supplemental Fig. S7; Supplemental Table S9). We also identified the recombining pseudoautosomal region (PAR) of the Z Chromosome as the distal 7.2 Mb region that shows equal male-female genomic read coverage (Fig. 2A). The PAR is GC-rich relative to the genomic background and the remaining Z chromosome (42.9%; Supplemental Fig. S8), similar to the PAR of the collard flycatcher (Smeds et al. 2014), suggesting that common processes may drive increased PAR GC content in independently evolved snake and avian sex chromosomes. The rattlesnake PAR also exhibits distinctive patterns of repeat element content (Supplemental Fig. S9) and a higher density of genes than the remaining Z Chromosome (Fisher's exact test: $p = 4.46 \times 10^{-7}$; Supplemental Fig. S10). Adjacent to the PAR, we identified an evolutionary stratum ('Recent Stratum') that shows near-autosomal female genomic read coverage (Fig. 2A, top panel). We hypothesize that recombination was most recently suppressed in this region, and that substantial homology is retained between Z and W Chromosomes. Consistent with this hypothesis, we observe elevated nucleotide diversity (π) across this region specifically in females (Fig. 2A, Supplemental Fig. S11), likely due to reads mapping to divergent Z and W-linked gametologs. These results suggest that a number of W-linked gametologs have either been retained during Z/W divergence, or are still in the process of degeneration, as has been suggested

for birds (Bellott et al. 2017). To further understand the evolutionary origins of the Recent Stratum, we compared mappings of female and male resequencing data for the Prairie Rattlesnake with those from the Pygmy Rattlesnake (*Sistrurus catenatus*) and Five Pace Viper (*Deinagkistrodon acutus*). Both species exhibit similar patterns of intermediate female normalized coverage across the Recent Stratum (Supplemental Fig. S7), suggesting that this evolutionary stratum evolved prior to the divergence between the Prairie Rattlesnake and Five Pace Viper greater than 30 million years ago (Zheng and Wiens 2016). Collectively, the features of the Recent Stratum suggest that recombination suppression and degeneration are ongoing processes in pitvipers, despite the already high differentiation between Z and W Chromosomes (Matsubara et al. 2006).

Patterns of gene expression between heterogametic and homogametic sexes in organisms with differentiated sex chromosomes are of broad interest because of the diversity of mechanisms that can result in dosage compensation (Graves 2016). To investigate dosage compensation in the rattlesnake we compared female and male RNA-seq data from kidney and liver tissues across the rattlesnake Z Chromosome (Supplemental Table S9). We find evidence from both tissues for lower overall expression in the female (Fig. 2B, left panel; Supplemental Fig. S12), consistent with previous conclusions that female colubroids lack complete dosage compensation (Vicoso et al. 2013), but also that this ratio is higher than expected if there were no dosage compensation (i.e., $\log_2$ female/male expression > -1, Wilcoxon signed-rank tests, p-values < 2.2×10^{-16}). We also find that chromosome-wide gene expression is higher on the Z than on autosomes for males (Mann-Whitney *U* tests, p-values < 0.0002), yet lower on the Z than on autosomes for females (p-values < 0.02; Fig. 2B). Consistent with this, the Z is also enriched for male-biased genes, and

depauperate in female-biased genes, relative to autosomes (Supplemental Fig. S13; Fisher's exact tests, p-values $< 2 \times 10^{-5}$).

To understand how patterns of gene expression on the Z have evolved, we compared current Z gene expression in our rattlesnake samples to inferred ancestral (i.e., proto-Z) expression, based on expression levels in autosomal orthologs of the rattlesnake Z genes in the anole and chicken (following Julien et al. 2012; Marin et al. 2017). We find that current male Z expression has not changed from the inferred male proto-Z expression level (Fig. 2C; Supplemental Fig. S12), but that current female Z expression is lower than ancestral female expression (Mann-Whitney *U* tests, p-values < 0.005). This finding suggests that female Z expression diminished after the establishment of sex chromosomes in the rattlesnake. Combined with evidence that current male Z expression is higher than autosomes, these findings raise the question of whether ancestral expression levels predisposed the proto-Z (e.g., *Anolis* Chromosome 6) to become the rattlesnake Z. We addressed this by comparing inferred ancestral Z and autosomal expression (Fig. 2C), and find that the ancestor of the rattlesnake Z shows higher expression in both sexes than ancestral autosomes (Mann-Whitney *U* tests, p-values < 0.02). These findings suggest that, due to the enrichment of male-specific function and the overall elevated level of expression, characteristics of the rattlesnake Z ancestor may have favored its transition from autosome to sex chromosome.

No mechanisms underlying partial dosage of genes or regions have been identified in snakes. The ratio of female/male gene expression is regionally variable across the rattlesnake Z, suggesting partial dosage compensation driven by regional or gene-specific mechanisms (Fig. 2A, bottom panel). We hypothesized that an inherently female-biased regulatory mechanism, estrogen response elements (EREs), might explain dosage compensated regions, and tested for a

relationship between the ratio of female/male expression and the number of predicted EREs in 100 kb windows of the Z chromosome. There is a positive relationship between ERE density and female/male expression for rattlesnakes on the rattlesnake Z (Fig. 2D), yet we do not find this relationship for the analogous comparison of *Anolis* female/male expression and ERE density on *Anolis* Chromosome 6 (Supplemental Fig. S14). We also find that that the rattlesnake Z Chromosome has a much higher density of EREs than *Anolis* Chromosome 6 (two-sample Z test, p-value $< 2.2 \times 10^{-16}$; Fig. 2D) and is enriched for EREs compared to the genomic background (Fisher's exact test, p-value $< 2.2 \times 10^{-16}$), despite a much higher density of EREs in the *Anolis* genome overall. To further understand if ERE accumulation is a general feature of snake Z Chromosome evolution, we also analyzed Z Chromosome and autosomal sequences of the Five Pace Viper, and find consistent evidence for ERE enrichment on the Z Chromosome compared to *Anolis* Chromosome 6 (Supplemental Fig. S15; Fisher's exact test, p-value = 0.00016) Our results illustrate that the evolution of the pitviper Z Chromosome has involved regional accumulation of EREs, which may be an important mechanism underlying regional dosage compensation.

Hi-C exposes unique microchromosome biology

Our analyses of the first chromatin contact data for a non-mammalian vertebrate (Fig. 3A) demonstrate broad similarities in chromatin structure across vertebrate macrochromosomes, yet unique features of snake microchromosomes. We find that patterns of intra- and interchromosomal chromatin contacts across rattlesnake macrochromosomes are consistent with patterns observed in mammals, such that when interchromosomal contact frequencies are normalized by chromosome length, they show a consistent negative linear relationship across

species (Fig. 3B). Rattlesnake microchromosomes deviate significantly from this macrochromosomal pattern, and share disproportionately high frequencies of contacts with other chromosomes, including other microchromosomes (Fig. 3A-B). Indeed, the initial over-assembly (Supplemental Fig. S16) of microchromosomes into a single scaffold was likely driven by these unexpected high contact frequencies among microchromosomes, which significantly exceed assumptions used for assembly that are based on mammalian macrochromosomes (*t*-test, $p < 0.000001$). These findings highlight the uniqueness of microchromosome interactions within the nucleus of the rattlesnake venom gland, and beg the question of whether distinctive chromatin contacts are a consistent feature of microchromosomes in other amniotes.

Venom evolution and regulation

While numerous studies have characterized the diversity of venom composition among snake species (e.g., Mackessy 2008; Casewell et al. 2009; Casewell et al. 2012; Rokyta et al. 2012), the chromosomal location of venom genes and mechanisms underlying the regulation of venom remain poorly understood. Our rattlesnake genome provides the genomic location and context for snake venom genes (Fig. 4A; Supplemental Fig. S17; Supplemental Tables S10, S11) and demonstrates that microchromosomes are enriched for these genes (i.e., 37% of all venom genes are found on microchromosomes which represent 10% of the genome; Fisher's exact test, $p = 0.0017$; Fig. 4A). Moreover, microchromosome-linked venom gene families include three of the most abundant and well-characterized components of rattlesnake venom (Fig. 4A, snake venom metalloproteinases, *SVMPs*; snake venom serine proteinases, *SVSPs*; and type IIA phospholipases A2, *PLA2s*) – each of these families is located on a different microchromosome (Fig. 4A; Supplemental Fig. S17). The other major component of prairie rattlesnake venom,

myotoxin (crotamine), is located on Chromosome 1 (Fig. 4A). To identify patterns of venom gene family evolution we conducted phylogenetic estimates of each of the microchromosome-linked families listed above (including non-venom paralogs). We inferred that each venom family represents a distinct set of tandemly-duplicated genes derived from a single ancestral homolog that gave rise to a monophyletic cluster of venom paralogs (Supplemental Figs. S18, S19). While this has been proposed previously (Ikeda et al. 2010; Vonk et al. 2013), the contiguity of our genome provides new definitive evidence that this duplicative mechanism explains the origin of multiple unlinked snake venom gene clusters.

The depletion of venom is followed by the rapid expression, synthesis, and storage of proteins in the venom gland lumen over the course of several days. To investigate the regulation of venom production we compared gene expression between venom glands and body tissues, and identified a set of 12 transcription factors (TFs) with significantly higher expression in the venom gland (Fig. 4B; Supplemental Fig. S20; Supplemental Table S12). Many of these TFs were linked to the secretory demands of the venom gland (e.g., the unfolded protein response of the endoplasmic reticulum: *ATF6* and *CREB3L2*) or repair of the glandular epithelium (e.g., *ELF5*). While the potential involvement of these TFs in regulating venom production cannot be entirely ruled out, we did not find evidence of predicted binding sites that would suggest a role in directly regulating venom genes (Supplemental Table S12). Five transcription factors, however, stood out as candidates for regulating venom gene expression based on their known regulatory functions, links to established mechanisms of venom production, and the proximity of their predicted binding sites to venom genes (Fig. 4B).

Though neither TFs or transcriptional mechanisms regulating venom production have been precisely identified, there is evidence that following venom depletion, venom production is triggered by a1-adrenoceptors that activate the *ERK* signaling pathway (Kerchove et al. 2008). One of the venom-gland upregulated TFs was *GRHL1*, which is known to function in epidermal barrier formation and repair (Ting et al. 2005), and is regulated directly by *ERK* (Kim and McGinnis 2011). We also identified a set of four Nuclear Factor 1 (*NFI*) TFs, all of which share the same predicted binding site and are classified as RNA polymerase II core promoter binding TFs. *NFI* TFs are known to drive tissue-specific expression (Gronostajski 2000), and function in chromatin remodeling and transactivation (Fane et al. 2017). Predicted binding sites of *GRHL1* tend to occur in close proximity to venom genes (average within 79 kb of a venom gene), and predicted *NFI* binding sites are present in the promoter regions of a large proportion (~72%) of venom genes (Fig. 4C; Supplemental Fig. S19; Supplemental Tables S12, S13). We also found that genes flanking venom clusters (and lacking venom-specific expression) lacked *NFI* binding sites and were on average further (86 kb) away from the nearest *GRHL1* binding site; binding sites for either set of TFs were not, however, statistically enriched in venom gene regions compared to the genomic background (Supplemental Table S13, Supplementary Methods). The upregulation of *GRHL1* and *NFI* TFs upon venom depletion and the presence of their predicted binding sites in venom gene clusters suggests these TFs may play a direct role in the regulation of venom, although the distribution of their binding sites does not entirely explain variation in venom gene expression (e.g., Fig. 4C), suggesting other TFs and potentially other mechanisms also contribute to venom regulation.

Because our results indicated that the specificity of venom gene expression is not fully explained by venom-specific TF activity, we tested for evidence that venom is also regulated by specific chromatin structure and organization. We performed Hi-C sequencing of a 1-day post-extraction venom gland, which enabled us to capture chromatin contacts associated with venom production. Genomic regions containing venom clusters show a specific structure within discrete high-frequency chromatin contact regions, representing venom-specific topologically-associated chromatin domains (TADs (Dixon et al. 2016); Fig. 4D; Supplemental Fig. S21). These 'venom TADs' are flanked by predicted binding sites of *CTCF*, which coordinates DNA looping and insulates transcriptional activity. Consistent with our chromatin data, we find that genes flanking venom TADs exhibit varied expression profiles across tissues, while genes within venom TADs show high venom gland specificity (Fig. 4C-D), indicating a strong insulating regulatory effect of TAD boundaries surrounding venom cluster regions. Collectively, these findings suggest that venom gene regulation is driven by synergistic interactions between tightly-regulated chromatin structure and highly expressed TFs that are responsive to venom depletion.

Discussion

Our results provide new perspectives on the structure and function of amniote genomes, mechanisms and evolution of dosage compensation, and the biology and regulation of snake venom. These findings further demonstrate the potential for a new generation of well-assembled genomes to facilitate advances in our understanding of the diversity of genome biology across otherwise poorly characterized lineages of the tree of life. Much of what is known about reptile genome biology comes from studies of lizards and birds (e.g., Hillier et al. 2004; Warren et al. 2010; Alfoldi et al. 2011), thus a primary motivation of this study was to use the highly

contiguous rattlesnake genome to compare and contrast aspects of snake genome biology with those of other reptiles and amniotes. For example, studies of bird genomes have shown that avian microchromosomes are gene-rich, and therefore functionally important. Despite the semi-independent origins of microchromosomes in squamate reptiles and birds, snake microchromosomes exhibit many of the same compositional patterns (i.e., gene density, GC and repeat content) as microchromosomes in birds (Fig. 1). Moreover, as the first species with microchromosomes to be examined using Hi-C, we find that rattlesnakes microchromosomes exhibit fundamentally different patterns of chromatin contact, with proportionally higher inter-chromosomal contact frequencies than macrochromosomes in snakes or mammals (Lieberman-Aiden et al. 2009; Rao et al. 2015; Darrow et al. 2016); Fig. 3). This discovery highlights the unique structure and function of microchromosomes, and raises the question of whether the uniqueness of snake microchromosome chromatin structure is a feature common to all amniote vertebrate microchromosomes. Future analyses using Hi-C or other data to compare microchromosome structure and nuclear contact patterns will be key to address the generality of links between microchromosome structure, organization, and function across vertebrate lineages.

A major goal of comparative genomics is to understand the patterns and mechanisms that lead to the observed variation in genome structure and function across species. Previous comparative analyses have demonstrated unique patterns of genome structure and content in squamate reptiles that are distinct from those observed in other major amniote lineages (i.e., birds and mammals; Alfoldi et al. 2011; Fujita et al. 2011; Pasquesi et al. 2018). Our comparative analyses of 12 squamate genomes provide new insight and context for understanding the evolution of unique genomic features of squamates (Fig. 1C-E). For example, our results indicate that snakes have

re-evolved genomic GC-isochore structure while also evolving reduced overall genomic GC content. These confluence of these patterns raise the intriguing possibility that snake isochore structure has evolved not through an accumulation of GC content (i.e., GC-biased gene conversion) as observed in mammals and birds (Duret and Galtier 2009; Weber et al. 2014), but through the accumulation of AT content via AT-biased substitutions (Castoe et al. 2013) or other mechanisms. This observation in snakes, together with the extremely varied GC landscapes across squamates (Fujita et al. 2011, Castoe et al. 2013; Fig. 1C-D), raise a number of questions, including whether mechanisms outside of GC-biased gene conversion contribute to GC isochore structure in vertebrates, and whether GC-biased gene conversion plays a major role in squamate genome evolution.

Due to the independent origins of distinct sex determination systems (Gamble et al. 2017) and variation in differentiation between sex chromosomes among lineages (Matsubara et al. 2006; Vicoso et al. 2013), snakes have become an important model system for investigating sex chromosome evolution. Through our analyses of the rattlesnake Z Chromosome, we identified the recombining pseudoautosomal region of the highly-differentiated Z and W Chromosomes, and an evolutionary stratum bearing the hallmarks of recombination suppression and degeneration on the W Chromosome. These findings indicate that even through the rattlesnake Z and W are highly differentiated, further differentiation and recombination suppression between the Z and W are ongoing (Fig. 2). Despite the independent origins of Z/W Chromosomes in rattlesnakes and birds, there are similarities in the patterns of GC-richness of the pseudoautosomal regions of sex chromosomes in both lineages, suggesting that common

processes may drive increased pseudoautosomal region GC content across divergent amniote lineages.

Although previous studies have found evidence of a lack of global dosage compensation on the Z Chromosome in females (Vicoso et al. 2013; Yin et al. 2016), the evolution of gene expression and incomplete dosage compensation as the snake Z Chromosome evolved has not be studied. Our comparison of female and male Z Chromosome expression with inferred ancestral expression provide new evidence that, in comparison to the ancestral proto-Z autosome, male expression has remained largely constant, while female expression has become reduced after the establishment of the sex chromosomes (Fig. 2C). We also found that chromosome-wide gene expression on the proto-Z was higher in both sexes than on other autosomes, raising the possibility that autosomes with these expression characteristics may be more likely (e.g., predisposed) to become sex chromosomes. We further demonstrated high gene-specific or regional variation in dosage compensation in the rattlesnake, and provide the first report that a female-biased transcriptional regulatory mechanism that modulates expression in other reptiles (Rice et al. 2017), estrogen response elements, does explain some of the variation in dosage compensation across the Z Chromosome. Specifically, we found that the density of estrogen response elements positively correlates with female gene expression across the Z Chromosome (Fig. 2D), and that these elements have accumulated on the Z Chromosome following its divergence from its autosomal homolog (Chromosome 6) in the anole lizard. Evidence for ERE accumulation on the Z chromosome of the rattlesnake and the Five Pace Viper further indicate that ERE accumulation occurred early in the evolution of the snake Z Chromosome, and provides evidence for the potential role of EREs in dosage compensation in ZW systems.

Despite venom representing the most intensively studied feature of snake biology, previous fragmentary snake genome assemblies have provided limited genomic context for snake venom evolution and regulation. Leveraging the first chromosome-level genome assembly for a snake, our precise chromosomal localization of genome genes revealed that numerous important components of snake venom (Mackessy 2008) are located on snake microchromosomes (Fig. 4A), further underscoring the functional importance of snake microchromosomes. Our integrated analysis of venom systems provides new evidence for a role of *GRHL1* in venom gene regulation, thereby linking a transcriptional regulatory mechanism to a previously known regulatory stimulus (*ERK* signaling; Kim and McGinnis 2011) shown to trigger venom production (Kerchove et al. 2008). Analyses of Hi-C chromatin contact information from recently depleted venom glands provided new evidence for the tight regulation of chromatin in and around venom gene clusters, to the extent that venom genes occupy venom-specific topologically-associated domains (venom TADs) bounded by *CTCF* binding sites, and genes within versus outside the boundaries of these venom TADs show distinct expression profiles (Fig. 4). Collectively, our results provide new evidence for the coordinated roles of chromatin organization and transcription factor activity in the process of venom gene regulation.

Methods

Genome assembly and annotation

Animal procedures were conducted with approved and registered IACUC protocols. Chicago and Hi-C libraries were constructed from genomic DNA from the liver and venom gland of a single male *Crotalus viridis viridis*, and assembly was performed using the Dovetail Genomics HiRise v2.1.3-59a1db48d61f assembler. A previous version of this assembler (Putnam et al. 2016) is

available as an open-source distribution at https://github.com/DovetailGenomics/HiRise_July2015_GR), however, Dovetail Genomics has not made the HiRise version used on this assembly available as open source software at this time. Chicago and Hi-C data were used to improve an existing fragmentary assembly (CroVir2.0; NCBI accession SAMN07738522; Supplemental Tables S1-S2), which was constructed using multiple short-read sequencing libraries in combination with long-insert mate-pair libraries (Supplemental Table S1). Information about input assembly breaks and Chicago assembly scaffold joins can be found in Supplemental Material 2). Genomic DNA for these libraries was extracted from snap frozen liver tissue using standard phenol-chloroform-isoamyl DNA extraction methods. We generated 24 transcriptomic libraries from 16 different tissue types (Supplemental Table S3) to generate a *de novo* rattlesnake transcriptome, which we assembled using Trinity v.2014.07.17 (Grabherr et al. 2011) with default settings. *De novo* transcriptome assembly resulted in 801,342 transcripts, including 677,921 Trinity-annotated genes, with an average length of 559 bp and and N50 length of 718 bp.

We annotated repeat elements present in the improved genome assembly using libraries from complete squamate genomes (Supplementary Methods) constructed using RepeatModeler v.1.0.9 (Smit and Hubley 2008-2017). *De novo* and homology-based predictions were then performed using RepeatMasker v.4.0.6 (Smit et al. 2015-2019). We used MAKER v.2.31.8 (Cantarel et al. 2008) to annotate protein-coding genes using empirical evidence for gene prediction from our *de novo* transcriptome assembly detailed above and protein datasets of all annotated protein-coding genes for *Anolis carolinensis* (Alfoldi et al. 2011), *Python molurus bivittatus* (Castoe et al. 2013), *Thamnophis sirtalis* (Perry et al. 2018), *Ophiophagus Hannah* (Vonk et al. 2013), and

Deinagkistrodon acutus (Yin et al. 2016). Prior to running MAKER, we used BUSCO v. 2.0.1 (Simao et al. 2015) and the full *C. viridis* genome assembly to iteratively train AUGUSTUS v. 3.2.3 (Stanke and Morgenstern 2005) HMM models based on 3,950 tetrapod vertebrate benchmarking universal single-copy orthologs (BUSCOs; Supplemental Table S4). The resulting annotation consisted of 17,486 genes and we ascribed gene IDs based on homology using reciprocal best-BLAST (with e-value thresholds of 1×10^{-5}) and stringent one-way BLAST (with an e-value threshold of 1×10^{-8}) searches against protein sequences from NCBI for *Anolis*, *Python*, and *Thamnophis*.

Hi-C sequencing analysis

We dissected the venom glands from the genome animal 1 day and 3 days after venom was initially extracted in order to track a time-series of venom production. A subsample of the 1-day venom gland was sent to Dovetail Genomics where DNA was extracted and replicate Hi-C sequencing libraries were prepared according to their protocol (see above). We also extracted total RNA from both 1-day and 3-day venom gland samples, along with tongue and pancreas tissue from the Hi-C genome animal. mRNA-seq libraries were generated and sequenced at Novogene on two separate lanes of the Illumina HiSeq 4000 platform using 150 bp paired-end reads (Supplemental Table S3).

Raw Illumina paired-end reads were mapped and processed using the Juicer pipeline (Durand et al. 2016) to produce Hi-C maps binned at multiple resolutions, as low as 5 kb resolution, and for the annotation of contact domains. All contact matrices used for further analysis were KR-normalized in Juicer. Topologically-associated chromatin domains (TADs) were called using Juicer's Arrowhead algorithm for finding contact domains at various resolutions (5 kb, 10 kb, 25

kb, 50 kb and 100 kb) with default settings (Durand et al. 2016). 175 TADs were identified at 5 kb resolution, 16 at 10 kb, 53 at 25 kb, 175 at 50 kb, and 126 at 100 kb. Additionally, TADs were annotated at 20 kb resolution using the HiCExplorer software (Ramirez et al. 2018). Raw reads were mapped and processed separately through HiCExplorer and 1,262 TADs were called at 20 kb resolution using the default settings with the p-value set to 0.05. We further identified TADs by eye at finer scale (i.e., 5 kb) resolution.

We compared intra- and interchromosomal contact frequencies in the rattlesnake venom gland to the following mammalian Hi-C datasets: human lymphoblastoma cells (Rao et al. 2015) and human retinal epithelial cells, mouse kidney, and rhesus macaque tissue (Darrow et al. 2016).

Chromosome identification and synteny analysis

We determined the identity of chromosomes using a BLAST search of the chromosome-specific markers linked to snake chromosomes from Matsubara et al. (2006), downloaded from NCBI (accessions SAMN00177542 and SAMN00152474). We kept the best alignment per cDNA marker as its genomic location in the *C. viridis* genome, except when a marker hit two high-similarity matches on different chromosomes. The vast majority of markers linked to a specific macrochromosome (i.e., Chromosomes 1-7; Supplemental Tables S6, S7) in *Elaphe quadrivirgata* mapped to a single genomic scaffold.

We identified a single 114Mb scaffold corresponding to the Z Chromosome, as 10 out of 11 Z-linked markers mapped to this scaffold. To further vet this as the Z-linked region of the genome, we mapped reads from male and female *C. viridis* (Supplemental Table S8) to the genome using BWA (Li and Durbin 2009) with default settings, quantified coverage in 100 kb windows, and

normalized windowed coverage by the median autosomal value per sex. The female exhibited roughly half the coverage of the male for much of the candidate Z Chromosome, and nowhere else in the genome (Supplemental Fig. S7).

To explore broad-scale structural evolution across reptiles, we used the rattlesnake genome to perform *in silico* painting of the chicken (*Gallus gallus* version 5) and green anole (*Anolis carolinensis* version 2) genomes. Briefly, we divided the rattlesnake genome into 2.02 million potential 100 bp markers. For each of these markers, we used BLAST to record the single best hit in the target genome requiring an alignment length of at least 50 bp. This resulted in 41,644 potential markers in *Gallus* and 103,801 potential markers in *Anolis*. We then processed markers on each chromosome by requiring at least five consecutive markers supporting homology to the same rattlesnake chromosome. We consolidated each group of five consecutive potential markers as one confirmed marker. We also performed a traditional gene-based synteny analysis for comparison (Supplemental Methods, Supplemental Fig. S5), which yielded results consistent with our *k*-mer based approach.

Sex chromosome analyses

The Z Chromosome was identified using the methods above and the pseudoautosomal region (PAR) was identified based on an equal ratio of female:male genomic read coverage. The 'Recent Stratum' was identified using a comparison of female and male nucleotide diversity (π). To quantify gene expression on the rattlesnake Z Chromosome and across the genome, we prepared RNA-seq libraries from liver and kidney tissue from two males and females and sequenced them on an Illumina HiSeq using 100bp paired-end reads (Supplemental Table S9). Per gene female-to-male (F/M) ratios of expression on the Z Chromosome were normalized by

taking the $\log_2$ of the ratio of female and male Z expression values, each scaled first to the median expression level of autosomal genes in female and male, respectively. To explore regional variation in the current F/M gene expression ratio across the Z Chromosome, we performed a sliding window analysis of the $\log_2$ F/M expression ratio with a window size of 30 genes and a step size of 1 gene. Comparisons of current gene expression to inferred ancestral autosomal expression were performed using kidney and liver RNA-seq data from anole lizard and chicken males and females, following previously described methods (Julien et al. 2012; Marin et al. 2017). Additional details of these analyses are provided in the Supplementary Methods.

We predicted estrogen response elements (EREs; i.e., *ESR1* binding sites) using the conserved *ESR1* position weight matrix and binding site prediction using PoSSuM Search (Beckstette et al. 2006). We quantified the number of predicted EREs and the average current F/M gene expression ratio (see above) along the Z Chromosome in 100 kb windows, and tested for a relationship between these variables using a Pearson's correlation coefficient. We also quantified the number of predicted EREs in the entire genome, as well as the entire *Anolis* genome. We then compared the density of EREs (i.e., number of EREs divided by total scaffold length) between the rattlesnake and *Anolis* genomes, and between the rattlesnake Z Chromosome and *Anolis* Chromosome 6, specifically. We tested for ERE enrichment on the Z Chromosome compared to *Anolis* Chromosome 6 using a Fisher's exact test.

Venom analyses

Venom homologs in the rattlesnake genome were identified and annotated using representatives from 38 known venom gene families (Supplementary Methods; Supplemental Table S10). Three

venom gene families that are especially abundant, both in terms of presence in the venom proteome (Fig. 4a) and in copy number, in the venom of *C. viridis* are phospholipases A2 (*PLA2s*), snake venom metalloproteinases (*SVMPs*), and snake venom serine proteases (*SVSPs*). Rattlesnakes possess multiple members of each of these gene families, and the steps taken above appeared to underestimate the total number of copies in the *C. viridis* genome. Therefore, for each of these families, we performed an empirical annotation using the Fgenesh++ (Solovyev et al. 2006) protein similarity search.

To detect potential tandem duplication events in venom gene families, we used LASTZ (Harris 2007) to align the genomic regions containing *PLA2*, *SVMP*, and *SVSP* genes to themselves. We used program defaults, with the exception of the 'hspthresh' command, which we set to 8,000. This was done to only return very high similarity matches between compared sequences. We then performed Bayesian phylogenetic analyses to further evaluate evidence of tandem duplication and monophyly among members of the *PLA2*, *SVMP*, and *SVSP* venom gene families. We generated protein alignments of venom genes with their most similar homologs, which we identified using tBLASTx searches between venom genes and our whole gene set using MUSCLE (Edgar 2004) with default parameters, with minor manual edits to the alignment to remove any poorly aligned regions, and analyzed protein alignments using BEAST2 (Bouckaert et al. 2014).

Gene expression of annotated genes was compared between the venom gland and multiple body tissues. To test for significant expression differences between venom gland and body tissues, we performed pairwise comparisons between combined venom gland (i.e., 1 day venom gland, 3 day venom gland, and unextracted venom gland) and body (all other tissues, except for accessory

venom gland) tissue sets using an exact test of the binomial distribution estimated in edgeR, integrating tagwise dispersion (Robinson and Oshlack 2010). Genes with differential expression at an FDR value ≤ 0.05 were considered significant.

To identify candidate transcription factors regulating venom gene expression, we searched the genome annotation for all genes included on the UniProt (http://www.uniprot.org) reviewed human transcription factor database. 12 candidate transcription factors included in this list were found to be significantly upregulated in the venom gland (Supplemental Tables S12, S13). Because four transcription factors of the *NFI* family each showed evidence of venom gland-specificity, we tested that their binding motifs are also upstream of venom genes by quantifying the number of predicted *NFI* binding sites using predictive searches analogous to those used for *ESR1* (detailed above) in the 1 kb upstream region of each venom gene. We also searched for proximity of *GRHL1* binding sites to venom gene regions, as well as all non-venom genes. Here, we did not confine our search only to promoter regions. To test for enrichment of *NFI* binding sites in the upstream regions of venom genes, we divided the number of predicted binding sites upstream of venom genes by the total length of upstream regions and compared this value to the analogous proportion for upstream regions of all non-venom genes using a Fisher's exact test (Supplemental Table S13). We performed a similar analysis for *GRHL1* at each interval size, again comparing the density of predicted *GRHL1* binding sites within intervals of venom genes to non-venom genes (Supplemental Table S13).

Venom gene contact domains were identified using contact frequency heatmaps from venom gland Hi-C, and *CTCF* binding sites were again predicted using the PoSSuM Search approach detailed above using the conserved vertebrate *CTCF* position weight matrix. Because each

PSSM has a different probability distribution based on the relative frequencies of observed binding and the length of the element, we pre-calculated the complete probability distribution for each PSSM using PoSSuMdist. We then used the resulting distribution in conjunction with relative base frequencies for the genome calculated using PoSSuMfreqs to identify putative binding sites exceeding a significance threshold. This threshold necessarily varied for different PSSMs, but was never higher than $p < 1 \times 10^{-5}$.

Data access

The genome assembly has been deposited at DDBJ/ENA/Genbank under accession number PDHV02000000. The Chicago and Hi-C data generated in this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA413201. [This database also contains the previously published genome assembly (CroVir2.0) published in Pasquesi et al. (2018).] The genome resequencing data generated in this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA476794. The RNA-seq data generated in this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA477004.

Acknowledgments

Support for this work was provided by a National Science Foundation (NSF) grant DEB-1655571 to TAC and SPM, an NSF grant IOS-655735 to TAC, a Research Dissemination and Faculty Development grant from the University of Northern Colorado to SPM, and NSF DDIG grants to DRS and TAC (DEB-1501886) and to DCC and TAC (DEB-1501747).

Figures

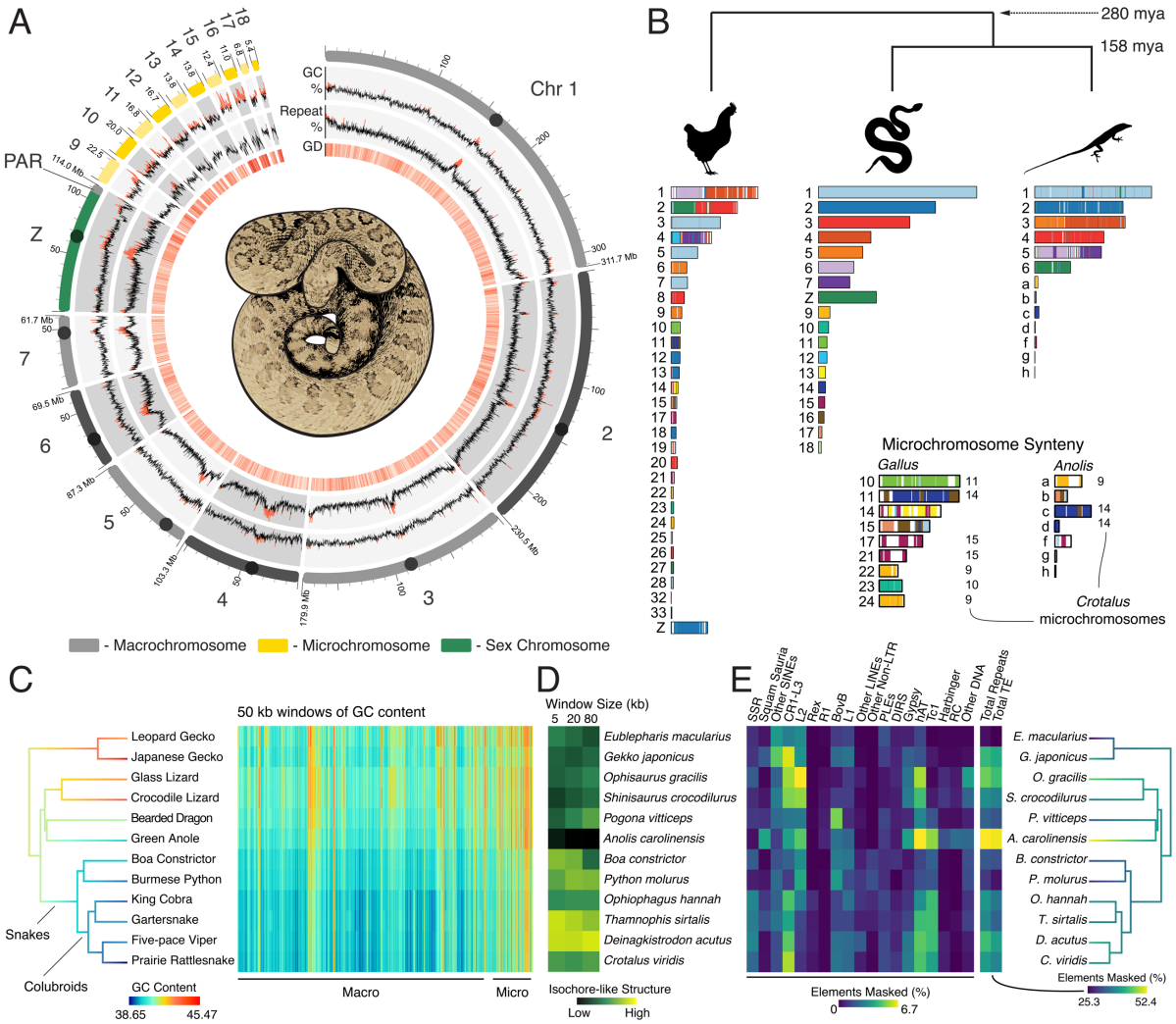


Figure 1. Structure and content of the rattlesnake genome. (A) Regional variation in GC content, genomic repeat content, and gene density (for 100 kb windows) are shown on to-scale chromosomes, with centromere locations represented by circles; values above the genome-wide median are red. GD is gene density, or the number of genes per 100 kb window; higher density shown in darker red. (B) Synteny between rattlesnake, chicken, and anole genomes. Colors on chicken and anole chromosomes correspond with homologous rattlesnake sequence. Numbers to the right of chromosomes in the microchromosome inset represent rattlesnake microchromosomes syntenic with a given chicken or anole chromosomes for greater than 80% of their length. Divergence times are shown in millions of years (mya). (C) Patterns of GC content from genome alignment of 12 squamate species, with tree branches colored according to genomic GC content. The heatmap to the right depicts GC content in 50 kb windows of aligned sequence, with macro- and microchromosome regions labeled below. (D) Genomic GC isochore structure measured by the standard deviation in GC content among 5, 20, and 80 kb windows. (*E*) Genomic repeat content among 12 squamate species, with tree branches colored by total genomic repeat content.

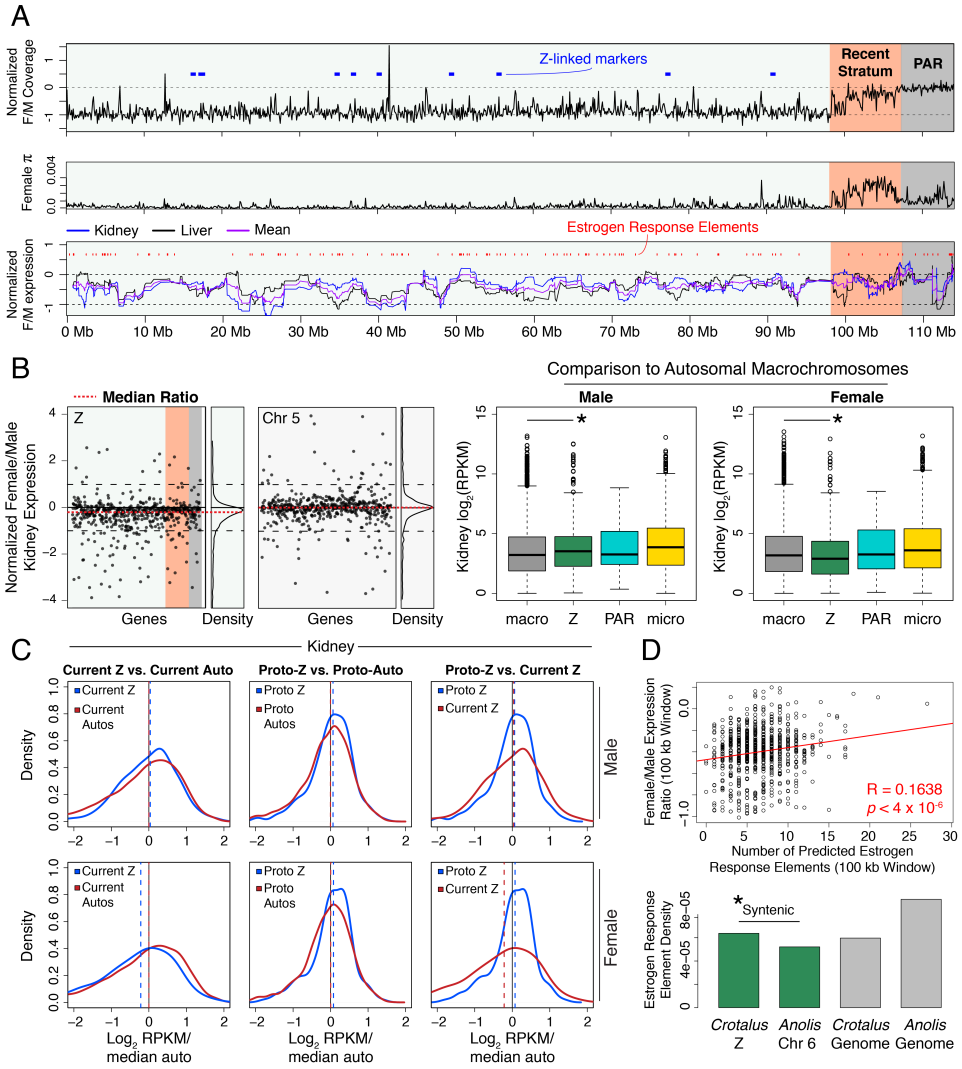


Figure 2. The Z Chromosome of the rattlesnake and the evolution of snake dosage compensation. (*A*) Normalized (log_2) female/male genomic read coverage, female π, and windowed (30 gene) log_2 normalized female/male gene expression. Known Z-linked markers (Matsubara et al. 2006) shown as blue blocks. In expression plot, red marks represent predicted estrogen response elements (EREs). On each plot, the pseudoautosomal region (PAR) and Recent Stratum are highlighted in grey and orange, respectively. (*B*) Normalized (log_2) female/male kidney gene expression per gene (black dots) across the Z shown next to expression on Chromosome 5, a similarly sized autosome (left panels). The red dashed lines are the median ratios, and relative density is shown to the right of each panel. Gene expression (log_2 RPKM) distributions for male and female across macrochromosomes, Z Chromosome, the PAR, and microchromosomes (center and right panels). Asterisks depict significant differences between autosomal and Z Chromosome expression. (*C*) Density plots of current and inferred ancestral patterns of gene expression (log_2 RPKM) in male and female kidney, respectively. Dashed lines represent the median of each distribution. (*D*) EREs drive partial dosage compensation. The correlation (red line) between predicted EREs and female/male gene expression ratios in 100 kb windows (top panel) is shown with evidence for accumulation of EREs on the rattlesnake Z (bottom panel). Each bar shows the density of EREs found in specific chromosomes (rattlesnake Z and *Anolis* 6 shown in green) and genome-wide (grey bars). The asterisk depicts a significantly greater density of EREs on the rattlesnake Z than *Anolis* Chromosome 6.

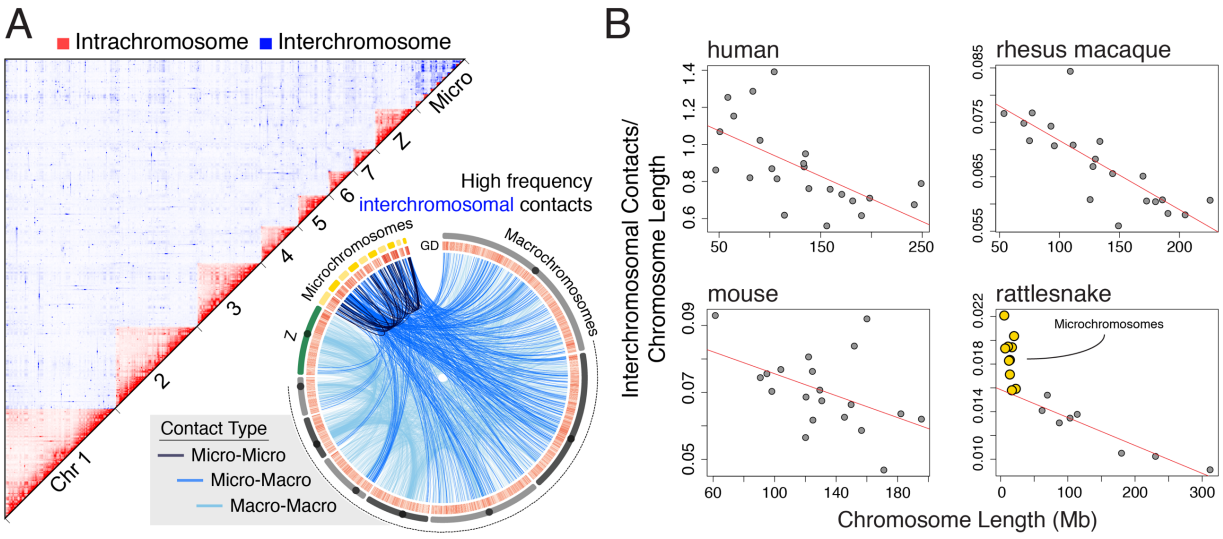


Figure 3. Genome-wide chromosomal contacts in the rattlesnake venom gland. (*A*) 2D heatmap of intrachromosomal (red) and interchromosomal (blue) contacts among rattlesnake chromosomes (top). Locations of interchromosomal contacts (bottom), where light blue lines are contacts between macrochromosomes, medium blue lines are micro-to-macrochromosome contacts, and dark blue lines are contacts between microchromosomes. (*B*) Comparison of interchromosomal contacts normalized by chromosome length versus chromosome length for different species from Hi-C datasets. Red lines depict negative linear relationships for macrochromosomes.

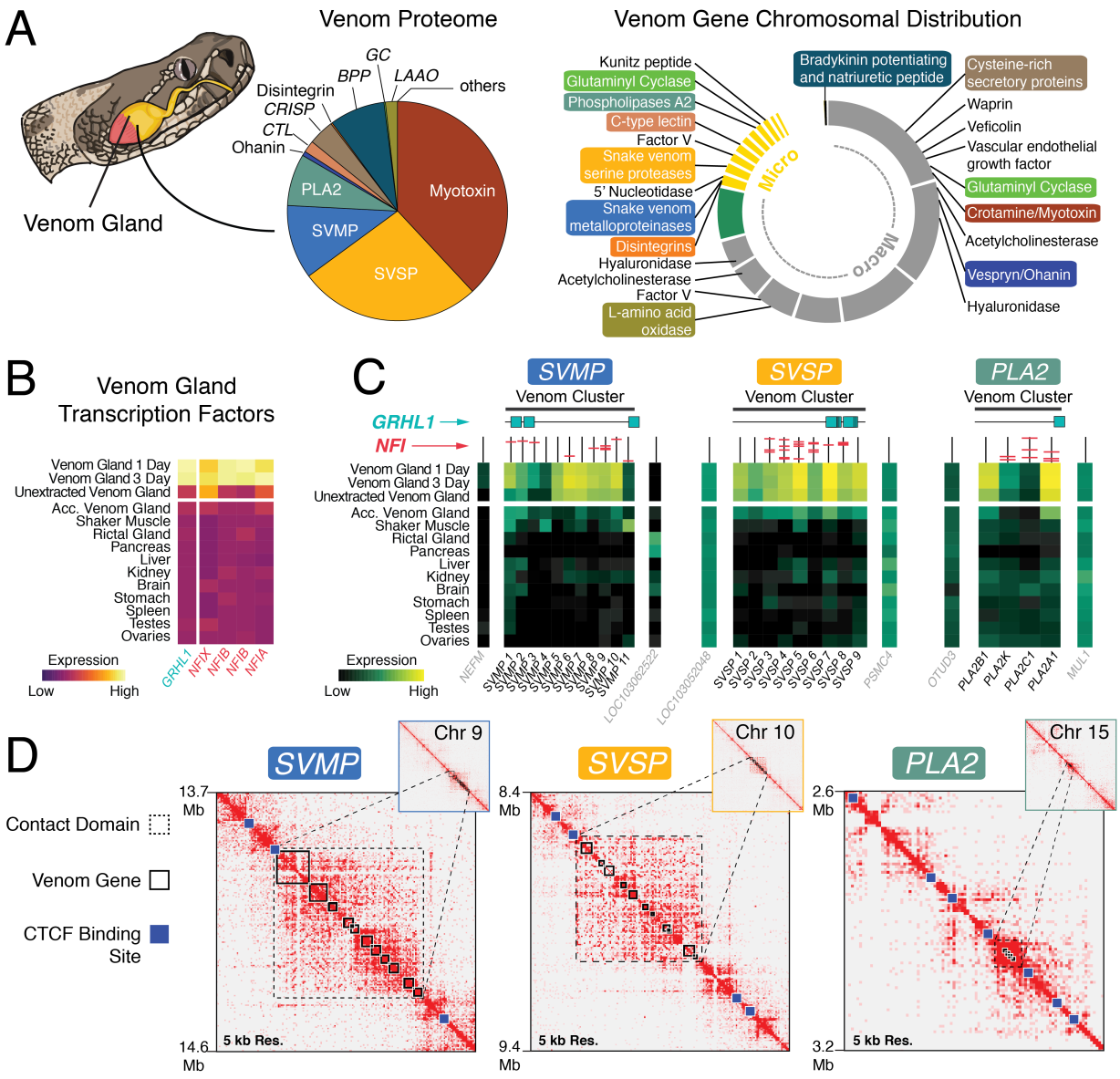


Figure 4. Genomic context for venom gene regulation and production. (*A*) Pie chart of the venom proteome with relative abundance of venom families (redrawn from Saviola et al. (2015). Chromosomal location of venom gene families (right); colored labels correspond to families from the proteome chart. (*B*) Gene expression across tissues of transcription factors (TFs) significantly upregulated in the venom gland. (*C*) Heatmaps of gene expression across tissues for venom genes in each of the three focal venom gene families, and the genes immediately flanking (i.e., outside of) each venom cluster (labeled in grey). Vertical lines above each gene represent their promoters, with predicted *NFI* binding sites shown in red. Predicted *GRHL1* binding sites in venom clusters are shown as turquoise squares. (*D*) Hi-C heatmaps showing contact domains (black dashed boxes), for the *SVMP*, *SVSP*, and *PLA2* venom genes (solid black boxes). Blue squares are predicted CTCF binding sites. Values to the left heatmaps are start and end coordinates (in Mb) of each region, visualized at 5 kb resolution.

Supplementary Methods

Prairie rattlesnake Genome Sequencing and Assembly

A male Prairie Rattlesnake (*Crotalus viridis viridis*) collected from a wild population in Colorado was used to generate the genome sequence. This specimen was collected and humanely euthanized according to University of Northern Colorado Institutional Animal Care and Use Committee protocols 0901C-SM-MLChick-12 and 1302D-SM-S-16. Colorado Parks and Wildlife scientific collecting license 12HP974 issued to S.P. Mackessy authorized collection of the animal. Genomic DNA was extracted using a standard Phenol-Chloroform-Isoamyl alcohol extraction from liver tissue that was snap frozen in liquid nitrogen. Multiple short-read sequencing libraries were prepared and sequenced on various platforms, including 50bp single-end and 150bp paired-end reads on an Illumina GAII, 100bp paired-end reads on an Illumina HiSeq, and 300bp paired-end reads on an Illumina MiSeq. Long insert libraries were also constructed by and sequenced on the PacBio platform. Finally, we constructed two sets of mate-pair libraries using an Illumina Nextera Mate Pair kit, with insert sizes of 3-5 kb and 6-8 kb, respectively. These were sequenced on two Illumina HiSeq lanes with 150bp paired-end sequencing reads. Short and long read data were used to assemble the previous genome assembly version CroVir2.0 (NCBI accession SAMN07738522). Details of these sequencing libraries are in Supplemental Table S1. Prior to assembly, reads were adapter trimmed using BBmap (Bushnell 2014) and we quality trimmed all reads using Trimmomatic v0.32 (Bolger et al. 2014). We used Meraculous (Chapman et al. 2011) and all short-read Illumina data to generate a contig assembly of the Prairie Rattlesnake. We then performed a series of scaffolding and gap-filling steps. First, we used L_RNA_scaffolder (Xue et al. 2013) to scaffold contigs using the complete

transcriptome assembly (see below), SSPACE Standard (Boetzer et al. 2011) to scaffold contigs using mate-pair reads, and SSPACE Longread to scaffold using long PacBio reads. We then used GapFiller (Nadalin et al. 2012) to extend contigs and fill gaps using all short-read data cross five iterations. We merged the scaffolded assembly with a contig assembly generated using the *de novo* assembly tool in CLC Genomics Workbench (Qiagen Bioinformatics, Redwood City, CA, USA).

We improved the CroVir2.0 assembly using the Dovetail Genomics HiRise assembly v2.1.3-59a1db48d61f method (Putnam et al. 2016), leveraging both Chicago and Hi-C sequencing. This assembly method has been used to improve numerous draft genome assemblies (e.g.,Jiao et al. 2017; Rice et al. 2017). Chicago assembly requires large amounts of high molecular weight DNA from a very fresh tissue sample. We thus extracted high molecular weight genomic DNA from a liver of a closely related male to the CroVir2.0 animal (i.e., from the same den site). This animal was collected and humanely euthanized according to the Colorado Parks and Wildlife collecting license and UNC IACUC protocols detailed above. Hi-C sequencing data were derived from the venom gland of the same animal (see details below on venom gland Hi-C and RNA-seq experimental design). The assembly was carried out using the existing CroVir2.0 draft genome assembly, short read data used in the previous assembly, Chicago, and Hi-C datasets. The HiRise assembly method then mapped Chicago and Hi-C datasets to the draft assembly and generated a model fit of the data based on insert size distributions (Supplemental Fig. S1; Supplemental Material 2). Models were generated with read pairs that mapped within the same scaffold and were used in successive join, break, and final join phases of the pipeline to perform

final scaffolding. Dovetail Genomics HiRise assembly resulted in a highly contiguous genome assembly (CroVir3.0) with a physical coverage of greater than 1,000× (Supplemental Table S2).

We estimated the size of the genome using k-mer frequency distributions (19, 23, and 27mers) quantified using Jellyfish (Marcais and Kingsford 2011). Raw Illumina 100bp paired-end reads (Supplemental Table S1) were quality trimmed using Trimmomatic (Bolger et al. 2014) using the settings LEADING:10, TRAILING:10, SLIDINGWINDOW:4:15, and MINLEN:36. The total number of output sequences and bases were 400,983,222 and 38,471,185,282, respectively. Quality trimmed reads were then used for Jellyfish *k*-mer counting, and the Jellyfish *k*-mer table output per *k*-mer was used to estimate genome size with GCE (Liu et al. 2013).

We generated transcriptomic libraries from RNA sequenced from 16 different tissues: two venom gland tissues; 1 day and 3 days post-venom extraction (see Hi-C and RNA sequencing of Venom Gland section below), one from pancreas, and one from tongue were taken from the Hi-C sequenced genome animal. Additional samples from other individuals included a third venom gland sample from which venom had not been extracted (‘unextracted venom gland’), three liver, three kidney, two pancreas, and one each of skin, lung, testis, accessory venom gland, shaker muscle, brain, stomach, ovaries, rictal gland, spleen, and blood tissues. Total RNA was extracted using Trizol, and we prepared RNAseq libraries using an NEB RNA-seq kit for each tissue, which were uniquely indexed and run on multiple HiSeq 2500 lanes using 100bp paired-end reads (Supplemental Table S3). We used Trinity v. 2014.07.17 (Grabherr et al. 2011) with default settings and the ‘--trimmomatic' setting to assemble transcriptome reads from all tissues. The resulting assembly contained 801,342 transcripts comprising 677,921 Trinity-annotated genes, with an average length of 559 bp and an N50 length of 718 bp.

Repeat Element Analysis

Annotation of repeat elements was performed using homology-based and *de novo* prediction approaches. Homology-based methods of transposable element identification (e.g., *RepeatMasker*) cannot recognize elements that are not in a reference database, and have low power to identify fragments of repeat elements belonging to even moderately diverged repeat families (Platt et al. 2016). Since the current release of the Tetrapoda RepBase library (v.20.11, August 2015; Bao et al. 2015) is unsuitable for detailed repeat element analyses of most squamate reptile genomes, we performed *de novo* identification of repeat elements on 6 snake genomes (*Crotalus viridis*, *Crotalus mitchellii*, *Thamnophis sirtalis*, *Boa constrictor*, *Deinagkistrodon acutus*, and *Pantherophis guttatus*) in RepeatModeler v.1.0.9 (Smit and Hubley 2008-2017) using default parameters. Consensus repeat sequences from multiple species were combined into a large joint snake repeat library that also includes previously identified elements from an additional 12 snake species (Castoe et al. 2013). All genomes were annotated with the same library with the exception of the green anole lizard, for which we used a lizard specific library that includes *de novo* repeat identification for *Pogona vitticeps*, *Ophisaurus gracilis*, and *Gekko japonicus*. To verify that only repeat elements were included in the custom reference library, all sequences were used as input in a BLASTx search against the SwissProt database (The UniProt Consortium 2018), and those clearly annotated as protein domains were removed. Finally, redundancy and possible chimeric artifacts were removed through clustering methods in CD-HIT (Li and Godzik 2006) using a threshold of 0.85.

Homology-based repeat element annotation was performed in RepeatMasker v.4.0.6 (Smit et al. 2015-2019) using a PCR-validated BovB/CR1 LINE retrotransposon consensus library (Castoe

et al. 2013), the Tetrapoda RepBase library, and our custom library as references. Output files were post-processed using a modified implementation of the ProcessRepeat script (RepeatMasker package).

Gene Annotation

We used MAKER v. 2.31.8 (Cantarel et al. 2008) to annotate protein-coding genes in an iterative fashion. Several sources of empirical evidence of protein-coding genes were used, including the full *de novo C. viridis* transcriptome assembly and protein datasets consisting of all annotated proteins from NCBI for *Anolis carolinensis* (Alfoldi et al. 2011), *Python molurus bivittatus* (Castoe et al. 2013), *Thamnophis sirtalis* (Perry et al. 2018), and *Ophiophagus hannah* (Vonk et al. 2013), and from GigaDB for *Deinagkistrodon acutus* (Yin et al. 2016). We also included 422 protein sequences for 24 known venom gene families that were used to infer *Python* venom gene homologs in a previous study (Reyes-Velasco et al. 2015). Prior to running MAKER, we used BUSCO v. 2.0.1 (Simao et al. 2015) and the full *C. viridis* genome assembly to iterative train AUGUSTUS v. 3.2.3 (Stanke and Morgenstern 2005) HMM models based on 3,950 tetrapod vertebrate benchmarking universal single-copy orthologs (BUSCOs). We also ran this analysis on the previous genome assembly (CroVir2.0) as a comparison, and provide the details of these analyses in Supplemental Table S4. We ran BUSCO in the 'genome' mode and specified the '--long' option to have BUSCO perform internal AUGUSTUS training. We ran MAKER with the 'est2genome=0' and 'protein2genome=0' options set to produce gene models using the AUGUSTUS gene predictions with hints supplied from the empirical transcript and protein sequence evidence. We provided the coordinates for all interspersed, complex repetitive elements for MAKER to perform hard masking before evidence mapping and prediction, and we

set the ‘model_org’ option to ‘simple’ to have MAKER soft mask simple repetitive elements. We used default settings for all other options, except ‘max_dna_len’ (set to 300,000) and ‘split_hit’ (set to 20,000). We iterated this approach an additional time and we manually compared the MAKER gene models with the transcript and protein evidence. We found very little difference between the two gene annotations and based on a slightly better annotation edit distance (AED) distribution in the first round of MAKER, we used our initial round as the final gene annotation. The resulting annotation consisted of 17,486 genes and we ascribed gene IDs based on homology using reciprocal best-BLAST (with e-value thresholds of 1×10^{-5}) and stringent one-way BLAST (with an e-value threshold of 1×10^{-8}) searches against protein sequences from NCBI for *Anolis*, *Python*, and *Thamnophis*.

Hi-C and RNA Sequencing of the Venom Gland

We dissected the venom glands from the Hi-C *Crotalus viridis viridis* 1 day and 3 days after venom was initially extracted in order to track a time-series of venom production. A subsample of the 1-day venom gland was sent to Dovetail Genomics where DNA was extracted and replicate Hi-C sequencing libraries were prepared according to their protocol (see above). We also extracted total RNA from both 1-day and 3-day venom gland samples, along with tongue and pancreas tissue from the Hi-C genome animal (see Sequencing and Assembly and Annotation sections above). mRNA-seq libraries were generated and sequenced at Novogene on two separate lanes of the Illumina HiSeq 4000 platform using 150 bp paired-end reads (Supplemental Table S3).

Chromosome Identification and Synteny Analyses

Genome assembly resulted in several large, highly-contiguous scaffolds with a relative size distribution consistent with the karyotype of *C. viridis* (Baker et al. 1972), representing nearly-complete chromosome sequences. We determined the identity of chromosomes using a BLAST search of the chromosome-specific markers linked to snake chromosomes from Matsubara et al. (2006), downloaded from NCBI (accessions SAMN00177542 and SAMN00152474). We kept the best alignment per cDNA marker as its genomic location in the Prairie Rattlesnake genome, except when a marker hit two high-similarity matches on different chromosomes. The vast majority of markers linked to a specific macrochromosome (i.e., Chromosomes 1-7; Supplemental Table S6) in *Elaphe quadrivirgata* mapped to a single genomic scaffold; only six of 104 markers did not map to the predicted chromosome from *E. quadrivirgata*. Possible reasons for unmatched chromosomal locations for these markers in *Elaphe* and the Prairie Rattlesnake include 1) original localizations in *Elaphe* that are unique to the species or were localized in error, 2) translocations have occurred, leading to divergent locations in each genome, or 3) misassembly errors in the rattlesnake genome assembly. To distinguish between these possibilities, we first identified the chromosomal location of each marker in the Anole Lizard (*Anolis carolinensis*) genome (Alfoldi et al. 2011) to determine if their locations are expected based on *Elaphe-Anolis* synteny. Three markers mapped to unexpected chromosomes in *Anolis* (*NOSIP*, *GNAI2*, and *P4HB*), which instead matched syntenic locations in the rattlesnake (Supplemental Table S7). *Anolis* synteny for a fourth marker (*UCHL1*) suggested correct assembly in the rattlesnake, but was unclear because it mapped to *Anolis* Chromosome 5, which is syntenic with both snake Chromosomes 6 and 7 (Fig. 1). To determine if the two remaining markers (*ZNF326* and *KLF6*) were placed on unexpected chromosomes due to misassembly, and

to identify further evidence that the other markers were assembled correctly, we leveraged our intrachromosomal Hi-C data to deeply investigate contact patterns around these markers. Specifically, we plotted heatmaps of $\log_{10}$ normalized contact frequencies in 10 kb bins using R (R Core Team). Regional dropout in intrachromosomal contact frequencies in the focal regions would be expected if mismatched chromosome locations were due to misassembly in the rattlesnake. We focused our searches on genomic intervals around each of the six focal genes and the nearest confirmed marker from Supplemental Table S6. The genomic region around each gene showed intrachromosomal contact frequencies consistent with correct assembly for five of six markers (Supplemental Fig. S2). Only *ZNF326* was adjacent to a region with intrachromosomal contact dropout that could have resulted from misassembly. All snake microchromosome markers mapped to a single 139Mb scaffold, which was later broken into 10 microchromosome scaffolds (scaffold-mi1-10; see below).

We identified a single 114Mb scaffold corresponding to the Z Chromosome, as 10 of 11 Z-linked markers mapped to this scaffold. To further vet this as the Z-linked region of the genome, we mapped reads from male and female *C. viridis* (Supplemental Table S9) to the genome using BWA (Li and Durbin 2009) using program defaults. Male and female resequencing libraries were prepared using an Illumina Nextera prep kit and sequenced on an Illumina HiSeq 2500 using 250bp paired-end reads. Adapters were trimmed and low-quality reads were filtered using Trimmomatic (Bolger et al. 2014). After mapping, we filtered reads with low mapping scores and quantified per-base read depths using SAMtools (Li et al. 2009). We then totaled read depths for consecutive 100 kb windows and normalized windowed totals for female and male by dividing the value for each window by the median autosomal 100 kb window value for each sex,

then determined the normalized ratio of female to male coverage by calculating $\log_2$(female normalized coverage/male normalized coverage) per window. Here, the expectation is that a hemizygous locus will show roughly half the normalized coverage, which we observe for females over the majority of the Z Chromosome scaffold length, and not elsewhere in the genome. To demonstrate Z Chromosome conservation among pit vipers and to further determine the identity of this scaffold, we mapped male and female Pygmy Rattlesnake (*Sistrurus catenatus*) reads from Vicoso et al. (2013) and female and male Five Pace Viper (*Deinagkistrodon acutus*) reads from Yin et al. (2016) to the genome using the same parameters detailed above (Supplemental Fig. S7). *Anolis* Chromosome 6 is homologous with snake sex chromsomes (Srikulnath et al. 2009), thus we aligned *Anolis* Chromosome 6 (Alfoldi et al. 2011) to the Prairie Rattlesnake genome using a chromosome painting technique described below. As expected, we found a large quantity of high-similarity hits to the rattlesnake Z Chromosome scaffold, specifically, which were organized in a sequential manner across the Z scaffold (Fig. 1B).

We used multiple sources of information to identify the best candidate breakpoints between microchromosomes within the 139Mb fused microchromosome scaffold in the initial Hi-C assembly. First, because Chicago scaffolds must be assembled from fragments that are physically linked (Rice et al. 2017), we used breakpoints between adjacent Chicago scaffolds on the microchromosome scaffold as candidate misjoins between microchromosomes, which identified 305 candidate break points. Second, intrachromosomal contact frequencies have been shown to be exponentially higher than contacts between chromosomes (Lieberman-Aiden et al. 2009), and we used shifts in intrachromosomal Hi-C data to further identify the nine most biologically

plausible candidate break points among microchromosomes (Supplemental Fig. S16). Here, we stress two things relevant to using Hi-C contact data for this purpose: 1) intrachromosomal contacts within candidate microchromosomes were far more frequent than contacts between candidate microchromosomes, as expected (Supplemental Fig. S16), and 2) the nine Hi-C derived breakpoints overlapped consistently with breaks between Chicago scaffolds. Because reptile microchromosomes are highly syntenic (Alfoldi et al. 2011), we also aligned the microchromosome scaffold to microchromosome scaffolds from chicken (Hillier et al. 2004) and *Anolis* using LASTZ (Harris 2007) to determine if likely chromosomal breakpoints also had shifts in synteny. To retain only highly similar alignments per comparison, we set the 'hspthresh' option equal to 10,000 (default is 3,000). We also set a step size equal to 20 to reduce computational time per comparison. We further validated candidate break points using genomic features that consistently vary at the ends of chromosomes. Here, we specifically evaluated if candidate breakpoints exhibited regional shifts in GC content and repeat content, similar to the ends of macrochromosomes (Fig. 1). Finally, if no annotated genes spanned this junction, we considered it biologically plausible. There were nine candidate breakpoints that met each of these criteria, equaling the number of boundaries expected given ten microchromosomes (Supplemental Fig. S16).

To explore broad-scale structural evolution across reptiles, we used the rattlesnake genome to perform in silico painting of the chicken (*Gallus gallus* version 5) and green anole *Anolis carolinensis* (version 2) genomes. Briefly, we divided the rattlesnake genome into 2.02 million potential 100 bp markers. For each of these markers, we used BLAST to record the single best hit in the target genome requiring an alignment length of at least 50 bp. This resulted in 41,644

potential markers in *Gallus* and 103,801 potential markers in *Anolis*. We then processed markers on each chromosome by requiring at least five consecutive markers supporting homology to the same rattlesnake chromosome. We consolidated each group of five consecutive potential markers as one confirmed marker. In *Gallus*, we rejected 12.4% of potential markers and identified 7,291 confirmed merged markers. In *Anolis*, we rejected 39.7% of potential markers and identified 12,511 confirmed merged markers.

This approach demonstrates considerable stability at the chromosomal level despite 158 million years of divergence between *Anolis* and *Crotalus* (Fig. 1B, Supplemental Fig. S5), and between squamates and birds, despite 280 million years of divergence between *Gallus* and *Crotalus* (and between *Gallus* and *Anolis*). This stability is evident not only in the macrochromosomes but also in the microchromosomes. In fact, 7 of 10 *Crotalus* microchromosomes had greater than 80% of confirmed markers associated with a single chromosome in the chicken genome (Fig. 1B, microchromosome inset). Comparisons among the three genomes suggest that the *Crotalus* genome has not experienced some of the fusions found in *Anolis*. Specifically, we infer that *Anolis* Chromosome 3 is a fusion of *Crotalus* Chromosomes 4 and 5. Likewise, *Anolis* Chromosome 4 is a fusion of *Crotalus* Chromosome 6 and 7. Divergence time estimates discussed above and shown in Fig. 1B were taken from the median of estimates for divergence between *Crotalus* and *Gallus* and between *Crotalus* and *Anolis* from Timetree (www.timetree.org; Kumar et al. 2017).

To validate the genome-wide *k*-mer based approach used to identify homology among reptile chromosomes, we also performed a more traditional analysis using only protein-coding genes. We first identified 2,190 three-way reciprocal best BLAST hits among rattlesnake, anolis, and

chicken protein-coding genes that we used as markers. Both the chicken and *Anolis* genomes contain genes that have not been placed on chromosomes and remain in unmapped scaffold or contigs, which reduced the number of markers available to 2,105 in chicken and 2,135 in *Anolis*. Results from this approach indicate that the *k*-mer approach is consistent with this more traditional approach but provides approximately three times the density of markers (Fig. 1B, Supplemental Fig. S5).

Genomic Patterns of GC Content

We quantified GC content in sliding windows of 100 kb and 1Mb across the genome using a custom Python script (https://github.com/drewschield/Comparative-Genomics-Tools/blob/master/slidingwindow_gc_content.py). GC content in 100 kb windows is presented in Fig. 1 in the Main Text.

To determine if there is regional variation in nucleotide composition consistent with isochore structures across the rattlesnake genome, we quantified GC content and its variance within 5, 10, 20, 40, 80, 160, 240, and 320 kb windows. The variation (standard deviation) in GC content is expected to decrease by half as window size increases four-fold if the genome is homogeneous (i.e., lacks isochore structures; Venter et al. 2001). By comparing the observed variances of GC content across spatial window scales to those from 11 other squamate genomes, including lizards (*Anolis* has been shown to lack isochore structure; Alfoldi et al. 2011), henophidian snakes, and colubroid snakes, we were able to determine the relative heterogeneity of nucleotide composition in the rattlesnake (Supplemental Table S8). To reduce potential biases from estimates from small scaffold sizes, we filtered to only retain scaffolds greater than the size of the window analyzed

(e.g., only scaffolds longer than 10 kb when looking at the standard deviation in GC content over 10 kb windows) and for which there was less than 20% of missing data.

To study patterns of molecular evolution across squamate evolution, we generated whole genome alignments of 12 squamates including the Green Anole (*Anolis carolinensis* v. anoCar2.0; Alfoldi et al. 2011), Australian Bearded Dragon (*Pogona vitticeps* v. pvi1.1; Georges et al. 2015), Crocodile Lizard (*Shinisaurus crocodilurus* GigaDb version; Gao et al. 2017), Glass Lizard (*Ophisaurus gracilis* v. O.gracilis.final; Song et al. 2015), Schlegel's Japanese Gecko (*Gekko japonicus* v. 1.1 ; Liu et al. 2015), Leopard Gecko (*Eublepharis macularius* v. 1.0; Xiong et al. 2016), Prairie Rattlesnake (*Crotalus viridis* v. CroVir3.0; current study), Five-pacer Viper (*Deinagkistrodon acutus* GigaDb version; Yin et al. 2016), Burmese Python (*Python bivittatus* v. Python_molurus_bivittatus-5.0.2; Castoe et al. 2013), Boa Constrictor (*Boa constrictor* v. 7C; Bradnam et al. 2013), Garter Snake (*Thamnophis sirtalis* NCBI version; Perry et al. 2018), and King Cobra (*Ophiophagus Hannah* v. OphHan1.0; Vonk et al. 2013). We obtained the repeat libraries for each species and softmasked each assembly. The repeat library was not available for *Deinagkistrodon*, so we annotated repeats in that assembly using RepeatMasker v4.0.5 (Smit et al. 2015-2019) with the vertebrate library from RepBase (Jurka et al. 2005). First, we generated pairwise syntenic alignments of each species as a query to the green anole genome (anoCar2.0) as a target using LASTZ v1.02 (Harris 2007) with the HoxD55 scoring matrix, followed by chaining to form gapless blocks and netting to rank the highest scoring chains (Kent et al. 2003) . The pairwise alignments were used to construct a multiple sequence alignment with MULTIZ v11.2 (Blanchette et al. 2004) with Green Anole as the reference species. We then filtered the multi-species whole genome alignment to retain only blocks for which information for all 12

species was available, and concatenated blocks according to their organization in the anole lizard genome. We then calculated GC content within consecutive 50 kb windows of this concatenated alignment using the 'slidingwindow_gc_content.py' script detailed above.

Comparative Microchromosome Genomics

To understand evolutionary shifts in microchromosome composition among amniotes, we compared measures of gene density, GC content, and repeat content of macro- and microchromosomes between the rattlesnake, anole (Alfoldi et al. 2011), bearded dragon (Georges et al. 2015; Deakin et al. 2016), chicken (Hillier et al. 2004), and zebra finch (Warren et al. 2010) genomes. These species were chosen because their scaffolds are ordered into chromosomes and because their karyotypes contain microchromosomes. For each species, we downloaded relevant data from Ensembl and quantified the total number of genes per chromosome, total number of G+C bases, and total bases masked as repeats in RepeatMasker. We then normalized each measure by the total length of macrochromosome and microchromosome sequences in each genome, then calculated the ratio of microchromosome:macrochromosome proportions. We then used Fisher's exact tests determine if one chromosome set possessed a significantly greater proportion of each measure. We generated a phylogenetic tree (Supplemental Fig. S4) for the five species based on divergence time estimates from TimeTree (Kumar et al. 2017), and plotted the ratio values calculated above onto the tree tips for between-species comparisons.

Hi-C analysis

Raw Illumina paired-end reads were mapped and processed using the Juicer pipeline (Durand et al. 2016) to produce Hi-C maps binned at multiple resolutions, as low as 5 kb resolution, and for

the annotation of contact domains. These data were aligned against the CroVir3.0 assembly. All contact matrices used for further analysis were KR-normalized in Juicer. TAD domains were called using Juicer's Arrowhead algorithm for finding contact domains at various resolutions (5 kb, 10 kb, 25 kb, 50 kb and 100 kb) using the default settings (Durand et al. 2016). 175 TADs were identified at 5 kb resolution, 16 at 10 kb, 53 at 25 kb, 175 at 50 kb, and 126 at 100 kb. Additionally, TADs were annotated at 20kb resolution using the HiCExplorer software (Ramirez et al. 2018). Raw reads were mapped and processed separately through HiCExplorer and 1,262 TADs were called at 20 kb resolution using the default settings with the p-value set to 0.05. We further identified TADs by eye at finer scale (i.e., 5 kb) resolution.

We compared intra and interchromosomal contact frequencies between the rattlesnake venom gland and various tissues from mammals. To do this we quantified the total intra- and interchromosomal contacts between chromosome positions from the rattlesnake and the following Hi-C datasets: human lymphoblastoma cells (Rao et al. 2015) and human retinal epithelial cells, mouse kidney, and rhesus macaque tissue (Darrow et al. 2016). To investigate patterns of intra- and interchromosome contact frequency, we normalized contact frequencies by chromosome length. In the case of the mouse, we removed the Y chromosome due to its small size and relative lack of interchromosomal contacts. We then performed linear regressions of chromosome length and normalized intra- and interchromosomal contact frequencies (i.e., contact frequency/chromosome length). In all cases we observed a positive relationship between normalized intrachromosomal contacts and chromosome size and a negative relationship between normalized interchromosomal contacts and chromosome size (Fig. 3B). We also tested

for significant differences in intra- and interchromosomal contact between the rattlesnake and mammals using *t*-tests.

Sex Chromosome Analysis

We identified the Prairie Rattlesnake Z Chromosome using methods described in the 'Chromosome Identification and Synteny Analyses' section above. We localized the candidate pseudoautosomal region (PAR) based on normalized female/male coverage (Fig. 2A; the PAR is the only region of the Z consistent with equal female and male coverage. We quantified gene content, GC content, and repeat content across the Z Chromosome and PAR (Supplemental Figs. S8, S9, and S10), and tested for gene enrichment in the PAR using a Fisher's exact test, where we compared the number of genes within each region to the total length of the region.

To compare nucleotide diversity (π) across the genome between male and female *C. viridis*, we called variants (i.e., heterozygous sites) from the male and female reads used in coverage analysis detailed above. With the mappings from coverage analysis, we used SAMtools (Li et al. 2009) to compile all mappings into pileup format, from which we called variant sites using BCFtools. We filtered sites to retain only biallelic variants using VCFtools (Danecek et al. 2011) and calculated the proportion of heterozygous sites using a custom pipeline of scripts. First, calcHet (https://github.com/darencard/RADpipe) outputs details of heterozygous site and window_heterozygosity.py (https://github.com/drewschield/Comparative-Genomics-Tools/blob/master/window_heterozygosity.py) uses this output in conjunction with a windowed bed file generated using BEDtools 'make_windows' tool to calculate π within a given window size. We then normalized π for each genomic window in the female and male by the median value of π for female and male autosomes, respectively.

Evolutionary patterns of the Z Chromosome were also analyzed by examining transposable element age and composition along the whole chromosome, and between the PAR and the Z, specifically (see Main Text). Since the length of the PAR is significantly smaller than the length of the Z, to rule out potential biases due to unequal sample size we also independently analyzed fragments of the Z with lengths equal to the PAR (total of 15 7.18 Mbp fragments). Each region was analyzed in RepeatMasker using a single reference library that included the squamate fraction of the RepBase Tetrapoda library, and the snake specific library clustered at a threshold of 0.75. The age distribution of TE families was estimated by mean of the Kimura 2-parameter distance from the consensus sequence per element (CpG corrected) calculated from PostProcessed.align outputs (see 'Repeat Analysis' section above), and using a modified Perl script from (Kapusta et al. 2017). We then merged estimates of repeat content from each of these regions for comparison to the PAR region, specifically.

To quantify gene expression on the rattlesnake Z Chromosome and across the genome, we prepared RNA-seq libraries from liver and kidney tissue from two males and females and sequenced them on an Illumina HiSeq using 100bp paired-end reads (Supplemental Table S9). Samples and libraries were prepared following the previously described methods of (Andrew et al. 2017). After filtering and adapter trimming using Trimmomatic v. 0.32 (Bolger et al. 2014), we mapped RNA-seq reads to the *C. viridis* genome using STAR v. 2.5.2b (Dobin et al. 2013) and counts were determined using featureCounts (Liao et al. 2014). To be comparable to anole and chicken RNA-seq data described below, we analyzed the rattlesnake RNA-seq reads as single-end data by ignoring the second read of each read pair. We normalized read counts across tissues and samples using TMM normalization in edgeR (Robinson et al. 2010) to generate both

counts per million (CPM) for use in pairwise comparisons between males and females, and reads per kilobase million (RPKM) normalized counts for comparisons of chromosome-wide expression within samples. All subsequent analyses of gene expression included only genes with expression information in both the male and female (>1 average RPKM in each sex; average overall for female and male were roughly equal). Mann-Whitney *U* tests in R (R Core Team) were used to compare median expression level between chromosomes and/or chromosomal regions (i.e., the PAR) within males and females. Per gene female-to-male ratios of expression in the Z Chromosome were normalized by taking the $\log_2$ of the female and male Z expression values scaled to the median expression level of autosomal genes in female and male, respectively:

$$\text{Current female/male Z} = log_2\left[\left(\frac{female\ Z}{median\ female\ Auto}\right) / \left(\frac{male\ Z}{median\ male\ Auto}\right)\right]$$

To explore regional variation in the current female-to-male (F/M) gene expression ratio across the Z Chromosome, we performed a sliding window analysis of the $\log_2$ F/M expression ratio with a window size of 30 genes and a step size of 1 gene.

To further investigate patterns of gene expression in females and males across the Z Chromosome, we compared current levels of female and male expression for Z-linked genes to inferred ancestral levels of expression using autosomal 1:1 orthologs in the anole lizard and the chicken. Comparisons of sex chromosome-linked genes to autosomal orthologs in outgroup species have been shown to provide robust information about global ancestral expression patterns in the 'proto-sex' chromosomes of the focal species (Julien et al. 2012; Marin et al. 2017), and can be used to determine if patterns of gene expression between sexes are consistent

with each other and with the evolution of dosage compensation mechanisms. We first filtered to retain only the 1,343 non-PAR genes on the rattlesnake Z Chromosome for comparison, and used reciprocal best BLAST searches to find putative 1:1 orthologs in the Ensembl anole (version 2) and chicken (version 5) cDNA datasets, respectively. This resulted in 682 1:1 orthologs between the rattlesnake and the anole, and 291 between the rattlesnake and the chicken, and 260 shared orthologs among the three species (i.e., 'proto-Z' genes). All putative orthologs are located on autosomes in both the anole and chicken. We also identified 3,059 1:1 orthologs that are autosomal in all three species (i.e., 'proto-autosomal' genes). We then obtained RNA-seq data from (Marin et al. 2017) for female and male kidney and liver tissue for the chicken and anole (at least two replicates per tissue per sex) and performed filtering, mapping, and normalization of counts using the methods described above for the rattlesnake.

We used female and male expression levels from rattlesnake Z autosomal orthologs in the anole and chicken to infer ancestral (i.e., proto-Z) female and male expression levels. To do this, we first calculated the average expression value per proto-autosomal gene between the anole and chicken for each sex, and then calculated the median expression value from each of these distributions. We used these median values to normalize female and male expression in the anole and chicken 1:1 rattlesnake Z orthologs (proto-Z genes) to a common scale (these values are analogous to the median female or male autosomal denominators in the equations above for current female/male expression).

$$\text{Proto-Z female} = \frac{female\ proto{-}Z\ gene}{median\ female\ proto{-}autosomal}$$

$$\text{Proto-Z male} = \frac{male\ proto{-}Z\ gene}{median\ male\ proto{-}autosomal}$$

We then calculated a weighted average of female and male proto-Z expression per gene between the anole and chicken designed to account for the more recent divergence between the anole and rattlesnake, which was equal to the reciprocal of the sum of branch lengths based on the divergence times in millions of years between rattlesnake and anole and between rattlesnake and chicken:

$$\text{Branch length weight} = \frac{branch\ length\ (rattlesnake\ to\ anole)=158}{branch\ length\ (rattlesnake\ to\ chicken)=402} = 0.393$$

$$\text{Weighted Proto-Z female} = log2\left[\frac{[(Proto{-}Z\ female\ anole*1)+(Proto{-}Z\ female\ chicken*0.393)]}{1.393}\right]$$

$$\text{Weighted Proto-Z male} = log2\left[\frac{[(Proto{-}Z\ male\ anole*1)+(Proto{-}Z\ male\ chicken*0.393]}{1.393}\right]$$

To further compare current and ancestral Z expression to the female and male distributions of proto-autosomal expression, we calculated the average expression between the anole and chicken per proto-autosomal gene, then normalized the averaged expression by the median of proto-autosomal expression detailed above:

$$\text{Proto-autosomal female} = log2\left[\frac{female\ proto{-}autosomal\ gene}{median\ female\ proto{-}autosomal}\right]$$

$$\text{Proto-autosomal male} = log2\left[\frac{male\ proto{-}autosomal\ gene}{median\ female\ proto{-}autosomal}\right]$$

We also calculated the distribution of current autosomal expression in the rattlesnake by normalizing the current female and male expression of rattlesnake autosomal genes by the median of female and male expression of all autosomal genes, respectively:

$$\text{Current autosomal female} = log_2 \left[\frac{female\ autosomal\ gene}{median\ female\ autosomal}\right]$$

$$\text{Current autosomal male} = log_2 \left[\frac{male\ autosomal\ gene}{median\ female\ autosomal}\right]$$

We tested for enrichment of male and female-biased gene expression on chromosomes by first characterizing genes as male or female biased if their current $\log_2$(female/male) expression ratio was less than -0.5 or greater than 0.5, respectively. We then compared proportions of male-biased, female-biased, and unbiased between the Z Chromosome and autosomes using Fisher's exact tests to determine if the Z Chromosome is enriched or depleted for sex-biased gene expression.

A potential mechanism for upregulation of Z-linked genes in females is positive regulation through estrogen response elements (EREs), which can enable binding of enhancers and promote transcription of genes over long distances (Lin et al. 2007). Rice et al. (2017) identified that the binding domain of *ESR1* is completely conserved among humans, chickens, and alligators, thus we obtained a position weight matrix for the *ESR1* binding motif (ERE) of humans (Lin et al. 2007) from the CisBP database, and performed binding site prediction using PoSSuM Search (Beckstette et al. 2006). For more details on PoSSuM Search parameters, see the 'Transcription Factor Binding Site Prediction' section below. We quantified the number of predicted EREs and the average current female/male gene expression ratio (see above) along the Z Chromosome in 100 kb windows, and tested for a relationship between these variables using a Pearson's correlation coefficient in R.

We also quantified the number of predicted EREs in the entire genome, as well as the entire *Anolis* genome. We then compared the density of EREs (i.e., number of EREs divided by total scaffold length) between the rattlesnake *Anolis* genomes, and between the rattlesnake Z Chromosome and *Anolis* Chromosome 6, specifically. We tested for ERE enrichment on the Z Chromosome compared to *Anolis 6* using a Fisher's exact test in R. To test more broadly for an expansion of EREs in snakes, we repeated this analysis using Z-linked and autosomal scaffolds from the five pace viper (*Deinagkistron acutus*; Yin et al. 2016).

Transcription Factor Binding Site Prediction

To identify putative transcription factor binding sites throughout the rattlesnake genome, we obtained the TRANSFAC position weight matrix (PSSM) for transcription factors of interest from the CIS-BP database (Weirauch et al. 2014). The focal transcription factors (e.g., *CTCF*, *NFI, GRHL1*, *ESR1*, and the remaining transcription factors on Supplemental Table S12) have conserved DNA binding domains among vertebrates, and where possible we obtained the chicken binding PSSM. In some cases there was no curated PSSM for chicken, and we used the PSSMs for human, and in the case of *NCOA2* (Supplemental Table S12), there was no available PSSM for a close relative. We searched for putative binding sites throughout the genome using PoSSuM Search (Beckstette et al. 2006). Because each PSSM has a different probability distribution based on the relative frequencies of observed binding and the length of the element, we pre-calculated the complete probability distribution for each PSSM using PoSSuMdist. We then used the resulting distribution in conjunction with relative base frequencies for the genome calculated using PoSSuMfreqs to identify putative binding sites exceeding a significance

threshold. This threshold necessarily varied for different PSSMs, but was never higher than $p < 1 \times 10^{-5}$.

Venom Gene Annotation and Analysis

We took a multi-step approach toward identifying venom gene homologs in the rattlesnake genome. We first obtained representative gene sequences for 38 venom gene families from GenBank (Supplemental Table S10), comprising known enzymatic and toxin components of snake venoms. We then searched our transcript set using the venom gene family query set using a tBLASTx search, defining a similarity cutoff e-value of 1×10^{-5}. For each candidate venom gene transcript identified in this way, we then performed a secondary tBLASTx search against the NCBI database to confirm its identity as a venom gene. In the case of several venom gene families, such as those known only from elapid snake venom, we did not find any candidate genes. Three venom gene families that are especially abundant, both in terms of presence in the venom proteome (Fig. 4a) and in copy number, in the venom of *C. viridis* are phospholipases A2 (*PLA2s*), snake venom metalloproteinases (*SVMPs*), and snake venom serine proteases (*SVSPs*). Rattlesnakes possess multiple members of each of these gene families (Mackessy 2008; Casewell et al. 2011; Dowell et al. 2016), and the steps taken above appeared to underestimate the total number of copies in the *C. viridis* genome. Therefore, for each of these families, we performed an empirical annotation using the Fgenesh++ (Solovyev et al. 2006) protein similarity search. We first extracted the genomic region annotated for each of these families above plus and minus a 100 kb flanking region. We used protein sequences from Uniprot (*PLA2*: APD70899.1; *SVMP*: Q90282.1; and *SVSP*: F8S114.1) to query the region and confirm the total number of copies per family. Each gene annotated in this way was again searched against NCBI to confirm its identity

and manual searches of aligned protein sequences (see phylogenetic analyses below) further confirmed their homology to each respective venom gene family. Genomic locations and details of annotated venom genes in the rattlesnake genome are provided in Table S9. We tested for venom gene enrichment on microchromosomes versus macrochromosomes using a Fisher's exact test, where numerator for each category was the number of venom genes located on each chromosome type, and the denominator in each category was the background number of genes, which allowed us to account for different levels of gene density on microchromosomes and macrochromosomes.

We used LASTZ (Harris 2007) to align the genomic regions containing *PLA2*, *SVMP*, and *SVSP* genes to themselves. We used program defaults, with the exception of the 'hspthresh' command, which we set to 8,000. This was done to only return very high similarity matches between compared sequences. Here the expectation is that when alignments are plotted against one another, we will observe a diagonal line demonstrating perfect matches between each stretch of sequence and itself. In the case of segmental duplications, we also expect to see parallel and perpendicular (if in reverse orientation) segments adjacent to the diagonal 'self' axis. We plotted LASTZ results for each of the regions using the base plotting function in R (R Core Team).

We then performed Bayesian phylogenetic analyses to further evaluate evidence of tandem duplication and monophyly among members of the *PLA2*, *SVMP*, and *SVSP* venom gene families. We generated protein alignments of venom genes with their closest homologs, which we identified using tBLASTx searches between venom genes and our whole gene set) using MUSCLE (Edgar 2004) with default parameters, with minor manual edits to the alignment to remove any poorly aligned regions. We analyzed the protein alignments using BEAST2

(Bouckaert et al. 2014), setting the site model to 'WAG' for each analysis. We ran each analysis for a minimum of 1×10^8 generations, and evaluated whether runs had reached stationarity using Tracer (Drummond and Rambaut 2007). After discarding the first 10% of samples as burnin, we generated consensus maximum clade credibility trees using TreeAnnotator (distributed with BEAST2).

Analyses of Venom Gland Gene Expression

To explore venom gland gene expression in comparison to other body tissues, raw Illumina RNA-seq reads from all tissues (Supplemental Table S3) were quality trimmed using Trimmomatic v. 0.36 (Bolger et al. 2014) with default settings. We used STAR (Dobin et al. 2013) to align reads to the genome. Raw expression counts were estimated by counting the number of reads that mapped uniquely to a particular annotated transcript using HTSeq-count (Anders et al. 2013). These raw counts were then normalized and filtered in edgeR using TMM normalization (Oshlack et al. 2010; Robinson et al. 2010), and all subsequent analyses were done using these normalized data. To test for significant expression differences between venom gland and body tissues, we performed pairwise comparisons between combined venom gland (i.e., 1 day venom gland, 3 day venom gland, and unextracted venom gland) and body (all other tissues, except for accessory venom gland) tissue sets using an exact test of the binomial distribution estimated in edgeR, integrating tagwise dispersion (Robinson and Oshlack 2010). Genes with differential expression at an FDR value ≤ 0.05 were considered significant. Heatmaps were generated in R using the heatmap function from the R Stats package (R Core Team).

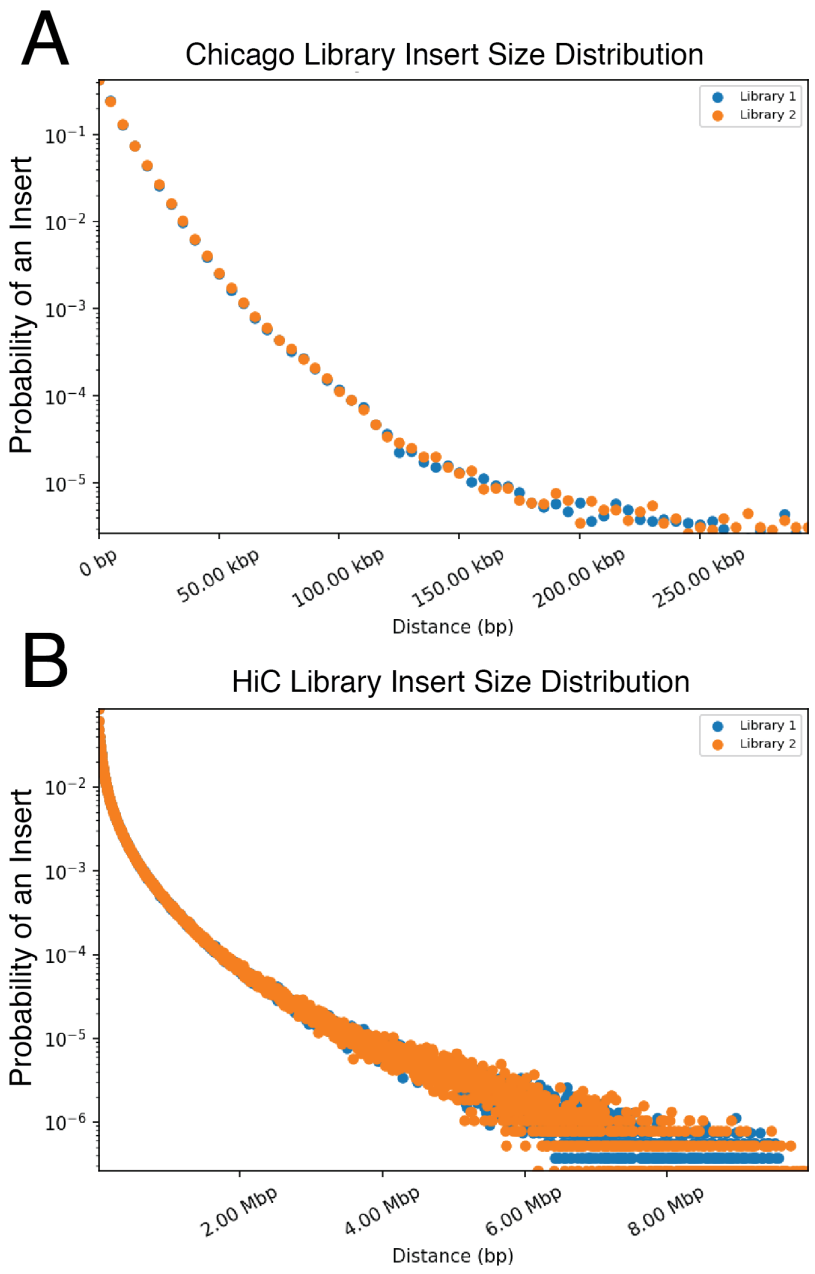
To identify candidate transcription factors regulating venom gene expression, we searched the genome annotation for all genes included on the UniProt (http://www.uniprot.org) reviewed

human transcription factor database, by specifying species = 'Homo sapiens' and reviewed = 'yes' in the advanced search terms. Using this list, we parsed our significant venom gland expressed gene results detailed above for candidate venom gland transcription factors, which showed a pattern of overall low body-wide expression and statistically significant evidence of higher expression in the venom gland, specifically. We identified 12 candidates using this approach, including four members of the *CTF*/*NFI* family of RNA polymerase II core promoter-binding transcription factors (*NFIA,* two isoforms of *NFIB*, and *NFIX*). *NFI* binding sites have been identified upstream of venom genes in several venomous snake taxa, including viperids, elapids, and colubrids (e.g., crotamine/myotoxin in *Crotalus durissus* (Radis-Baptista et al. 2003) and three finger toxins in *Naja sputatrix* (Lachumanan et al. 1998) and *Boiga dendrophila* (Pawlak et al. 2008)). *NFI* family members were also found to be expressed in the venom glands of several species in a previous study exploring putative venom gland transcription factors (Hargreaves et al. 2014), but information about whether they showed venom gland-specific expression was not provided. This set also included the grainyhead-like homolog 1 (*GRHL1*) transcription factor Other significantly up-regulated transcription factors in the venom gland appear to be involved in the unfolded protein stress response of the endoplasmic reticulum and in glandular epithelium development and maintenance (Fig. 4B; Supplemental Table S12). We quantified the distance between predicted binding sites of all transcription factors upregulated in the venom gland (Supplemental Table S12) from 1) venom genes and 2) non-venom genes and compared these distance distributions using *t*-tests.

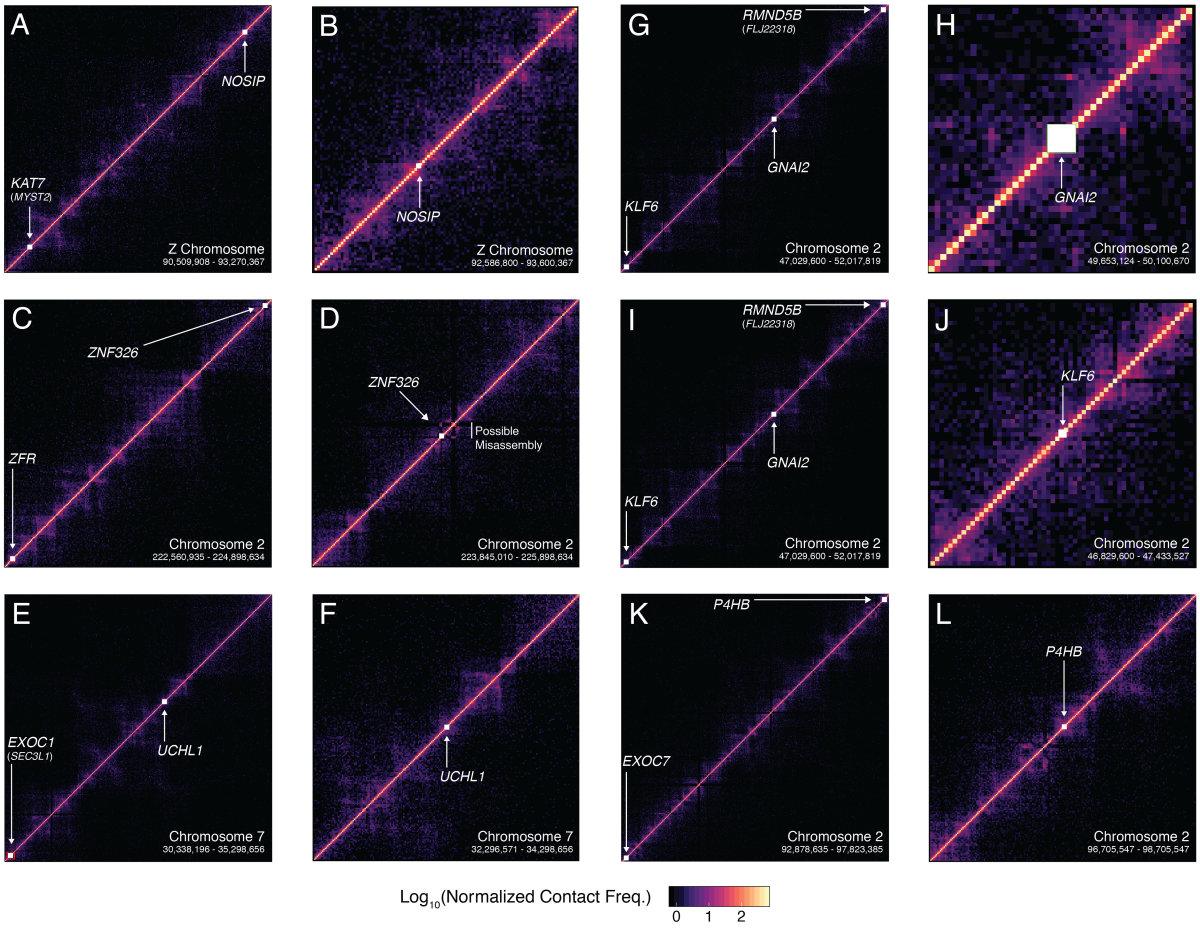
Because four transcription factors of the *NFI* family each showed evidence of venom gland-specificity, we tested the hypothesis that their binding motifs are also upstream of venom genes

by quantifying the number of predicted *NFI* binding sites from PSSM analyses detailed above in the 1 kb upstream region of each venom gene. We also searched for proximity of *GRHL1* binding sites to venom gene regions, as well as all nonvenom genes, using BEDtools (Quinlan and Hall 2010) to calculate the number of predicted binding sites within 100 kb, 50 kb, 10 kb, and 5 kb intervals up and downstream of each gene. Here, we did not confine our search only to promoter regions. To test for enrichment of *NFI* binding sites in the upstream regions of venom genes, we divided the number of predicted binding sites upstream of venom genes by the total length of upstream regions and compared this value to the analogous proportion for upstream regions of all nonvenom genes using a Fisher's exact test (Supplemental Table S13). We performed a similar analysis for *GRHL1* at each interval size, again comparing the density of predicted *GRHL1* binding sites within intervals of venom genes to nonvenom genes (Supplemental Table S13). We also used the Bedtools 'closest' function (Quinlan and Hall 2010) to calculate the distribution of distances between genes and predicted *GRHL1* binding sites.

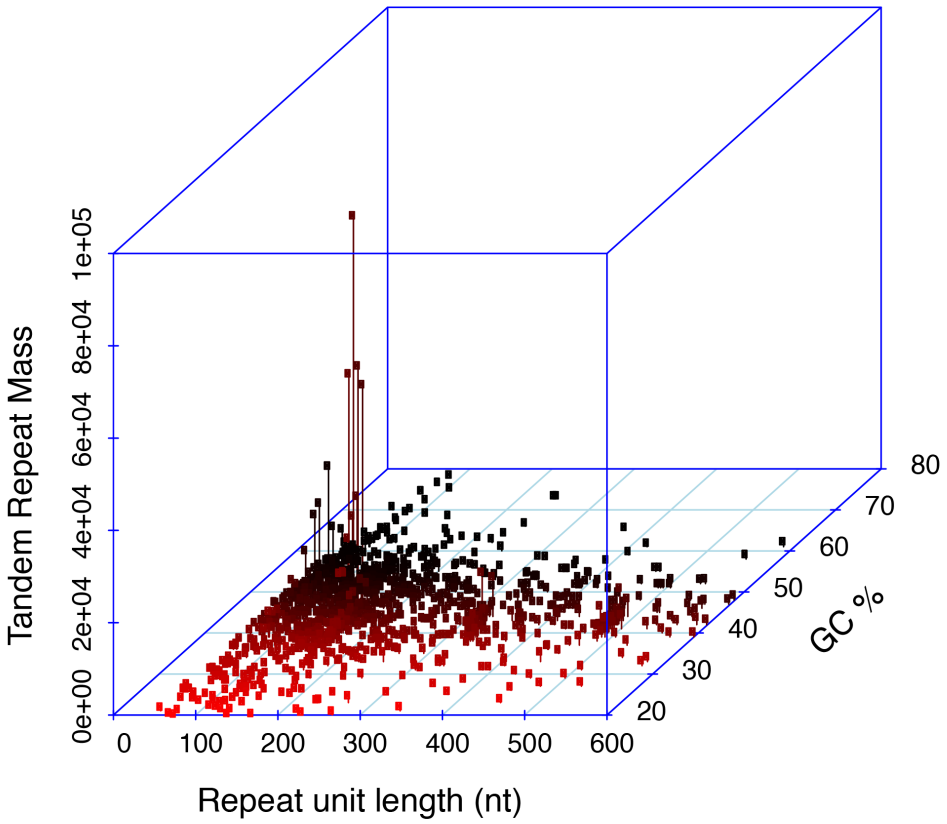
SUPPLEMENTARY FIGURES



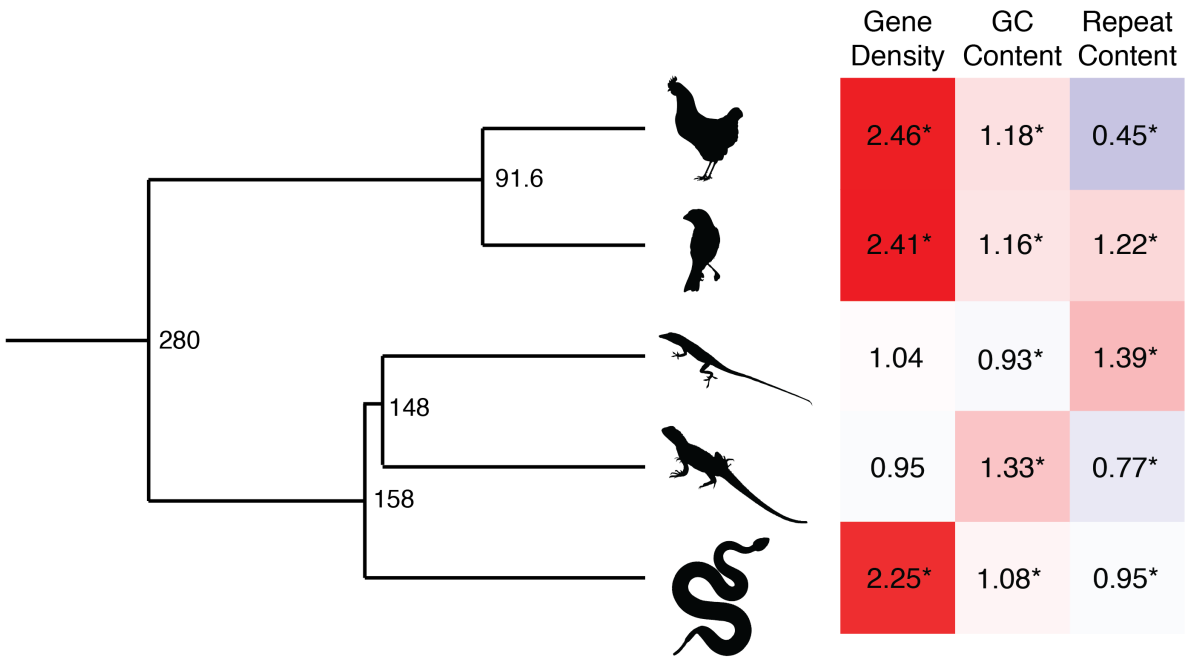
Supplemental Figure S1. Insert size probability distributions used in the Dovetail Genomics HiRise assembly method from paired Chicago (A) and Hi-C (B) datasets.



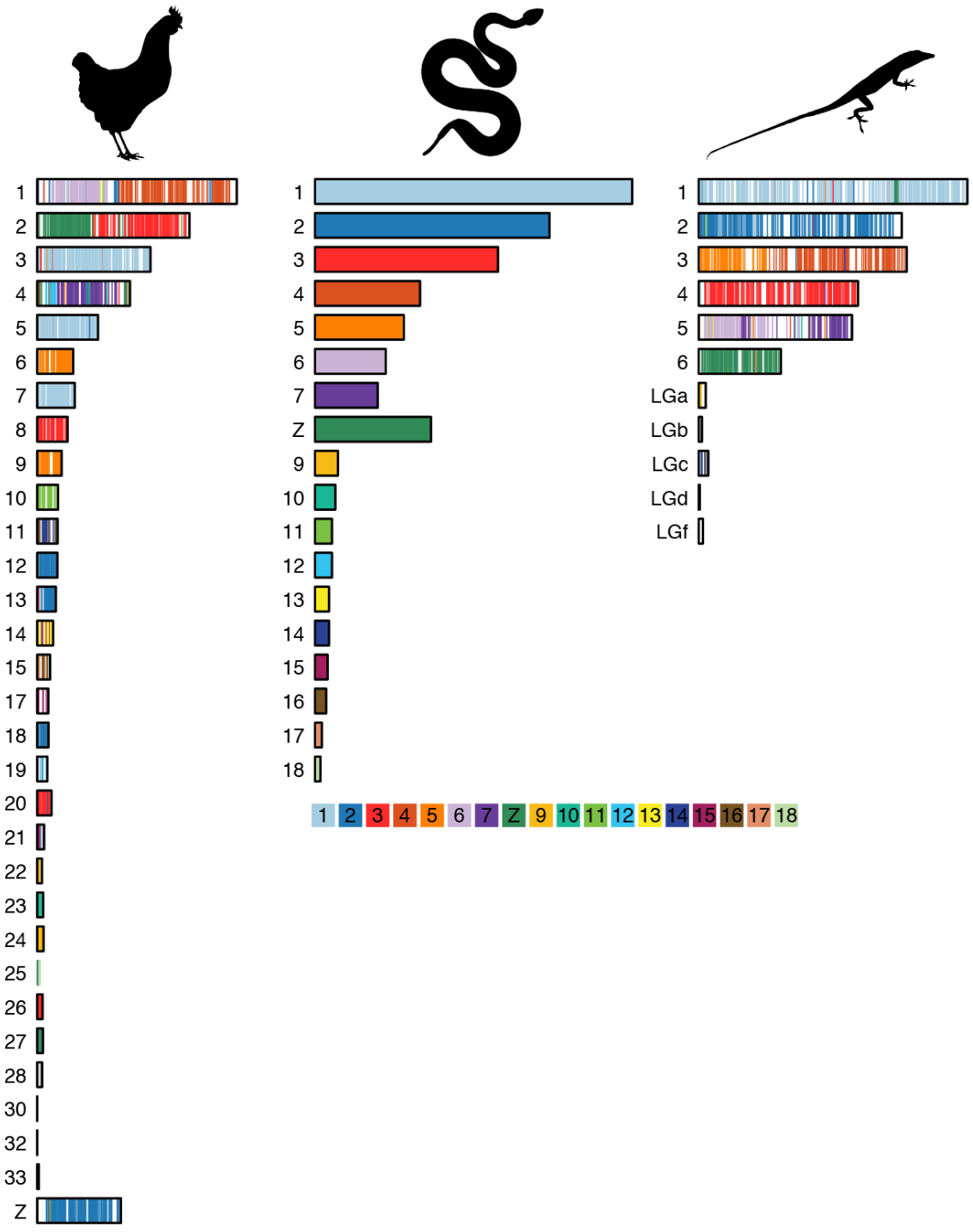
Supplemental Figure S2. Heatmaps of Log_{10} normalized intrachromosomal Hi-C contact frequencies around mapping locations for cDNA markers from *Elaphe quadrivirgata* (Matsubara et al. 2006) in the rattlesnake genome. For each of the six markers, panels showing the contact frequencies between the focal marker and its nearest confirmed marker (see Supplemental Methods), and panels zoomed to the region immediately around the focal marker are shown: *NOSIP* (A-B), *ZNF326* (C-D), *UCHL1* (E-F), *GNAI2* (G-H), *KLF6* (I-J), and *P4HB* (K-L). Marker locations are shown with white squares, and chromosomal coordinates for each panel are shown in the bottom right corner. The location of a potential misassembly error is shown in panel D.



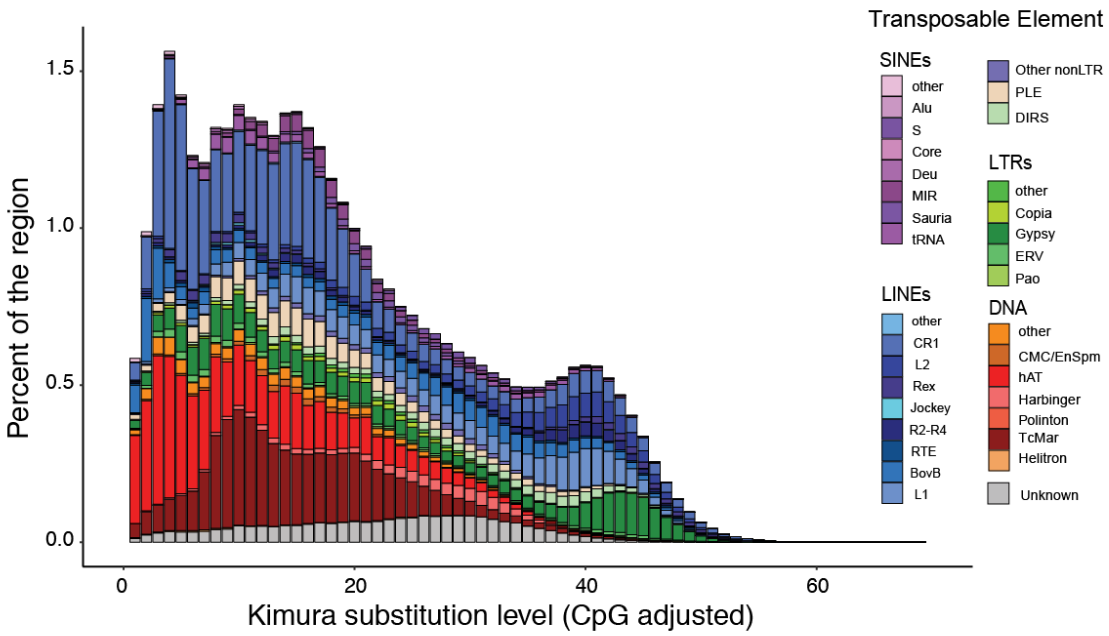
Supplementary Figure S3. Centromeric tandem repeat motif characterized using tandem repeats finder. Analysis of high frequency tandem repeats identified a 164-mer with high relative GC to the genomic background. The y-axis, tandem repeat mass, represents the relative abundance of tandem repeats of a given unit length and GC content.



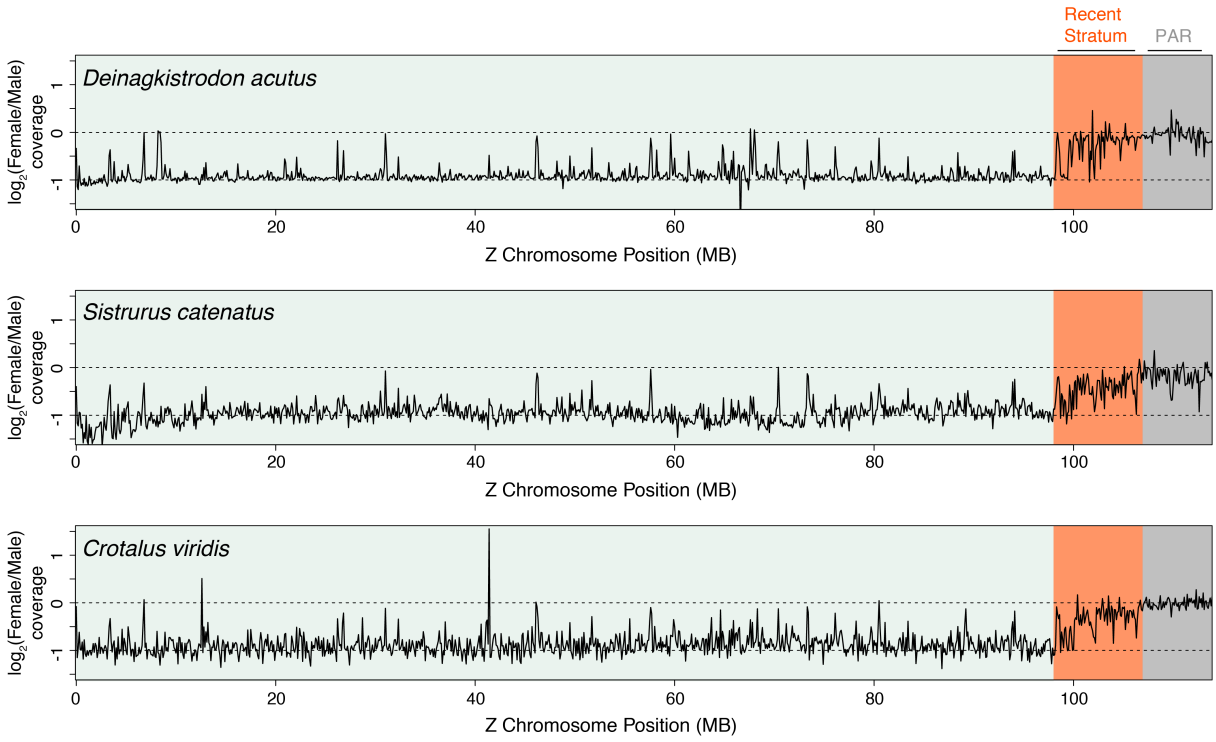
Supplementary Figure S4. Evolutionary patterns of genomic features of microchromosomes among reptiles. Values at nodes on the phylogenetic tree represent the node age in millions of years, and were obtained using median estimates from TimeTree. The heatmap to the right represents the relative abundance of a given measure on microchromosomes versus macrochromosomes within each species (blue values represent greater abundance on macrochromosomes and red values represent greater abundance on microchromosomes). Values in each heatmap cell equal the ratio of each measure on microchromosomes:macrochromosomes, and values with asterisks represent significant differences between microchromosomes and macrochromosomes.



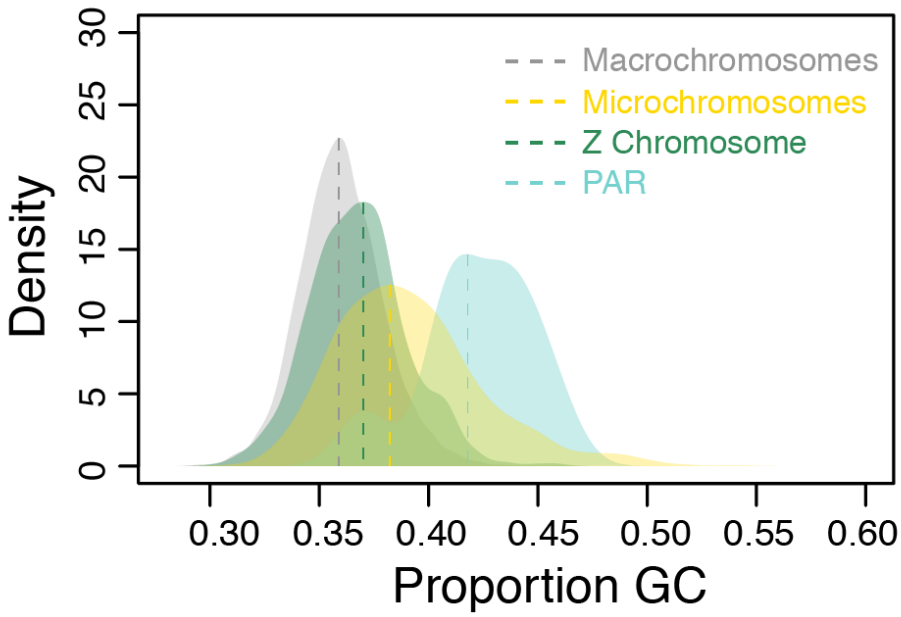
Supplemental Figure S5. Results of gene-based synteny analyses between the chicken (left), rattlesnake (center), and anole lizard (right). Chromosome numbers for each species are shown to the left of the chromosome ideograms, which are scaled by total length. Colors for chromosome paints are based on the rattlesnake genome.



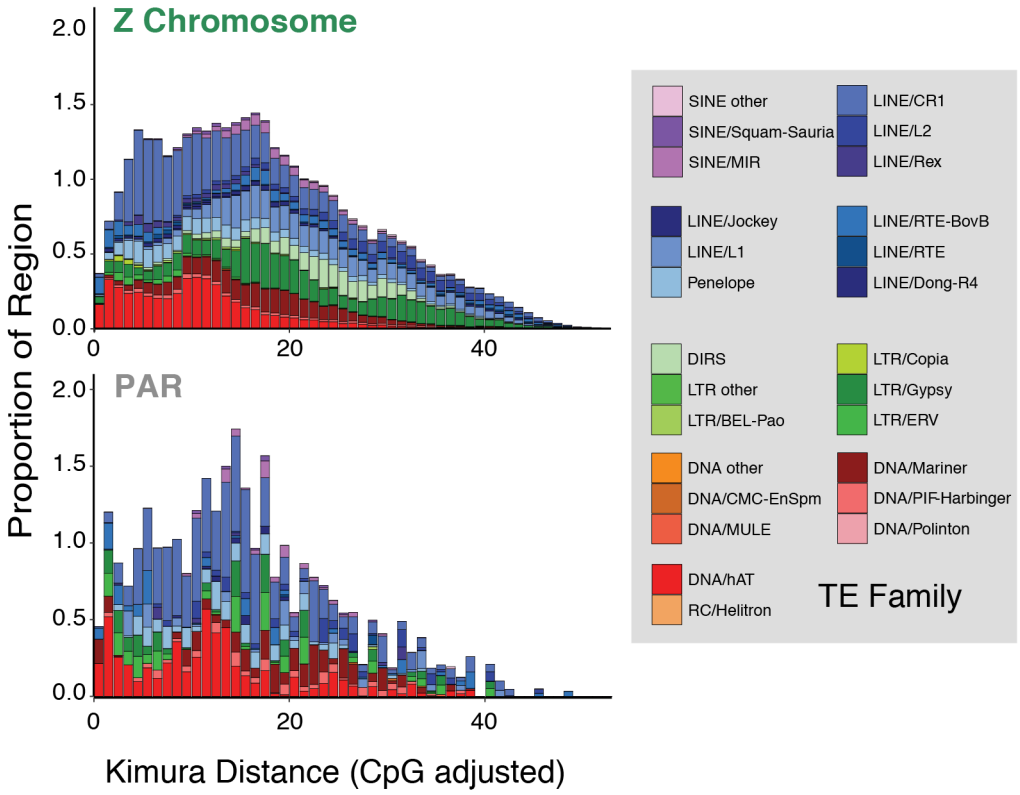
Supplementary Figure S6. Genomic repeat element abundance at a range of relative age values. Age is measured using the Kimura substitution level of transposable elements when compared to a consensus sequence.



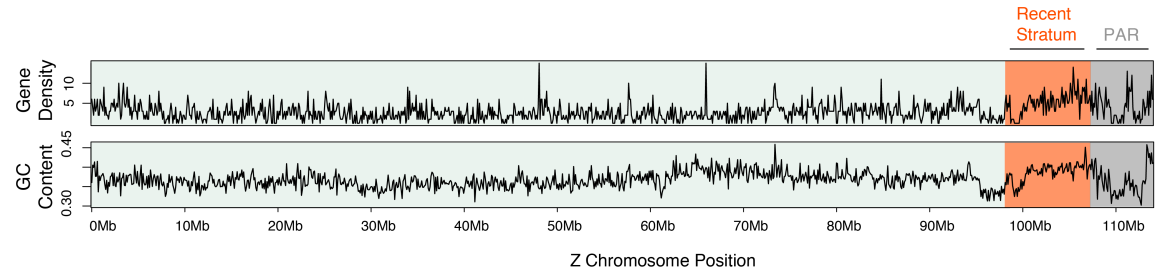
Supplemental Figure S7. Log_2 normalized female/male coverage ratio of pitviper species (Five Pace viper (*Deinagkistrodon acutus*), Pygmy Rattlesnake (*Sistrurus catenatus*), and Prairie Rattlesnake (*Crotalus viridis*), when mapped to the prairie rattlesnake reference genome. The dashed line at zero represents the normalized coverage expectation for diploid loci, and the dashed line at -1 represents the expectation of a hemizygous locus. Ratios are shown show values for each 100 kb window in a sliding window analysis of coverage. Colored backgrounds depict the major regions discussed in the Main Text.



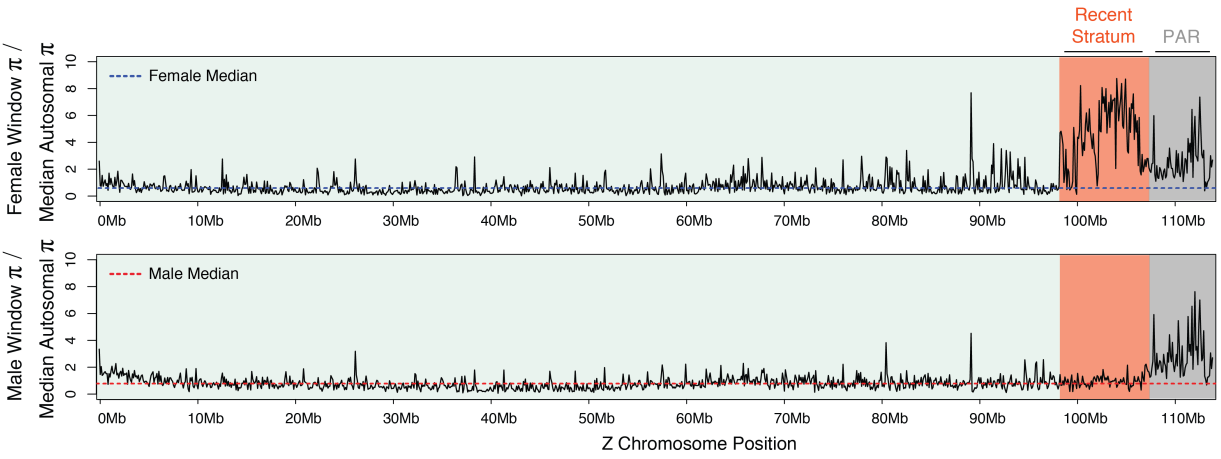
Supplemental Figure S8. Density distributions of GC content across Prairie Rattlesnake chromosomes, showing specific distributions of macrochromosomes, microchromosomes, the Z Chromosome, and the pseudoautosomal region (PAR) of the sex chromosomes, specifically.



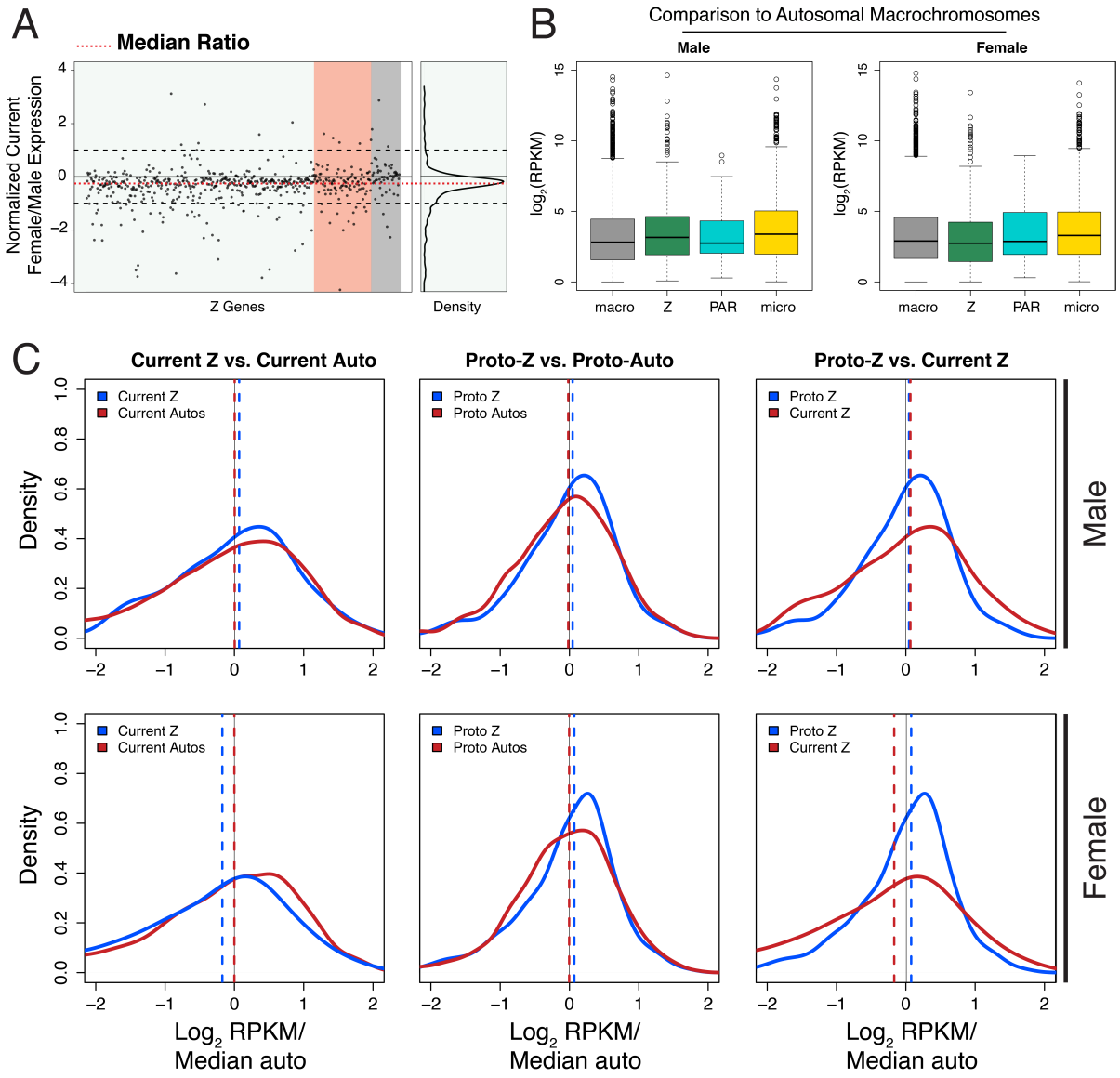
Supplemental Figure S9. Comparative age distributions of proportions of transposable elements (TEs) across the Z Chromosome (upper) and the pseudoautosomal region (PAR; lower) of the rattlesnake Z Chromosome. TE families contributing to proportions in each region at each age are shown at the right.



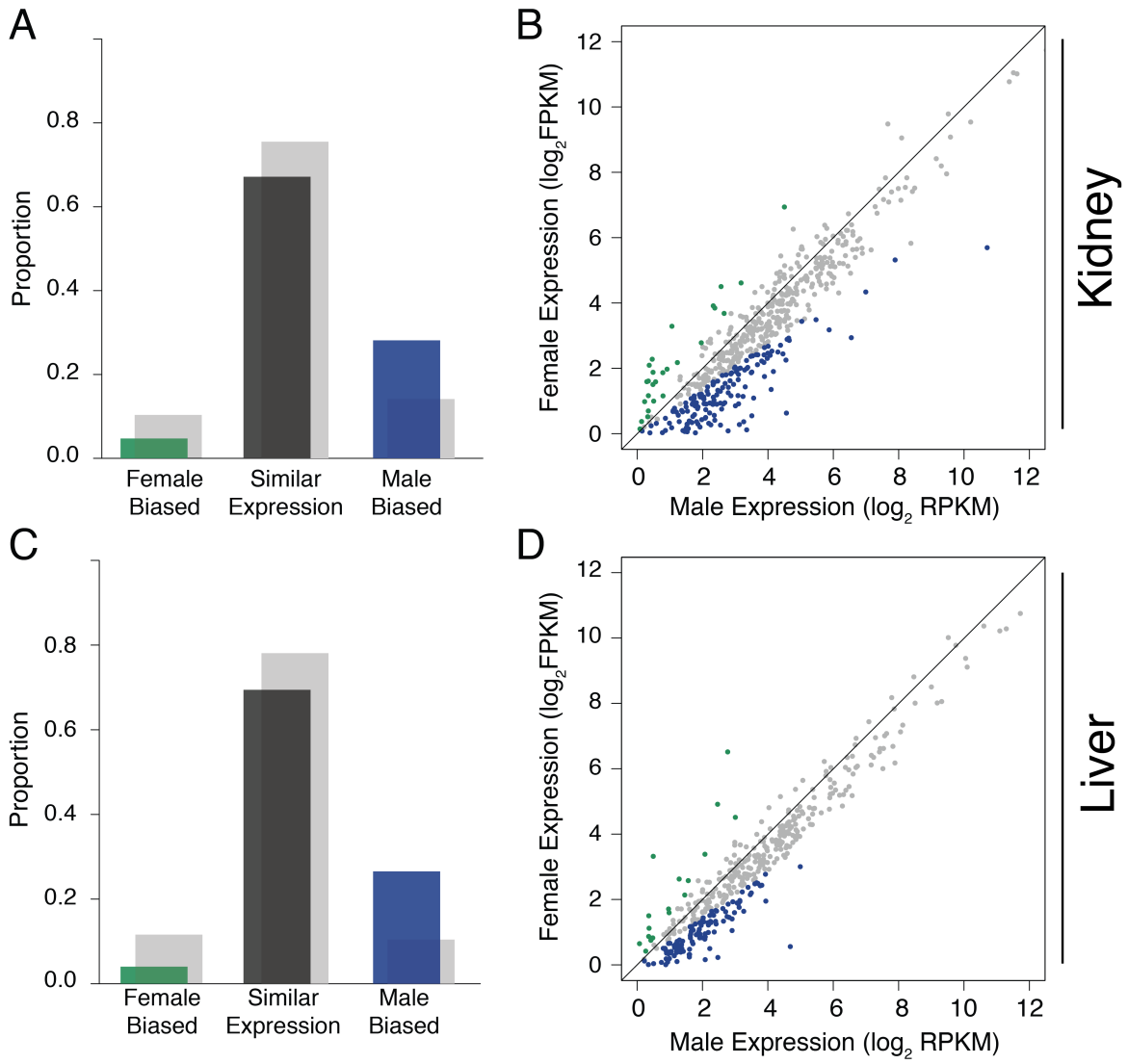
Supplementary Figure S10. 100 kb windowed scans of gene density (measured as number of genes per window) and GC content (i.e., proportion of GC bases within each window) across the Z Chromosome of the prairie rattlesnake. The regions on the Z correspond to those demarcated in Fig. 2 in the main text.



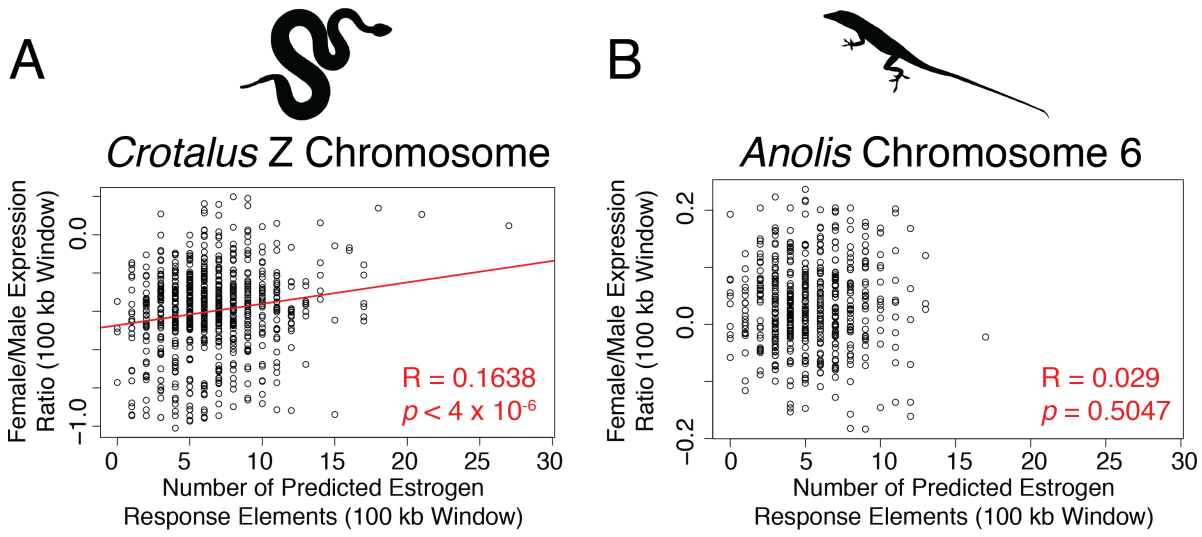
Supplemental Figure S11. 100 kb windowed scans of nucleotide diversity (π) for each sex across the Z Chromosome of the Prairie Rattlesnake. The regions on the Z correspond to those demarcated in Fig. 2 in the main text. Blue and red dashed lines correspond to median female and male values, respectively.



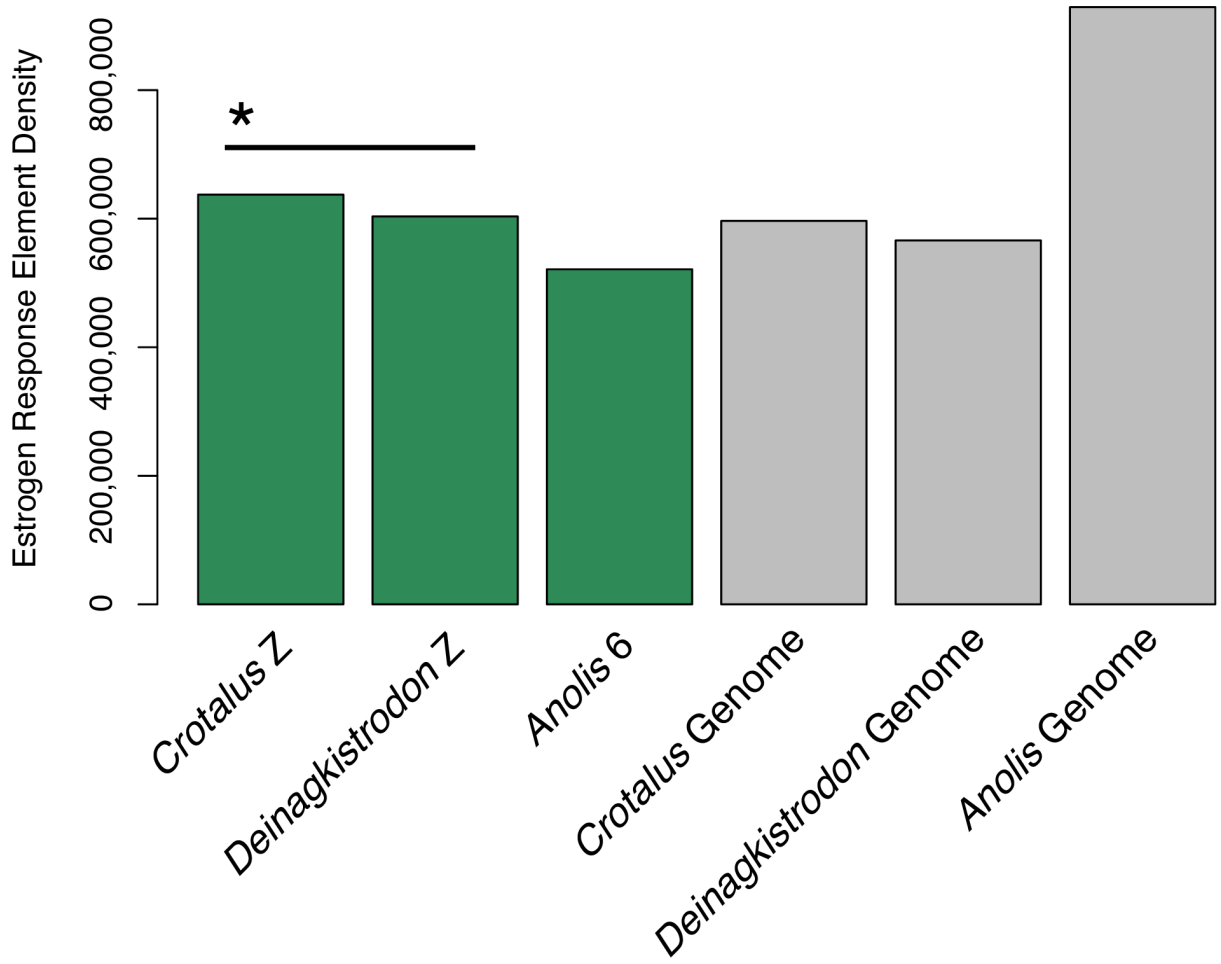
Supplemental Figure S12. Patterns of liver gene expression in females and males across the Z chromosome. (*A*) Log_2 normalized female/male gene expression per gene (black dots) across the Z. The red dashed line is the median ratio, and relative density is shown to the right. (*B*) Gene expression (log_2 RPKM) distributions for male and female across macrochromosomes, Z chromosome, the PAR, and microchromosomes. (*C*) Density plots of current and inferred ancestral patterns of gene expression (log_2 RPKM) in male and female, respectively. Dashed lines represent the median of each distribution.



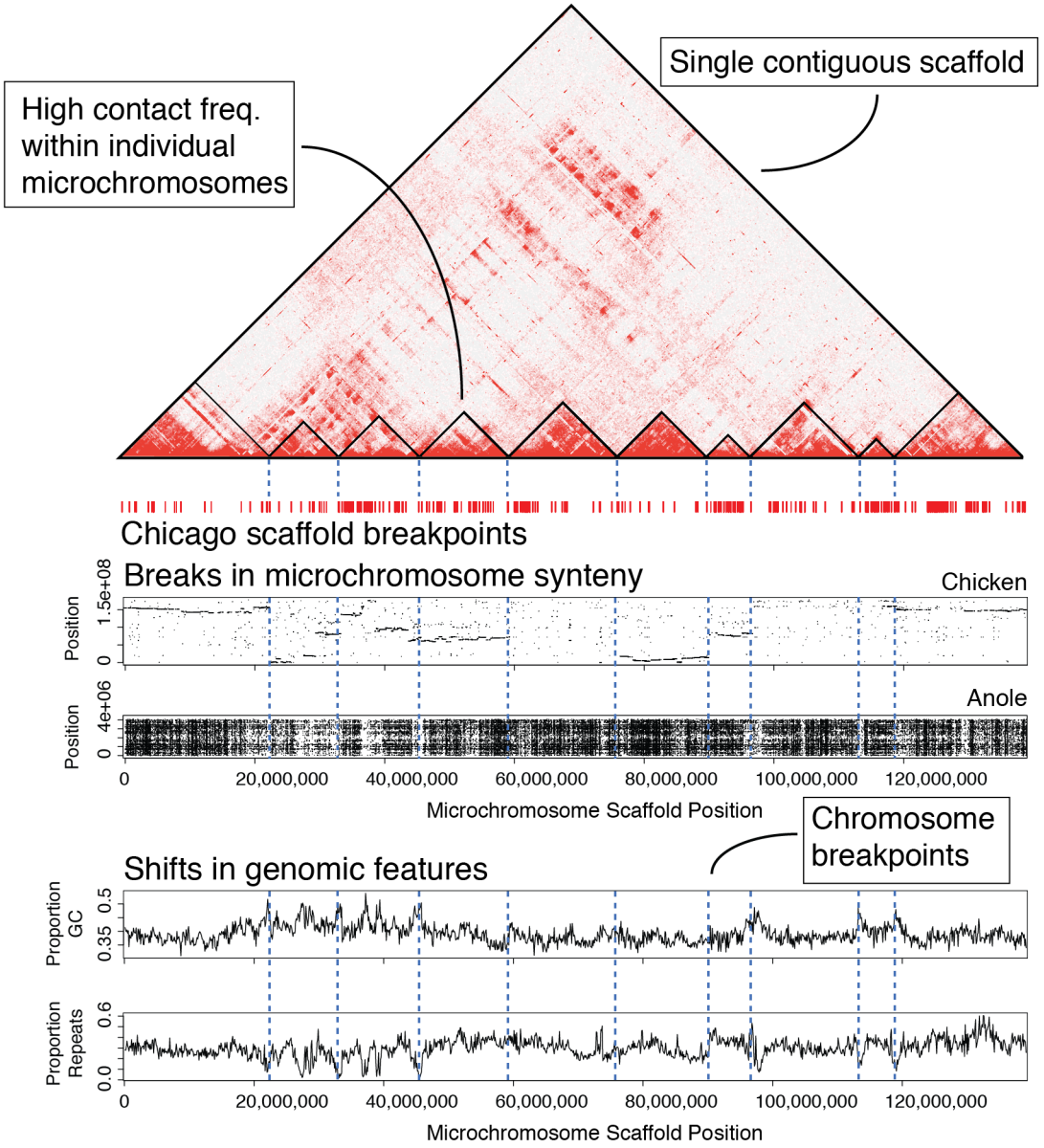
Supplemental Figure S13. Proportions of genes on the Z that exhibit female-biased (i.e., $\log_2$ female/male RPKM > 0.5; green bars), unbiased (dark grey bars), and male-biased (i.e., $\log_2$ female/male RPKM < -0.5; blue bars) expression in the kidney (*A*) and liver (*C*). Light grey bars in the background represent proportions of autosomal genes meeting the same criteria. Scatterplots of male versus female gene expression ($\log_2$ RPKM), with points showing expression of male-biased (blue), unbiased (grey), and female-biased (green) genes for kidney (*B*) and liver (*D*).



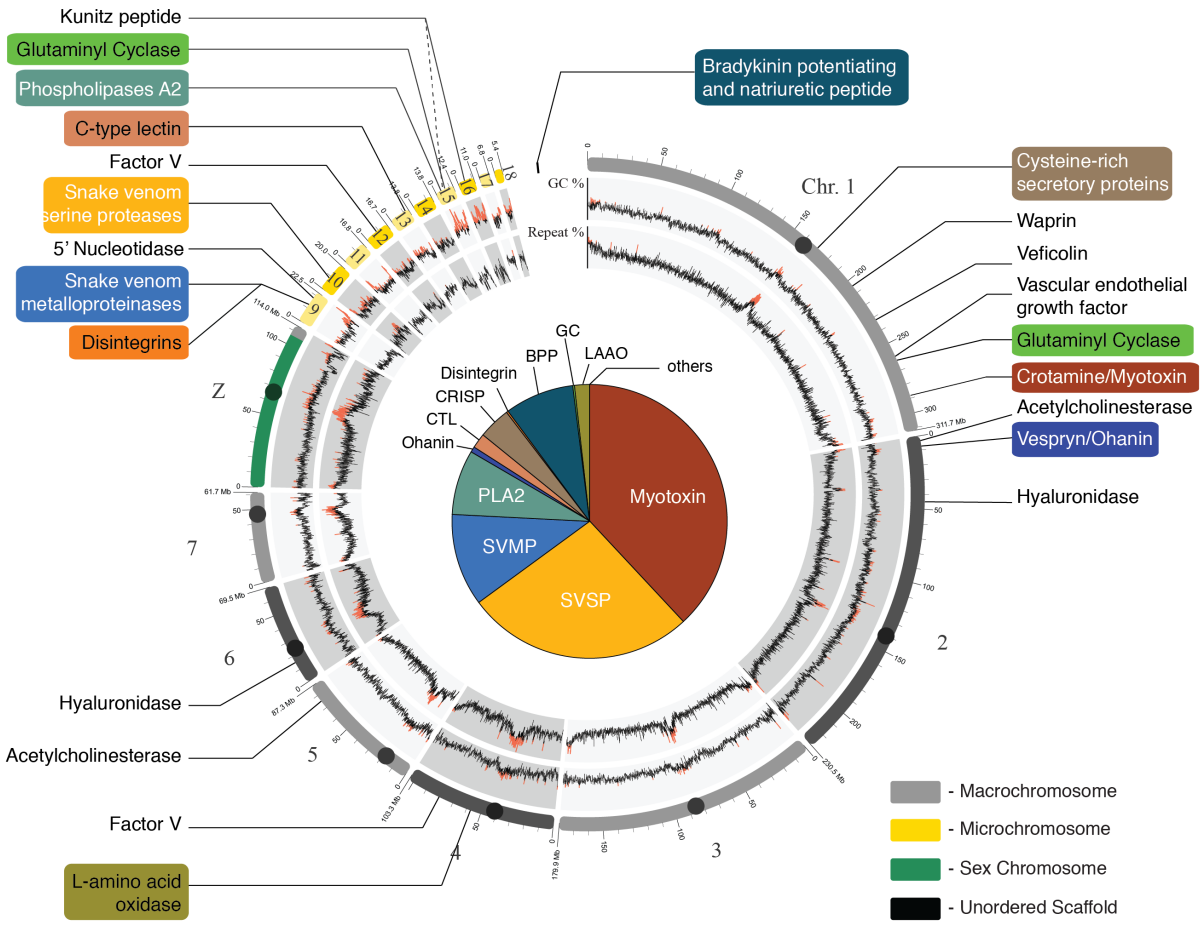
Supplemental Figure S14. Scatterplots of the number of predicted estrogen response elements versus the ratio of $\log_2$(female/male) gene expression in 100 kb windows across the rattlesnake Z Chromosome (*A*) and *Anolis* Chromosome 6 (*B*). The significant positive correlation between variables on the rattlesnake Z is shown by the red line.



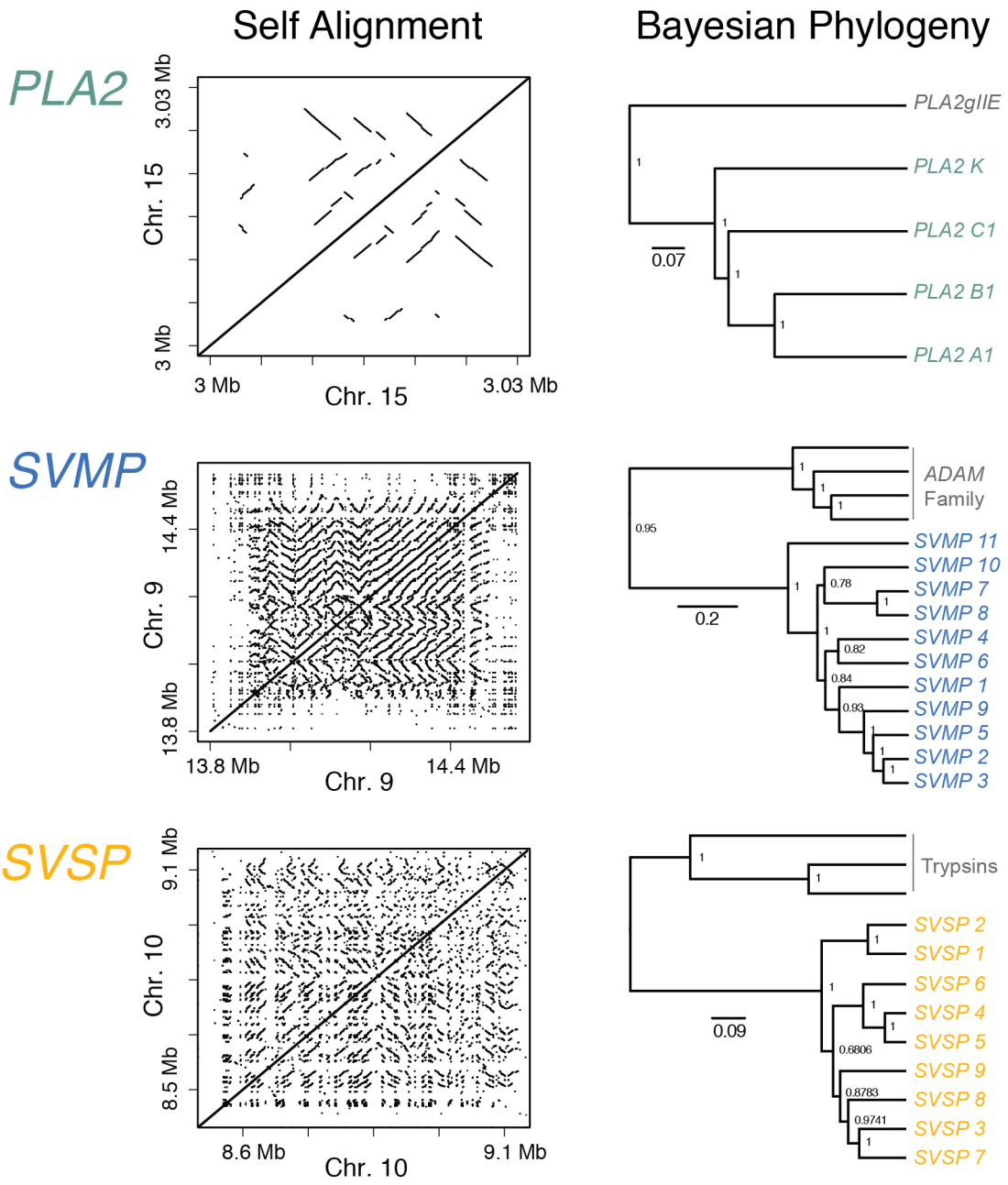
Supplemental Figure S15. Density of estrogen response elements (EREs) across the genomes of squamate species. Density in Z-linked regions of the Prairie Rattlesnake (*Crotalus*) and Five Pace Viper (*Deinagkistrodon*) and the syntenic Anole lizard (*Anolis*) Chromosome 6 regions are depicted in green, and the genomic background for each species is shown in grey bars. The black bar and asterisk depict that EREs are enriched on the pitviper Z Chromosome relative to the homologous autosome in *Anolis* (Chromosome 6).



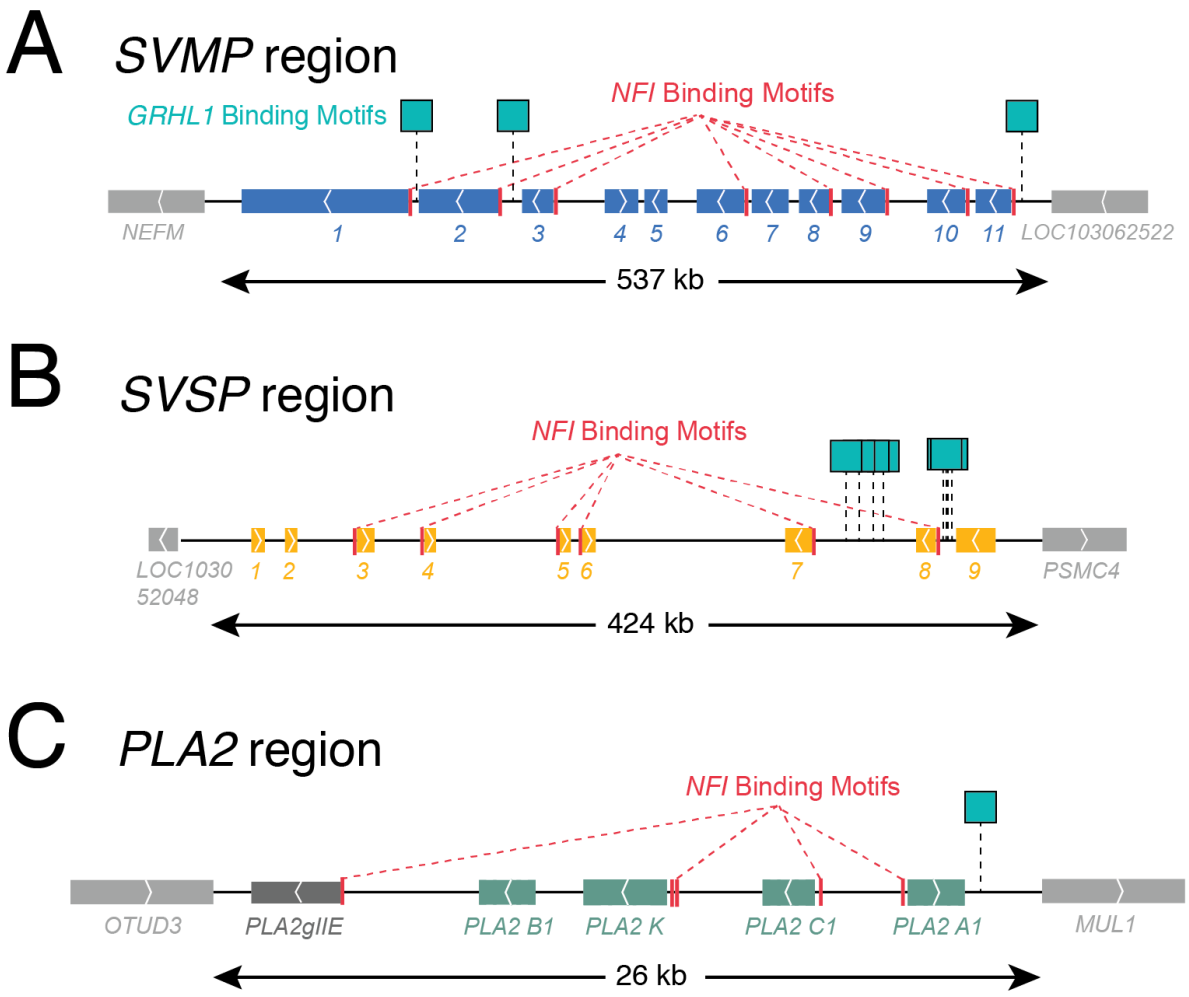
Supplemental Figure S16. Schematic of the initial misassembled microchromosome scaffold. The heatmap panel at the top depicts the high frequency intrachromosomal contacts within individuals microchromosomes, and black triangles depict boundaries between microchromosomes. Breakpoints between Chicago scaffolds used as initial microchromosome breakpoint hypotheses are shown as red dashes below the Hi-C heatmap. The middle two panels show synteny alignments between rattlesnake, chicken, and anole microchromosomes. The bottom two panels show windowed GC and repeat content across microchromosomes. Blue dashed lines in the lower panels show breakpoints between individual microchromosomes.



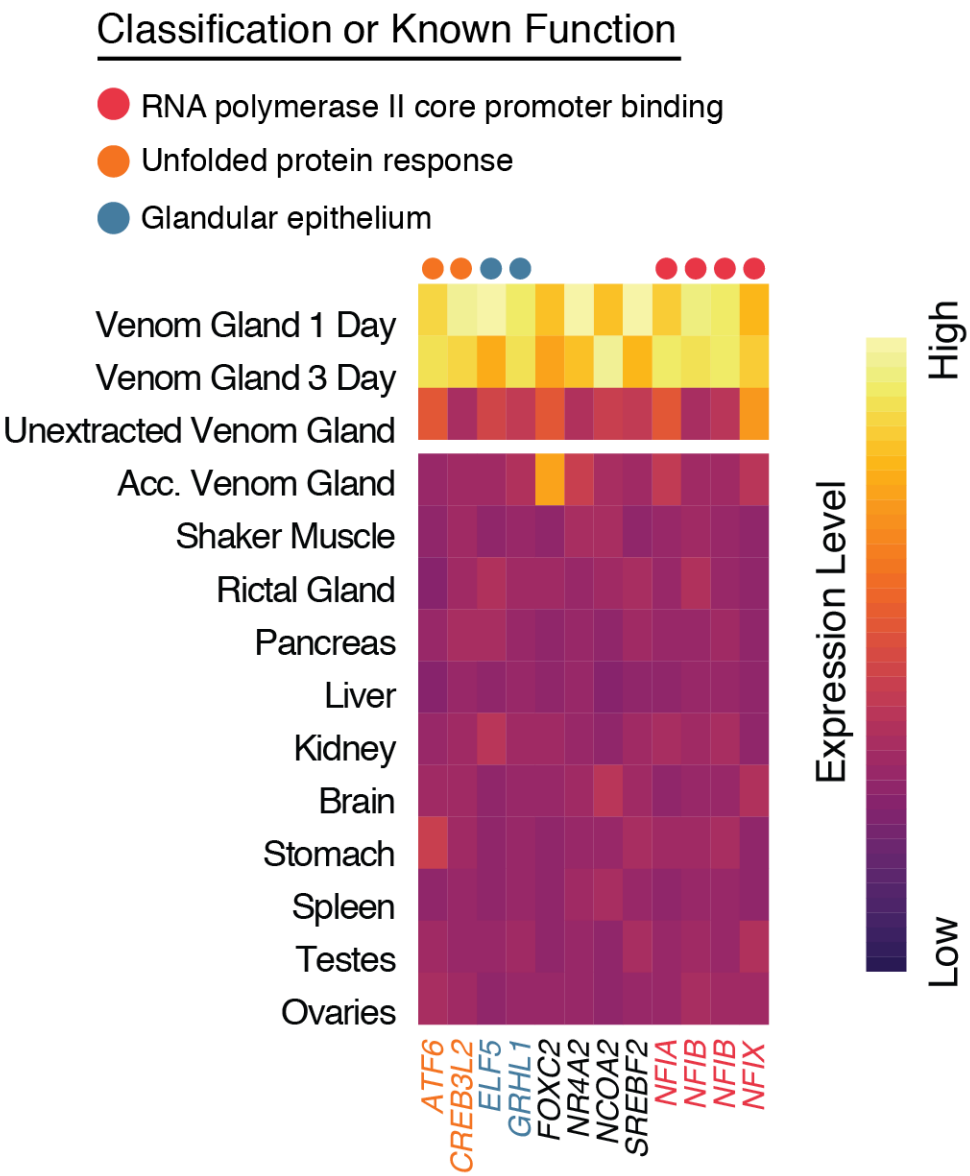
Supplemental Figure S17. Chromosomal locations of snake venom gene families in the prairie rattlesnake. The pie chart in the center depicts the relative abundance of venom families in the prairie rattlesnake proteome. Chromosomal ideograms and windowed scans of GC content (%) and repeat content (%) correspond to those described in Fig. 1 in the main text).



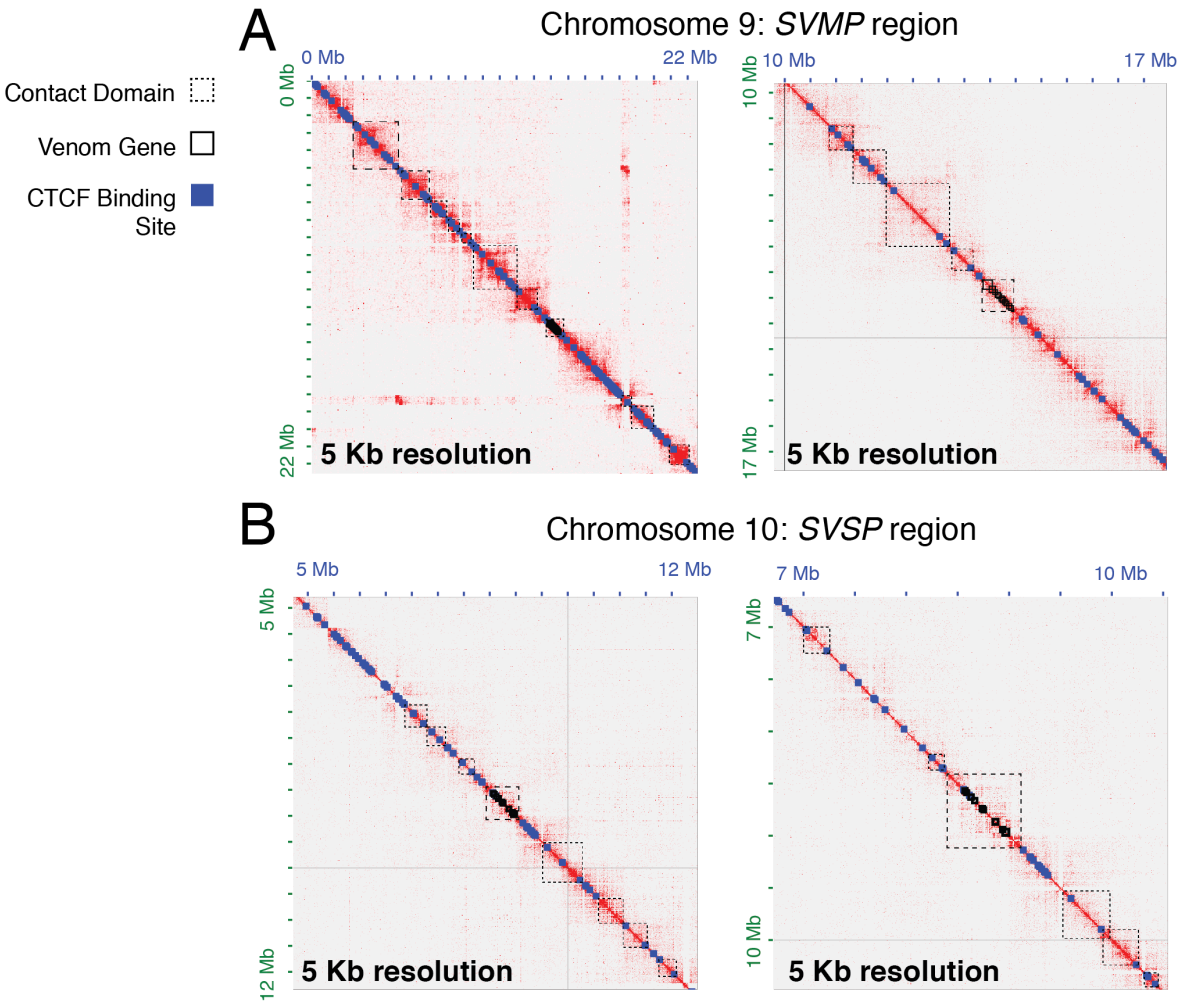
Supplemental Figure S18. Regional self alignment of phospholipase A2 (*PLA2*), snake venom metalloproteinase (*SVMP*), and serine proteinase (*SVSP*) venom gene clusters (left). Parallel and perpendicular lines off of the central diagonal line indicate segmental duplications. Bayesian phylogenetic tree estimates for each of the three gene families constructed based on protein alignments (right), with venom gene paralogs shown in color, and non-venom paralogs in grey. Values at nodes represent posterior probabilities.



Supplemental Figure S19. Structure of annotated *SVMP* (*A*), *SVSP* (*B*), and *PLA2* (*C*) venom gene clusters in the prairie rattlesnake genome. Strandedness (i.e., +/-) of genes is summarized by arrows in the center of each gene. The length of each cluster is shown at the bottom of each panel. Non-venom genes flanking each cluster are shown in grey. In the *PLA2* region, *PLA2gIIE* (non-toxin) is depicted in dark grey. Predicted *NFI* transcription factor binding sites within the 1 kb upstream region of venom genes are shown in red, and locations of predicted *GRHL1* binding sites between genes are shown as turquoise squares.



Supplemental Figure S20. Gene expression across tissues of 12 transcription factors (TFs) significantly upregulated in the venom gland. Broad classifications of known TF functions are annotated at the top of each gene, where applicable.



Supplemental Figure S21. Zoomed out Hi-C heatmaps of the SVMP (*A*) and SVSP (*B*) venom gene regions at two scales (left and right) on microchromosomes, depicting chromatin contact domain structure. Inferred contact domains are represented by dashed black boxes, venom genes in each venom gene region are depicted by solid black boxes, and predicted *CTCF* binding sites are represented by blue squares. Zoomed in versions of these schematics are presented in Fig. 4 in the main text.

Chapter 4

Vertebrate lineages exhibit diverse patterns of transposable element regulation and expression across tissues

Giulia I.M. Pasquesi[1], Blair W. Perry[1], Mike W. Vandewege[2], Drew R. Schield[1], Robert P. Ruggiero[1,3], and Todd A. Castoe[1 ‡]

[1]Department of Biology, 501 S. Nedderman Dr., University of Texas at Arlington, Arlington, TX 76019 USA

[2]Department of Biology, ENMU Station 33, 1500 S Ave K, Eastern New Mexico University, Portales, NM 88130 USA

[3]Department of Biology, One University Plaza, MS 6200, Southeast Missouri State University, Cape Girardeau, MO 63701 USA

Abstract

Transposable elements (TEs) comprise a major fraction of vertebrate genomes, yet little is known about their expression and regulation across tissues, and how this varies across major vertebrate lineages. We present the first comparative analysis integrating TE expression and TE regulatory pathway activity in somatic and gametic tissues for a diverse set of 12 vertebrates. We conduct simultaneous gene and TE expression analyses to characterize patterns of TE expression and TE regulation across vertebrates, and examine relationships between these features. We find remarkable variation in the expression of genes involved in TE negative regulation across tissues and species, yet consistently high expression in germline tissues, particularly in testes. Most vertebrates show comparably high levels of TE regulatory pathway activity across germline tissues, except for mammals which show reduced TE regulatory pathway activity in the ovary. We also find that all vertebrate lineages examined exhibit remarkably high levels of TE-derived transcripts in somatic and gametic tissues, with recently-active TE families showing higher expression in gametic tissues. Although most TE-derived transcripts originate from inactive TE families (and are likely incapable of transposition), such high levels of TE-derived RNA in the cytoplasm may have secondary, unappreciated biological relevance.

Introduction

Transposable elements (TEs) represent the largest identifiable fraction of vertebrate genomes (Smit et al. 2015-2019; Chalopin et al. 2015) despite the fact that they are fundamentally mutagens that propagate through the insertion of new copies. Though ubiquitous, the composition and abundance of TEs is highly variable across vertebrate genomes (Chalopin et al. 2015; Kapusta et al. 2017; Pasquesi et al. 2018; Platt et al. 2018). This variability is the result of complex processes acting at both the levels of TEs and the host genome, including population demography (Lynch and Conery 2003; Neafsey et al. 2004; Xue et al. 2018), the evolutionary history of TEs that have infected host genomes (Kordis and Gubensek 1998; Gilbert et al. 2012; Pasquesi et al. 2018), and the ability of the host to repress TE mobilization (Aravin et al. 2008; Ozata et al. 2019). TE insertions may negatively impact the fitness of their host (Boissinot et al. 2006; Lynch and Walsh 2007) and have been shown to disrupt open reading frames and regulatory regions, alter chromosome structure, and promote genomic rearrangements (Callinan and Batzer 2006; Gasior et al. 2006; Sen et al. 2006; Beck et al. 2011; Vogt et al. 2014; Burns 2017). Yet, increasing evidence for the roles of TEs in rewiring regulatory networks and driving evolutionary innovation (Agrawal et al. 1998; Bourque et al. 2008; Lynch et al. 2015; Chuong et al. 2016; Makałowski et al. 2017; Zeng et al. 2018) counterbalances a simplistic view that TEs are exclusively associated with deleterious impacts on host genomes.

Host genomes have evolved multiple mechanisms to negatively regulate TE activity (reviewed in Goodier 2016), with the primary mechanism being epigenetic modification to silence TE-containing chromatin (Reik 2007; Slotkin and Martienssen 2007; Jacobs et al. 2014). Gonadal germ cell development, however, requires genome-wide erasure of methylation patterns in

primordial germ cells to establish cell potency (Surani et al. 2007). This leaves transposons temporarily unsuppressed by chromatin silencing and thus capable of generating heritable insertions until chromatin structure is reestablished (Hajkova et al. 2002; Kato et al. 2007; Molaro et al. 2014). Safeguarding of the genome against this TE propagation in the germline is primarily accomplished by the PIWI:piRNA (PIWI interacting RNAs) pathway (Aravin and Tuschl 2005; Lim and Kai 2015), a specific small RNA interference mechanism that limits TE proliferation at both the transcriptional level through *de novo* methylation of TE loci and the post-transcriptional level by targeting and degrading TE transcripts (Aravin et al. 2008; Siomi et al. 2011; Weick and Miska 2014).

Previous studies of TE expression and regulation have primarily focused on analyses of germline cell populations, and testes in particular (Shi et al. 2007; Handel and Schimenti 2010). Fewer studies have examined the extent of somatic TE activity (Faulkner et al. 2009; Soumillon et al. 2013; Garcia-Perez et al. 2016; Loreto and Pereira 2017; Faulkner and Billon 2018), although there is evidence for biologically-relevant levels of TE transposition in certain somatic tissues, such as the brain, and for elevated levels of TE activation in somatic tissues associated with ageing or disease (Callinan and Batzer 2006; De Cecco et al. 2013; Bedrosian et al. 2016; Anwar et al. 2017; Faulkner and Garcia-Perez 2017; Kreiling et al. 2017). Currently, our understanding of variation in TE expression and TE regulation across somatic and gametic tissues is based primarily on studies of mammal and bird species (Soumillon et al. 2013), and remarkably little is known about how TE expression and TE regulation may vary across the vertebrate tree of life.

Here, we examine patterns of TE expression and regulation in somatic and gametic tissues from 12 species that represent a sampling of all major vertebrate lineages (Supplementary file 1). We

leverage this sampling to (i) quantify the effects of conserved TE regulatory mechanisms on TE expression levels within and across vertebrate lineages; and (ii) evaluate whether non-mammalian vertebrate species follow mammalian patterns of TE regulation and expression. Our integrated analyses provide new evidence for the uniqueness of mammalian germline biology compared to that of other vertebrates, highlight many features of TE regulation shared across vertebrate lineages, and raise new questions about the biological significance of broad expression of TE-derived transcripts in somatic and gametic tissues that appears to be ubiquitous across vertebrates.

Results

TE regulatory mechanisms are active in somatic and gametic tissues across vertebrate lineages

Our analysis of gene expression for a combined set of 77 genes known to be involved in different TE regulatory mechanisms (Supplementary file 2) demonstrates substantial variation in expression across tissues and species. We find that all categories of negative regulators (i.e., repressors of TE activity) are expressed in both somatic and germline vertebrate tissues at widely varying levels, with the germline tending to show higher average expression (about 2.5 times higher than somatic tissues; Figure 1A and Figure supplement 1). Of all regulatory pathways, the PIWI:piRNA pathway shows higher expression levels in the germline compared to both somatic tissues (i.e., average 16.85-fold higher), and other regulatory gene sets in the germline (1.65-fold higher; Supplementary file 3). In contrast, genes involved in the siRNA pathway show consistently low expression in somatic and germline tissues, while genes involved in transcriptional and post-transcriptional regulation of TE activity show wide variation in expression across species and tissues (Figure 1A; Figure supplement 1). We also find that

negative transcriptional regulators of TE expression on average are expressed at levels similar to the PIWI pathway in the germline, with nine-fold higher expression than in somatic tissues; this is consistent with elevated levels of chromatin modification and the deposition of histone and DNA methylation markers in germline tissues.

Patterns of TE regulatory mechanism activation across tissues and vertebrate lineages

To assess variation in expression patterns of TE regulatory pathways among tissues and across lineages, we used multivariate clustering methods to summarize and differentiate trends of expression. Within-species principal component analyses (PCAs) on gene expression of PIWI pathway genes show distinct, individual clustering of germline tissues in non-mammal species, such that expression patterns in testes and ovaries are distinct from each other and from somatic tissues. In contrast, only testes show a distinct profile in mammals, while PIWI pathway levels in mammalian ovarian tissues fall within the variance of somatic tissues (Figure 1B left panel; Figure supplement 2A). No clear tissue clustering patterns are observed in pathway-specific analyses of the siRNA, transcriptional, and post-transcriptional regulatory pathways (Figure supplement 2B-D), except for a consistent trend of tissue separation driven by the ovary among non-mammal species. Broadly, these other regulatory pathways show cross-tissue profiles similar to those of the PIWI pathway, but with greater variance among somatic tissues (Figure 1B right panel). We further measured the contribution of each gene to the principal component determination, and find that the five genes with the highest contribution scores all belong to the PIWI pathway for the majority of species.

To understand how vertebrate lineages may differ on the basis of how they regulate TEs in the germline, we directly compared variation in expression levels of TE regulatory pathways

between species in germline tissues, specifically. Phylogenetically-correct PCAs for the set of PIWI pathway genes, genes from the three other regulatory mechanisms (i.e., "other pathways"), and all mechanisms combined demonstrate distinct TE regulatory pathway expression patterns in mammals compared to non-mammalian species, largely driven by variation in TE regulatory activity in the ovaries (Figure supplement 4). Comparisons of the first principal components between the PIWI pathway and "other pathways" distinguish testes expression patterns in the alligator and snake species from all other vertebrates (Figure 1C above). In contrast, we find that ovary expression patterns in eutherian mammals cluster independently from other vertebrate species, with the distinction being driven mostly by variation in expression of genes in the PIWI pathway (Figure 1C below).

Between-lineage variation in gametic tissue expression of TE regulatory pathway genes

To further characterize variation in TE regulatory activity across lineages, we calculated Z-scores of expression relative to the mean expression of all genes for a subset of TE regulatory genes with orthologs identified in at least 8 of 12 species (Figure 2). Hierarchical clustering of Z-scores across tissues identified five distinct clusters: vertebrate testes, ovary of non-mammal species, vertebrate brain, mammalian ovary, and a mixed cluster of somatic tissues from diverse lineages (Figure 2). This is particularly evident in mammals, which exhibit the highest Z-scores in the testes. The brain is the only non-germline tissue to exhibit similar expression profiles of TE-silencing genes across all vertebrates. Finally, in contrast to the single testes germline cluster, we find two groups of TE regulatory expression profiles among vertebrate ovaries. The first group includes all non-mammal species, in which expression profiles resemble TE regulation profiles in the testes. The other group includes ovary profiles for mammalian species, in which

expression levels are more similar to somatic tissues. The only exception to this pattern is the human ovary profile, which is clustered with brain. Differences in relative gene expression levels in vertebrate ovaries are further supported by comparative analyses of differential gene expression (DE) between germline and somatic tissues. Multiple genes are significantly differentially expressed in the ovaries of non-mammal species while none are differentially expressed in the mouse or human, and few genes show significant differential expression in the platypus and opossum (Figure supplement 5).

TE-derived transcript abundance across tissues and vertebrate lineages.

To characterize TE transcription levels and composition across vertebrate tissues, we compared expression levels of total TE-derived transcripts (total-TE dataset; Figure supplement 6), as well as transcripts derived only from recently inserted TEs in the genome (recent-TE dataset; Figure supplement 7). Total-TE expression is substantial in both germline and somatic tissues across all species analyzed, although at variable levels within and between species (Figure supplement 6 and 8). For example, while the mean proportion of total-TE derived transcripts is 6.68% across vertebrate tissue transcriptomes, values ranged from 0.26% in the chicken muscle to 23.44% in the opossum spleen (Figure supplement 8; Supplementary file 4). Among sampled species, the chicken and human are characterized by the lowest total-TE average expression levels (2.66% and 2.93% of the total transcriptome, respectively), due mainly to the very low TE transcription levels in somatic tissues (1.52% and 2.36% of the transcriptome on average, respectively). The highest average levels of total TE expression are found in the two snake species, the prairie rattlesnake and boa constrictor (13.75% and 12.16% of the transcriptome, respectively).

Our analyses also show that germline tissues do not always exhibit higher average total-TE expression levels than somatic tissues in vertebrates. For example, the clawed frog, prairie rattlesnake, platypus, and opossum have higher average total-TE expression in somatic tissues compared to germline tissues. In the prairie rattlesnake, platypus, and opossum, this is driven by expression levels that are generally elevated in all or several somatic tissues. In the case of the clawed frog, this pattern is driven by the comparatively low expression levels of total TE transcripts in the germline (which are the lowest across all vertebrate species analyzed). Despite high variance in TE expression levels across tissues, several tissues have relatively consistent trends across species. For example, the testes exhibit greater than two-fold the expression level compared to the ovary (9.63% vs. 4.13%) in all species except the opossum, where expression in the ovaries is higher than in the testes. Additionally, the brain has consistently high total-TE transcription levels across species, which is notably higher than expression in testes (average of 10.05% versus 9.13% of the transcriptomes made up by TE-derived transcripts in the brain and testes, respectively). Conversely, muscle, liver and ovary exhibit consistently low total-TE expression (Figure supplement 8; Supplementary file 4).

Recent-TEs are expressed in both germline and somatic tissues across vertebrates, although at lower levels (0.14% of the transcriptome on average across tissues and species) compared to all TE-derived transcripts (Figure supplement 8-10; Supplementary file 4). Similar to trends in total-TE transcript levels, proportional expression levels of recent-TEs are variable across species and tissues, although lower overall (e.g., from 0.003% in boa muscle to 1.94% in zebrafish testes). In contrast to the total-TE transcript dataset, average recent-TE expression is highest in the testes (0.24%, although though this is driven primarily by high testes expression in the zebrafish),

followed by the small intestine and the brain (0.22% and 0.19%, respectively). We found multiple examples of divergent levels of recent TE transcript expression among species within major vertebrate lineages. For example, while the mouse shows among the highest average recent TE expression levels, the human has low average recent-TE expression levels (Supplementary file 4; Figure supplement 7C and 9).

Overall, our analyses demonstrate that recent and total TE expression levels in somatic tissues are also poor predictors of one another. For example, the small intestine has a relative higher fraction of the transcriptome made up by recent-TEs, while the brain and the spleen have higher fractions of the transcriptome made by TE-derived transcripts that originated from more ancient (and presumably non-mobilizing) TE families (Figure supplement 9; Supplementary file 4). Such differences in the germline tend to be clade-specific. In the testes, mammal and non-mammal species have similar average total-TE expression levels (8.41% vs. 10.24%, respectively), but remarkably different recent-TE expression levels (0.14% and 0.33% respectively). With the exception of the zebrafish, however, recent-TE expression levels are very similar (0.14% and 0.10%). In contrast, mammalian ovaries exhibit more than two-fold greater TE expression than non-mammal species (2.56-fold for recent-TEs and 2.16-fold for the total-TE dataset; Figure supplement 9 and 10, Supplementary file 4). Additionally, there is a positive relationship between the fold-change in TE expression levels (total-TE/recent-TE) between testes and ovaries at the phylogenetic scale (Figure supplement 11), and TE-family composition in testis and ovary is very similar for total-TE transcripts. Yet, analyses of recent-TE transcriptional levels highlight sexually dimorphic TE expression, with some specific TE families appearing to be exclusively expressed in either ovaries or testes (e.g., CR1-LINEs are expressed in the python ovary but not

in the testis, and the reverse pattern is observed in the platypus; Figure supplement 10). Despite tissue-specific expression of some TE families in the recent-TE transcriptome of testes and ovaries, there is a significant association between the relative TE composition of the two germline tissues for both total and recent expression for each species (Figure supplement 12).

Relationships between genome and transcriptome TE composition in germline tissues

To test whether a stochastic model of genome-wide transcription, which predicts that a vast majority of the genome is transcribed at some level (Encode Consortium 2012; Djebali et al. 2012; Hangauer et al. 2013), applies to TEs across vertebrate lineages, we compared relative expression levels of 16 major TE families in the germline and the relative TE composition of the genome for each species analyzed (Supplementary file 5). Our analyses illustrate that each vertebrate species is characterized by a strong significant linear relationship between gametic tissue total-TE expression and the relative genomic abundance of TEs for each respective genome (Figure 3; Figure supplement 13 and Supplementary file 6). We also observe similar trends in relative recent-TE transcriptome composition and relative abundance of recently inserted TE-copies in the genome (Supplementary file 6). However, regression coefficients are generally lower for recent-TEs than for total-TEs, and in some species we find a lack of support for the relationship between genome TE content and TE transcriptional levels in the recent-TE matched comparisons (e.g., chicken, anole, and mouse ovary). This likely stems from multiple instances of TE subfamilies being entirely absent in germline transcriptomes but detectable in the genomes of these species, a trend that is observed in particular in mammals and birds (Figure 3). Finally, comparisons of the relative total genomic TE composition to the relative abundance of recent-TEs in germline transcriptomes found no association in testes and ovaries for most

species. However, mammal species represent an exception to this general trend, as they do exhibit significant linear correlations between genomic TE composition and recent-TE expression in both tissues, although with low regression coefficients (Figure supplement 13; Supplementary file 6).

Relationships between recent TE expression and TE regulatory activity

Considering multiple lines of evidence from our analyses of differential regulation of TE activity in germline tissues, we tested the relationships between the magnitudes of the host response against TEs (particularly the relative activation of the PIWI pathway) and recent-TE expression in germline and somatic tissues (Figure 4; Figure supplement 14A). Phylogenetically corrected (PIC) Spearman rank-order correlation analyses show no significant relationship across species, despite different but non-significant trends across tissues. When all vertebrate tissues are analyzed together, we find a positive association between recent-TE and PIWI pathway gene expression (correlation coefficient $\rho = 0.27$; Figure supplement 14B), as well as positive trends in germline tissues ($\rho = 0.51$; Figure supplement 14C). In contrast, somatic tissues are characterized by a negative correlative trend ($\rho = -0.47$; Figure supplement 14C). The germline exhibits a general trend where, when PIWI pathway genes are expressed at similar levels in testes and ovaries (e.g., in non-avian reptiles), recent-TE expression in the ovaries is lower than in the testes. In the mammalian ovaries, the very low expression level of the PIWI pathway correlates with higher than average recent-TE expression levels (Figure 4; Figure supplement 14).

We further analyzed relationships between recent TE expression and regulation in testes and ovaries across species using PIC linear regression and Spearman rank-order correlation analyses (Figure 5A; Figure supplement 14D and Supplementary file 7 and 8). In the testes we find

significant positive relationships between expression levels of recent-TEs and both PIWI pathway genes and the entire set of genes involved in TE regulation (p-value = 0.02 and 0.004, respectively). There are also weak positive relationships between recent-TE expression and activity of TE regulatory pathways, though these results were not statistically significant based on a Spearman correlation test (ρ= 0.12 and 0.27 for the total regulatory gene set and the PIWI pathway; Figure supplement 14D). We also find no significant relationship between recent-TE expression and regulatory activity in ovaries when all species were analyzed or when mammalian species were excluded. Although not significant, we observe a weak negative trend between PIWI pathway activity and recent-TE expression (ρ= -0.37). Finally, we tested if the response of the PIWI pathway is proportional to recent-TE expression between testes and ovaries across species, but found no evidence for a significant relationship (Figure 5B).

Discussion

A vertebrate-wide perspective on TE expression and TE regulatory pathway activity

To date, studies of TE expression have primarily focused on analysis of male germline and embryonic tissues in mammals (e.g., human and mouse) to understand the mechanisms that regulate TE activity during developmental windows associated with genome-wide DNA demethylation, which are critical for the vertical propagation of TEs (Hajkova et al. 2002; Surani et al. 2007; Ernst et al. 2017). Our integrated analyses of TE regulatory mechanisms and TE expression across germline and somatic tissues shed new light on the variation that exists in both TE expression and regulation among vertebrates, and highlight major differences between germline patterns in mammals compared to other vertebrate lineages. Our results also raise new questions about the relatively high, yet variable, levels of TE-derived transcripts across somatic

and gametic tissues in vertebrates, and underscore the poorly understood relationships between TE regulation and TE transcript expression.

Overlooked complexity of TE negative regulation in the vertebrate germline

Despite major differences in evolutionary history and genomic composition of vertebrate TEs, we find that active repression of TEs via multiple conserved regulatory pathways appears to be a shared feature of vertebrates in both somatic and gametic tissues. However, the activation of TE repression mechanisms is particularly variable in ovaries across vertebrate lineages compared to more conserved patterns of activation in the testes. Mammals in particular appear to regulate TE expression in the ovary at a low level comparable to that of somatic tissues, which directly contrasts the high regulation observed in the ovary of other vertebrates. This reduced level of TE regulation in mammalian ovaries may explain why polymorphic TE insertions that have developmental origins in the female early embryo and late germline exhibit the highest transmission rates in mice (Richardson et al. 2017).

These findings raise intriguing questions regarding the biological basis and selective drivers that underlie reduced ovarian TE regulation in mammals compared to other vertebrate lineages. One potential explanation may derive from differences in mitotic rates in mammals; previous studies of TE activity and repression have focused specifically on the male germline over the female germline due to higher mitotic and meiotic rates during spermatogenesis (Handel and Schimenti 2010), and other previous studies have indicated that TE activity positively correlates with tissue-specific cell mitotic rates (Navarro et al. 2019). Across animals, ovaries are characterized by a cell population in meiotic arrest (Sagata 1996), but differences likely exist in the frequency and magnitude of oocyte activation across lineages (Abrieu et al. 2001). Species characterized by

the deposition of numerous eggs (e.g., fish and amphibians) possess an ovarian germline stem cell (OGSCs) population in order to replenish the oocyte pool (Hanna and Hennebold 2014). Whether the presence of OGSCs is a shared feature among vertebrates is still controversial – their presence in mammals is debated, and information for other vertebrate lineages is not available (Hanna and Hennebold 2014). Future studies that advance our understanding of the variation in key features of ovarian biology across vertebrates, including mitotic and meiotic rates as well as the presence of OGSCs, may prove valuable for examining links between variation in characteristics of ovarian biology and the activity of TE regulatory mechanisms across vertebrate lineages.

Few previous studies have focused on TE regulatory mechanism outside of the mammalian germline (Watanabe et al. 2008; Lim et al. 2013; Malki et al. 2014), which limits the context for comparison of our results across tissues in vertebrates. Our conclusion that PIWI pathway genes are expressed at similar levels in testes and ovaries is broadly consistent with previous studies in the zebrafish, clawed frog, and anole (Houwing et al. 2007; Kirino et al. 2009; Zhang et al. 2017), while expression of PIWI mRNAs or piRNAs have not been detected in previous studies of chicken ovaries (Sun et al. 2017). Interestingly, the zebrafish is also known to produce sex-specific piRNAs from distinct genomic TE loci (Zhou et al. 2010); if this mechanism exists in other vertebrates, it may provide an explanation for sexually dimorphic expression of recent-TEs in the germline.

TE regulatory pathways do not clearly demarcate somatic and gametic tissues

Our comparative analyses illustrate that expression of genes involved in the negative regulation of TEs is not limited to the germline, and we find evidence for the activation of many of these

pathways across somatic tissues in diverse vertebrate species. Among the four categories of TE regulatory mechanisms analyzed, only expression levels of the PIWI:piRNA pathway consistently discriminated at least one germline tissue from somatic tissues based on variation in across-tissue gene expression. In contrast, endogenous small interfering RNA (siRNA), transcriptional, and post-transcriptional pathways are all characterized by relatively consistent expression levels across germline and somatic tissues. Our analyses therefore support the canonical view of PIWI pathway genes and associated piRNAs are a hallmark of gametic tissues, and the vanguard of germline genome integrity.

TE expression and repression mechanisms have been extensively studied in somatic tissues, but mostly in association with cancer, aging, and other diseases (Kazazian 1998; Burns 2017; Kreiling et al. 2017). Those studies led to the collective view that, because of the threat TE mobilization poses to genome integrity and structure, their expression is severely restricted at both transcriptional and post-transcriptional levels. Subsequent studies found exceptions to this pattern in the central nervous system and in specific developmental stages, where expression of specific elements promotes cellular mosaicism and the correct execution of cell specification pathways, respectively (Baillie et al. 2011; Weissman and Gage 2016; Hackett et al. 2017). In brain tissues, we find a single, distinct profile of TE regulation common to all vertebrates characterized by relatively higher expression of transcriptional regulators (e.g., TRIM28 and methyltransferases). This finding suggests that a shared pattern of TE activity may exist in the central nervous system of all vertebrates. Broadly, our findings indicate that genes traditionally associated with the germline (e.g., genes that belong to the PIWI:piRNA pathway; Ponnusamy et al. 2017) are expressed in somatic tissues, although often at relatively low levels, and vice-versa

(e.g., genes in the siRNA pathway; Stein et al. 2015). These results further support the roles of TE regulation in somatic tissues, possibly through the evolution of compensatory or reinforcing mechanisms, or the cooption of existing mechanisms for TE regulation (Levine et al. 2016).

Interpretations of TE-derived transcript abundance and relationships to TE regulation

Our analyses demonstrate that TE-derived transcripts comprise a notably large fraction of the transcriptomes of both germline and somatic tissues across vertebrate lineages. We expected that a large majority of TE-derived transcripts would originate from recent active TE families, but this is not what we found in any of the species analyzed. Instead, for each vertebrate analyzed, TE-derived transcripts originate from a variety of both recent and ancient TEs families. These findings, corroborated by the identification of similar relative composition of genomes and TE transcriptomes across species, support hypotheses from studies in mammalian systems evoking a stochastic transcription model, in which the majority of the genome is pervasively transcribed (Encode Consortium 2012; Hangauer et al. 2013). Although the majority of TE-derived transcripts may not have biological activity related to insertional mutagenesis or replication, it remains an open question whether this abundant pool of TE-derived cellular RNAs have other biologically relevant impacts in gene regulation (e.g., lnRNAs and microRNAs), or unappreciated roles due to their sheer abundance (e.g., mass-effect competition for RNA catabolic processes, RNA metabolism, and interference with translation) or due to their cooption as regulatory elements (van de Lagemaat et al. 2003; Lippman et al. 2004; Cordaux and Batzer 2009; Chuong et al. 2016).

To focus on TE-derived transcripts that are more likely to be biologically relevant sources of mutation and transpositional activity, we restricted our analyses to transcripts that originated

only from recently-active TEs (i.e., recent-TEs). These recent and active TEs are likely to be more strongly targeted by negative regulatory mechanisms (Vandewege et al. 2016; Sun et al. 2017; Xue et al. 2018). We find that recent-TEs are expressed in both germline and somatic tissues across vertebrates, although at far lower levels (that average 0.14% of the transcriptome) compared to total TE-derived transcripts (that average 6.86%). Recent-TE expression tends to be highest in the testes, followed by the small intestine and the brain. Our results also indicate that patterns of recent-TE expression in mammals are unique among vertebrates analyzed by having relatively higher levels of expression in the ovaries, such that mammalian ovaries and testes show similar expression levels. We also identified multiple examples of highly divergent levels of recent-TE transcript expression among species within major lineages, suggesting that substantial variation likely exists across species within major vertebrate lineages.

Our analyses of the relationships between recent-TE expression and TE regulatory pathway activity provide evidence for divergent patterns between gametic tissues across vertebrates. In the male germline, there is a positive relationship across vertebrates between expression levels of recent-TEs and TE regulatory pathway activity. This suggests that the relative activation levels of TE repressive mechanisms may be tuned to be proportional to the threat posed by TE activity, and is also consistent with the concept that higher TE activity elicits a stronger response in the host (Reznik et al. 2019). In contrast, this trend was non-significant and weakly negative across vertebrate ovarian tissues. These findings highlight further questions about the unique biology of ovaries, how this biology may vary across vertebrates, and how it might relate to differences in the potentially deleterious effects of TE activity.

Conclusions and future directions

Our comparative analyses of TE regulation and expression across vertebrate lineages suggests that active repression of TEs is accomplished by multiple conserved mechanisms, and represents a shared feature among germline and somatic vertebrate tissues. Our results also highlight highly unique sexually dimorphic TE-associated biology specific to gametic tissues. We find that patterns of TE regulation are remarkably distinct in mammalian ovarian tissues compared to other vertebrates, and that a shift towards decreased TE regulatory activity in ovaries occurred early in the evolution of the mammalian lineage. These findings, together with other differences in TE regulation and expression identified among vertebrate lineages underscore the importance of studies of diverse vertebrate lineages for understanding the uniqueness of mammalian biology, and demonstrate the potential shortcomings of broad assumptions that diverse vertebrate model systems share common biological features and regulatory mechanisms. Our findings also underscore challenges in understanding the relevance of TE-derived transcript abundance from analysis of RNAseq data alone, and argue for future integration of approaches that quantify transpositionally competent TE-derived transcripts (Deininger et al. 2017) and other functional data (Sun et al. 2017; Goubert et al. 2019; He et al. 2019).

Materials and Methods

We used previously published poly-A-selected RNAseq datasets to compare expression levels of TE-derived transcripts in testes, ovaries, and up to 6 somatic tissues (brain, heart, kidney, liver, muscle, spleen and small intestine) across 12 vertebrate species that included representatives of fish, amphibians, reptiles, and mammals (Supplementary file 1). Raw RNAseq data were first filtered for prokaryote and eukaryote rRNA transcripts using SortMeRNA v2.1 (Kopylova et al.

2012), and then quality and adapter trimmed in Trimmomatic 0.36 (Bolger et al. 2014). Detailed information for each analysis is provided in the supplemental methods. For each species, reads were mapped using STAR v2.7.0f (Dobin et al. 2013) to the latest genome version and annotation .gff files available on the NCBI Genome database (Sayers et al. 2019). STAR was run according to default parameters, allowing for a maximum of 100 mapped reads per locus.

Gene and TE-derived transcript expression levels were simultaneously estimated using *TEtranscript* (Jin et al. 2015). To assign mapped reads to a genomic locus, *TEtranscript* requires two annotation files that specify gene and repeat element coordinates, respectively. TE index structures were built from RepeatMasker *.out* files (Smit et al. 2015-2019), and gene index structures were built from the same gene annotation files used when running STAR (detailed information on the protocol used to build the input *.gtf* files are provided in the supplemental methods). *TEtranscript* was run using default parameters, the *–multi* multi-mapper mode flag, and specifying whether transcriptome data was stranded or not. Expression levels of TE-derived reads that originated from recently active (and thus, potentially capable of transposition) TE copies were estimated in a second, separate analysis. In this case, we provided *TEtranscript* with a filtered *.gtf* annotation file that contains only TE loci with less than 2% Kimura 2-parameter distance consensus (we refer to this as the "recent-TE" dataset). For each species, normalization of TE-derived and gene-derived raw read counts across tissues was performed using the *estimateSizeFactors-estimateDispersions-counts(normalized=TRUE)* functions in *DESeq2* v1.20 (Love et al. 2014) after removing elements with less than 10 mapped reads across samples.

To assess the relationships between TE expression levels and TE regulatory pathway gene levels, we compared recent-TE expression levels to 5 sets of TE regulatory genes: (i) genes participating

in the PIWI:piRNA pathway (Carbon et al. 2009; PIWI pathway hereafter); (ii) genes involved in the small RNA interference pathway (Carbon et al. 2009; siRNA pathway); (iii) genes involved in transcriptional regulation of TEs (e.g., responsible for de novo DNA or histone methylation; e.g., Hutchins and Pei 2015; Levine et al. 2016); (iv) other genes previously identified to negatively impact TE mobilization and/or insertion at the post-transcriptional level (e.g., Apobec; Goodier 2016); and (v) the combined magnitude all TE regulatory genes (which corresponds to all 77 conserved genes belonging the four previous sets).

Patterns of within-species variation in expression levels across tissues were assessed by performing principal component analyses (PCAs) on blind variance stabilizing transformed data (Anders and Huber 2010). Because of the heterogeneous nature of our data, between species comparisons were performed using percentages of the total transcriptome following normalization of read counts to limit biases due to different methods of tissue processing, library preparation, sequencing technology and dataset quality (Sudmant et al. 2015; Dunn et al. 2018). To investigate relationships in expression patters across vertebrates, we performed phylogenetic independent contrast (PIC) linear regressions and PCAs using the *phytools* package in R (Revell 2012). Additional methodological details for analyses performed in this study are provided in the supplemental methods.

Acknowledgements

Support was provided from startup funds from the University of Texas at Arlington to TAC and the Society for the Study of Evolution (to GIMP). We acknowledge the Texas Advanced Computing Center (TACC) for providing access to computational resources.

Figures

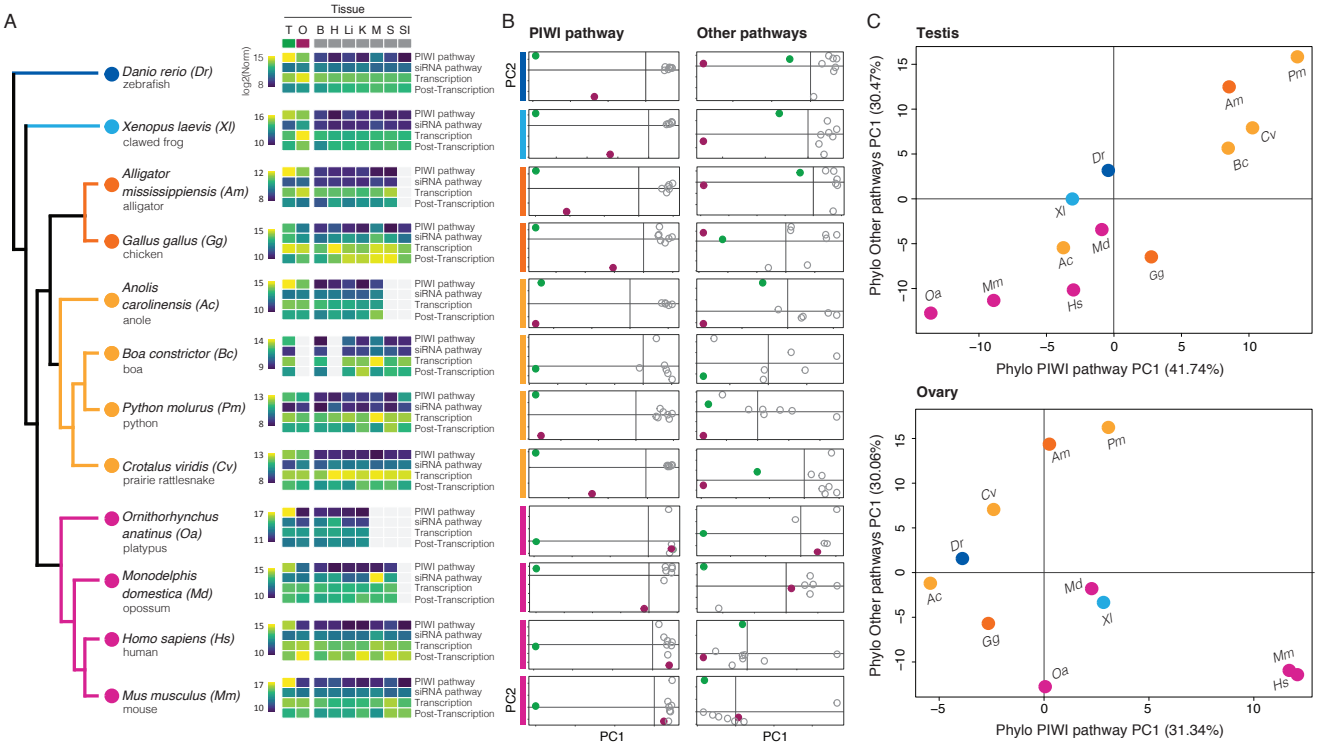


Figure 1. – Expression patterns of key genes involved in negative regulation of transposable element (TE) activity in germline and somatic vertebrate tissues. (A) For each species, heatmaps show log2-transformed within-species normalized expression levels of main pathways involved in TE silencing. Warm colors (yellow) represent higher total expression levels of genes in the pathway across tissues. (B) Principal component analyses (PCA) for the PIWI:piRNA pathway (left) and all other regulatory pathways (siRNA pathway, transcriptional and post-transcriptional TE silencing mechanisms; right) reflect variance in gene expression profiles across tissues for each species. While non-mammal species show discrimination of both germline tissues (testis in green and ovary in maroon) form somatic tissues (empty grey circles) and from each other in respect to PIWI pathway genes, gene expression in the mammalian ovary falls within the variability of somatic tissues. (C) PCA for the testis (above) and ovary (below) show species clustering based on the principal component of the PIWI pathway (x axis) and all other regulatory pathways (y axis). Per each species, coordinates were extracted from the corresponding phylogenetically independent contrast (PIC) PCAs. Cold colors represent non-amniote vertebrates, warm colors reptiles, and magenta mammal species.

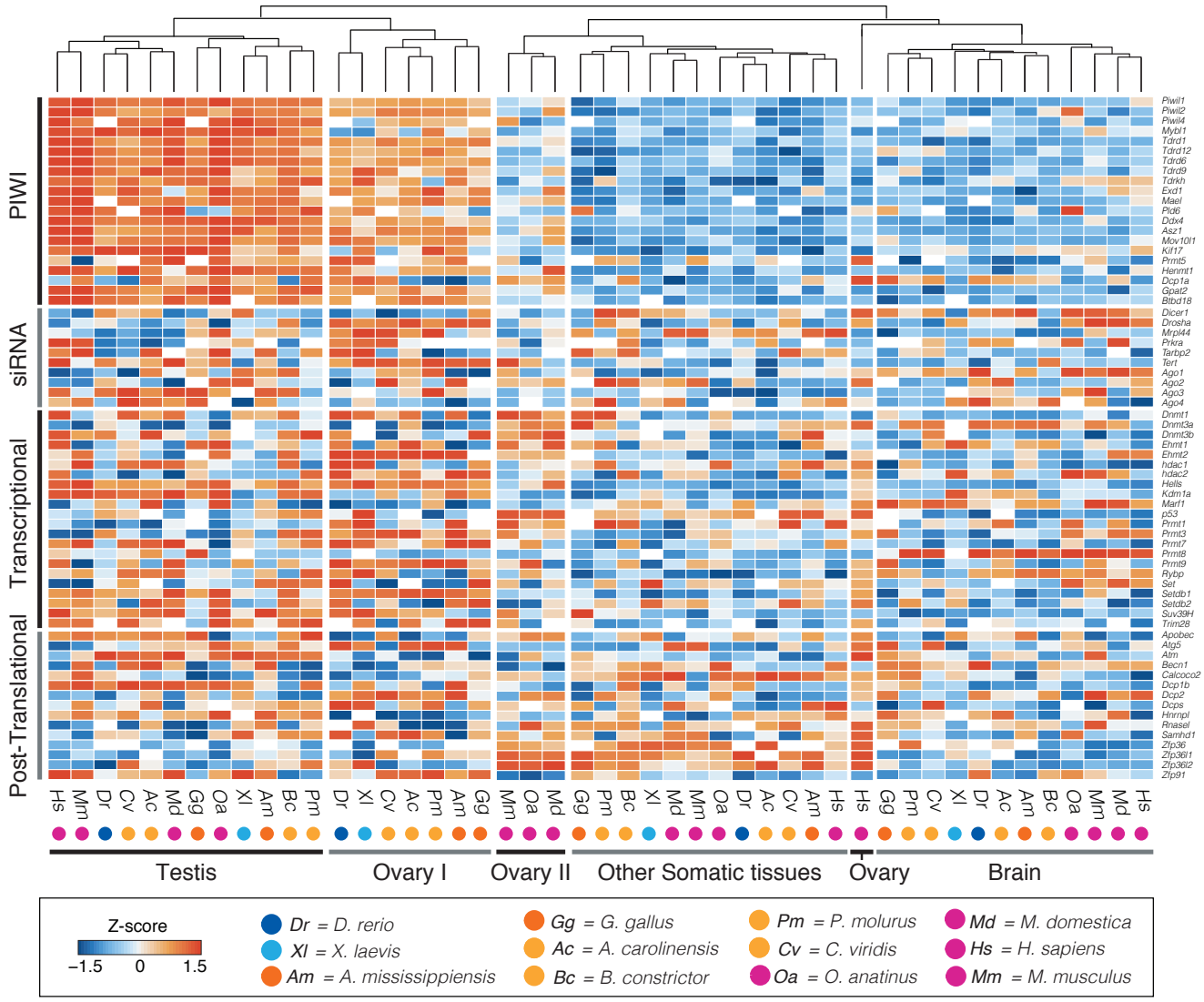


Figure 2. – Hierarchical clustering Z-score heatmap of TE regulatory genes in germline and somatic vertebrate tissues. Analysis of differential expression of key conserved genes involved in TE silencing suggests the existence of 5 main expression profiles across vertebrate tissues: vertebrate testis, characterized by the highest activation status of the PIWI:piRNA pathway and transcriptional regulators; ovary of non-mammal species, with expression patterns similar to the testis; mammalian ovary (to the exclusion of humans), which shows a sharp decreased expression of PIWI genes; other somatic tissues (average Z-scores across heart, kidney, liver, muscle, spleen and small intestine after individual tissue heatmap supported the existence of a single cluster); and vertebrate brain.

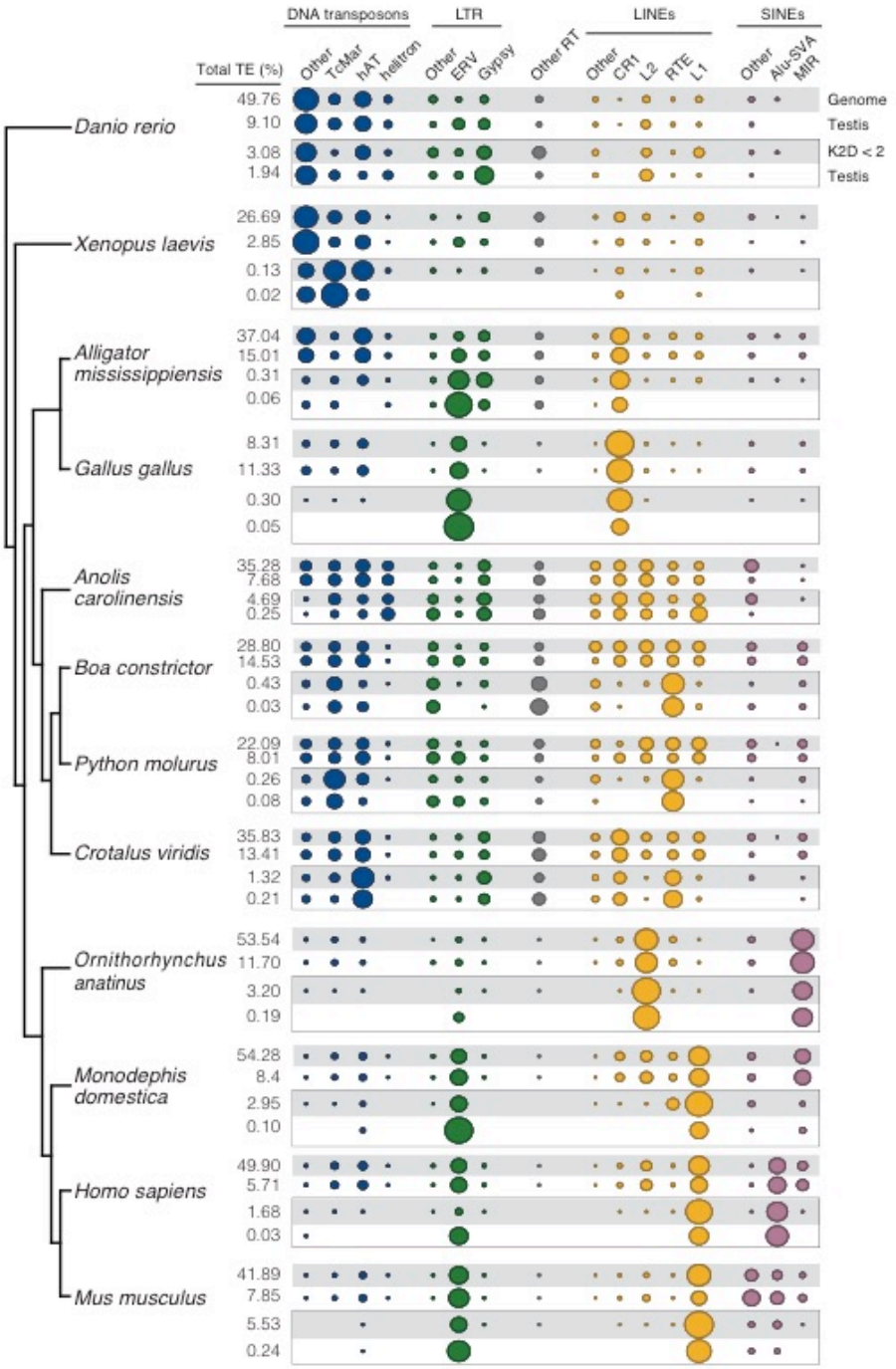


Figure 3. – Relationship between genomic and transcriptomic TE relative composition in the male germline. Area of the circles in the balloon plot reflects the percentage of major TE subfamilies (blue = DNA transposons, green = LTRs, grey = PLE and DIRS, yellow = LINEs, violet = SINEs) relative to the total genomic TE content (top row, grey background) and to the total TE transcriptome (white background). In the box, the same relationship is displayed for recently inserted TE copies (with a Kimura distance < 2%) and recent-TEs in the transcriptome. Values to the left report the real proportion of TEs (TE content %) in the genomes and transcriptomes. We find support for high TE transcription in testis transcriptomes (up to 15%), which perfectly match the relative composition of the genome. In contrast, for recent-TEs some families are entirely missing in the transcriptome despite their presence in the genomic background. Balloon plot additionally highlights variability in TE landscapes across vertebrates.

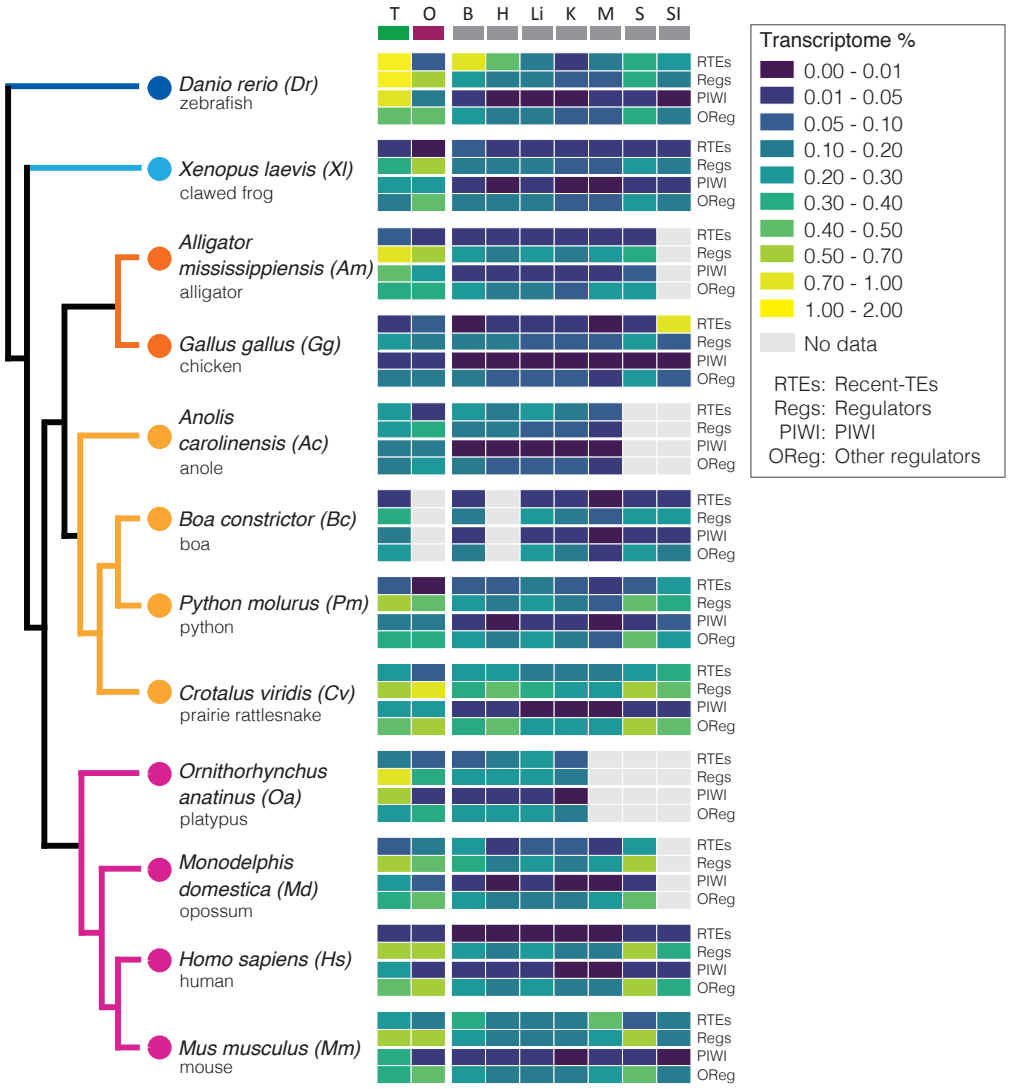


Figure 4. – Expression levels of recent-TEs and their negative regulatory mechanisms in vertebrate somatic and germline tissues. Heatmap shows comparative expression levels of recent-TEs (top row), total regulatory pathways (PIWI:piRNA, siRNA, transcriptional and post-transcriptional), and details of the contribution of PIWI:piRNA pathway and all remaining silencing mechanism (bottom section) across vertebrate tissues. Comparative gene expression is reported as percentage of the transcriptome following within species normalization. Whereas human, xenopus and chicken show the lowest levels of recent-TE expression in both germline and somatic tissues, vertebrate tissues show moderate to high contribution of TEs to tissue transcriptomes, which is consistently highest in the testis, and reduced in non-mammal ovary.

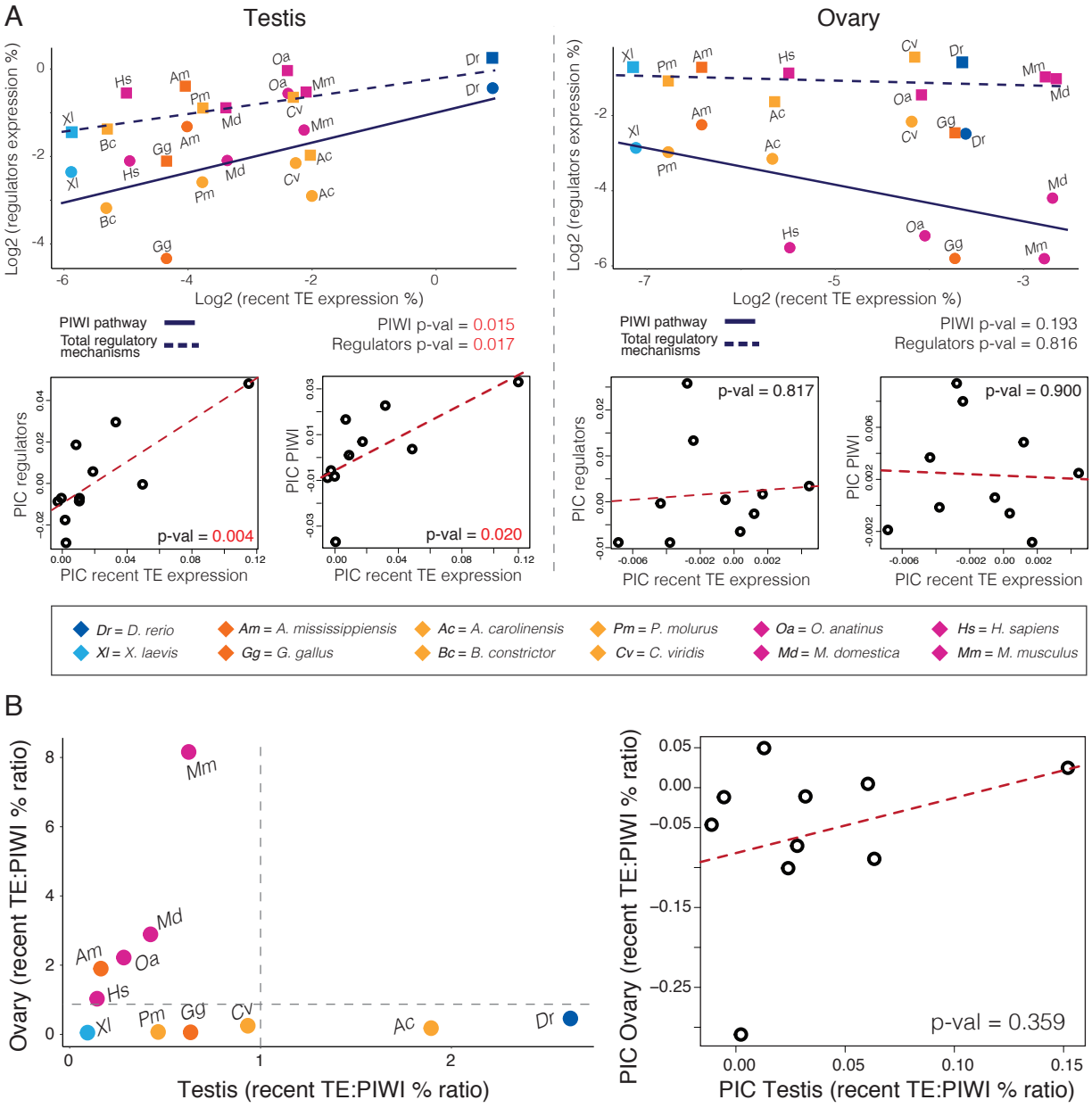


Figure 5. – Relationship between expression levels of recent-TEs and their negative regulatory pathways. (A) Linear regressions and PICs support a significant positive relationship between recent-TE expression and host response (PIWI pathway and total response) in the testis, whereas in the ovary they suggest the opposite, although not significant, trend. (B) Patterns of species TE expression levels in the testis (x-axis) and ovary (y-axis). Recent-TE transcriptome percentages were corrected by the PIWI pathway to test for a correlation in expression levels. Mammal species show a consistent trend in the ovary where lower regulatory activity brings to increased TE transcription, matched by the testis although in favor of the PIWI pathway, compared to non-mammal species. In contrast, non-mammal species show a consistent host response proportional to TE activity in the ovary (constant TE:PIWI ratio), but higher variability in the testis, with some species that are more efficient at contrasting TEs.

Supplementary Methods

In this study we used previously published poly-A-selected RNAseq datasets that are available on the NCBI SRA archive database (Leinonen et al. 2011); supplementary table 1). Raw RNAseq data were first filtered for prokaryote and eukaryote rRNA transcripts using SortMeRNA v2.1 (Kopylova et al. 2012), and then quality and adapter trimmed in Trimmomatic 0.36 (Bolger et al. 2014). We used FastQC v0.11.8 (Andrews 2010) to assess quality of the processed reads. Since quality assessment passed analyses for all parameters without warnings, no further read filtering was performed. Forward and reverse paired reads from Trimmomatic filtering were then used as input for estimation of gene and transposable element (TE) expression levels.

For each species, we used *STAR* v2.7.0f (Dobin et al. 2013) to map reads to the latest genome version and annotation *.gff* files available on the NCBI Genome database (Sayers et al. 2019) at the time of analyses (*Danio rerio*: assembly GRCz11, *Xenopus laevis*: assembly Xenopus_laevis_v2, *Alligator mississippiensis*: assembly ASM28112v4, *Gallus gallus*: assembly GRCg5, *Anolis carolinensis*: assembly AnoCar2.0, Boa constrictor: ERS218597, snake_7C available on GigaDB (Bradnam et al. 2013; Card et al. 2019), *Python molurus*: assembly Python_molurus_bivittatus-5.0.2, *Crotalus viridis*: assembly UTA_CroVir_3.0, *Ornithorhynchus anatinus*: mOrnAna1.p.v1, *Monodelphis domestica*: assembly MonDom5, *Homo sapiens*: assembly GRCh38.p12, *Mus musculus*: assembly GRCm38.p6). For human and mouse, primary genome assemblies and corresponding annotation files were used to avoid incorrect read alignment to loci on patches or alternative haplotype contigs as suggested by the authors (Dobin et al. 2013). STAR was run according to default parameters but using the variables -

winAnchorMultimapNmax 100 and *-outFilterMultimapNmax 100*, which allow for multiple alignments of maximum 100 reads, as specified in (Jin et al. 2015).

Expression levels were estimated using *TEtranscript* (Jin et al. 2015), a tool that allows for the simultaneous estimation of gene and TE expression levels. To assign mapped reads to a genomic locus, *TEtranscript* leverages two annotation files that specify gene and repeat element coordinates, respectively. We used the same gene annotation files provided as a reference when running *STAR* to build the gene index structure; to convert *.gff* files into the required *.gtf* files we used the *gff3ToGenePred* and *genePredToGtf* modules available on the UCSC website (Casper et al. 2018). TE index structures were built from RepeatMasker (Smit et al. 2015-2019) runs; for all species, we used the standard tetrapoda library as reference, except for human and mouse for which we used the mammal library. For squamate species we used the same strategy described in (Schield et al. 2019) instead. We used the script *makeTEgtf.pl* available made available from the Hammell lab (http://labshare.cshl.edu/shares/mhammelllab/www-data/TEToolkit/TE_GTF/makeTEgtf.pl.gz) to convert RepeatMasker *.out* files into *.gtf* files. *TEtranscript* was run using default parameters, using the *--multi* flag and specifying whether transcriptome data was stranded or not. To further focus our analyses on TE-derived reads that originated from recently active (and thus, potentially capable of transposition) TE copies, in a second, separate analysis we provided *TEtranscript* with a *.gtf* annotation file containing only TE loci that according to the RepeatMasker *.out* file had less than 2% Kimura 2-parameter distance from the consensus (we refer to this as the "recent-TE" dataset).

For each species, raw read counts for the entire transcriptome (genes and TEs) were normalized across tissues in DESeq2 v1.20 (Love et al. 2014) after removing elements with less than 10

mapped reads across samples. Normalization was run independently on the total-TE dataset and on the recent-TE dataset; since gene expression estimates were the same in the two datasets, we consistently used the normalized counts coming from the recent-TE dataset when analyzing gene expression levels. To examine patterns of within-species variation in expression profiles across tissues, we first applied a blind variance stabilizing transformation (Anders and Huber 2010) to the entire count matrix, and used the resulting transformed data to calculate tissue-wise variance and evaluate patterns of expression using principal component analyses (PCAs).

To assess the relationships between expression levels of TEs and genes involved in TE negative regulatory mechanisms, we compared recent-TE expression levels to 5 sets of TE regulators: (i) genes participating in the PIWI:piRNA pathway (Carbon et al. 2009; PIWI pathway hereafter); (ii) genes involved in the small RNA interference pathway (Carbon et al. 2009; siRNA pathway); (iii) genes involved in transcriptional regulation of TEs (e.g., responsible for *de novo* DNA or histone methylation; e.g., Hutchins and Pei 2015; Wylie et al. 2016); (iv) other genes previously identified to negatively impact TE mobilization and/or insertion at the post-transcriptional level (e.g., Apobec; Goodier 2016); and (v) the combined magnitude of the host response against TEs (all genes involved in negative TE regulation). The final gene dataset included a total of 77 conserved genes, for which we recovered expression values of all annotated orthologues (Supplementary table 2).

Because of the heterogeneous nature of our data, we chose to perform between species comparisons using percentages of the total transcriptome following normalization of read counts to limit biases due to different methods of tissue processing, library preparation, sequencing technology and dataset quality (Sudmant et al. 2015; Dunn et al. 2018). To compare differences

in patterns of gene expression levels across species and tissues, we calculated Z-score values of gene expression for each species using the *scale* function on log2 transformed normalized data in R (R Core Team). Z-scores were also used to perform hierarchical tissue clustering across species as part of the heatmap data visualization step generated using the *pheatmap* R package (Kolde 2012) (distance method="euclidean"; hclust "complete" clustering). To assess differential gene expression in the testis and in the ovary compared to somatic tissues, we used log2 fold change values and corresponding adjusted p-values as calculated in DESeq2. Finally, to investigate relationships in TE and gene expression patters across vertebrates, we performed phylogenetic independent contrast (PIC) linear regressions and PCAs using the *phytools* package in R (Revell 2012). Spearman rank correlation analyses were performed using the *rcorr* function in the Hmisc v4.2-0 R package (Harrell 2019).

SUPPLEMENTARY FIGURES

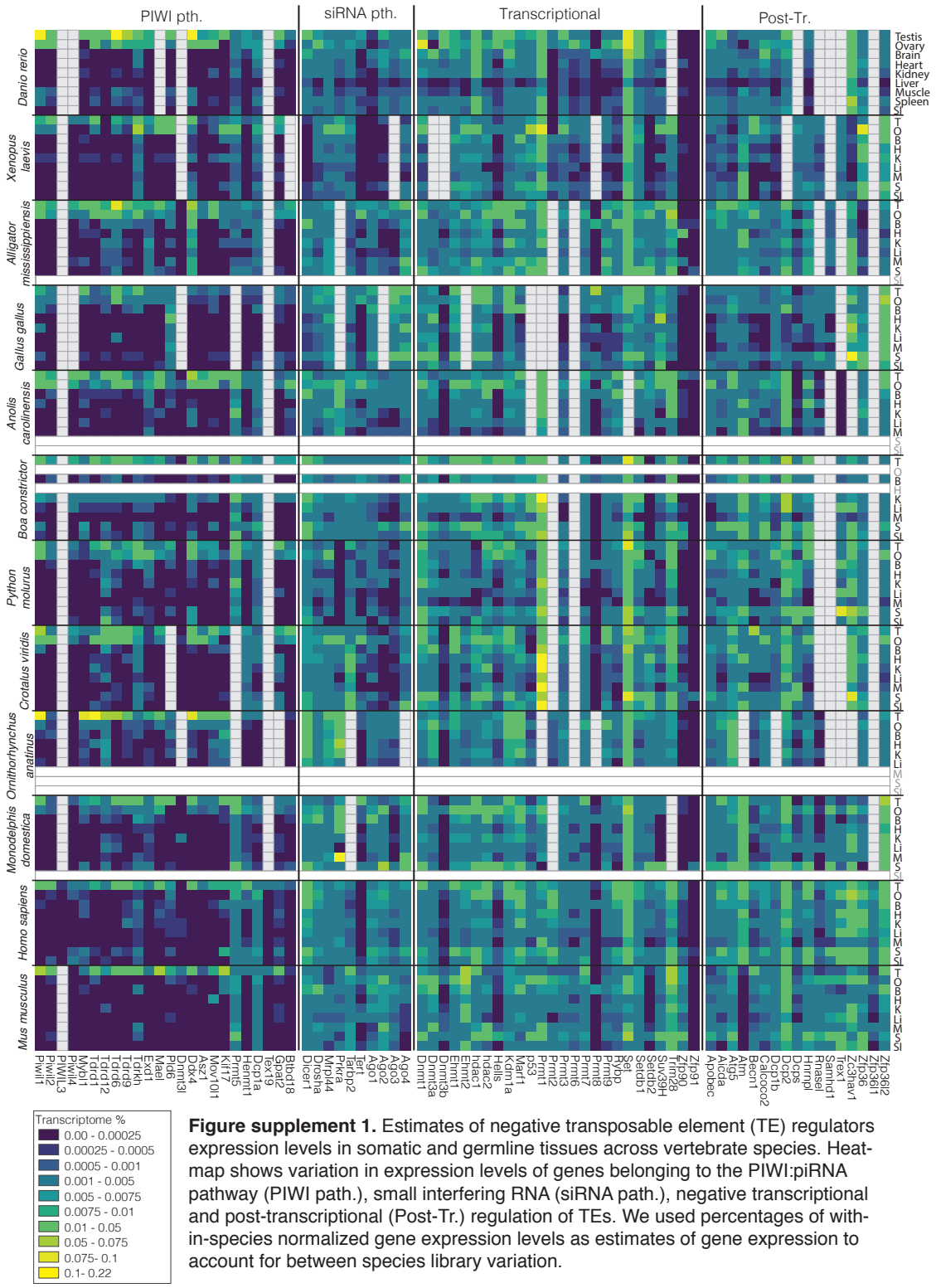


Figure supplement 1. Estimates of negative transposable element (TE) regulators expression levels in somatic and germline tissues across vertebrate species. Heat-map shows variation in expression levels of genes belonging to the PIWI:piRNA pathway (PIWI path.), small interfering RNA (siRNA path.), negative transcriptional and post-transcriptional (Post-Tr.) regulation of TEs. We used percentages of within-species normalized gene expression levels as estimates of gene expression to account for between species library variation.

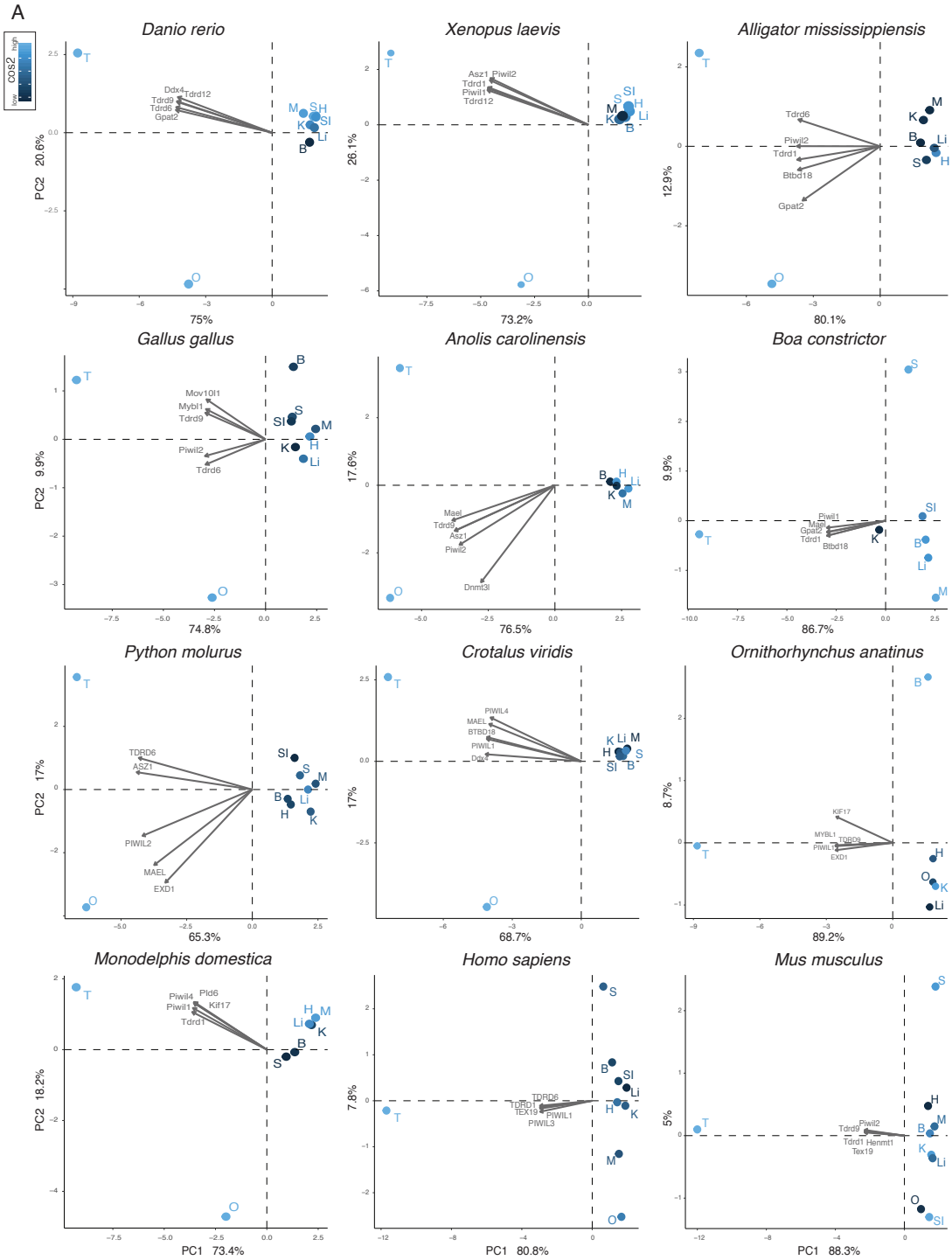


Figure supplement 2A. Principal Component Analyses (PCA) for genes taking part in the PIWI-piRNA pathway of negative TE regulation across tissues in vertebrate species. PCA plots show tissue clustering based on variance stabilized transformation (vst) of gene expression estimates assessed in DeSeq2. Tissues are colored according to their contributions (cos2 = quality of the sample on the factor map). Arrows represent the eigen vectors for the 5 most contributing variables in the variables factor map. Noticeable is the discrimination of both germline tissues in non-mammal species, whereas only the testis is discriminated in mammals.

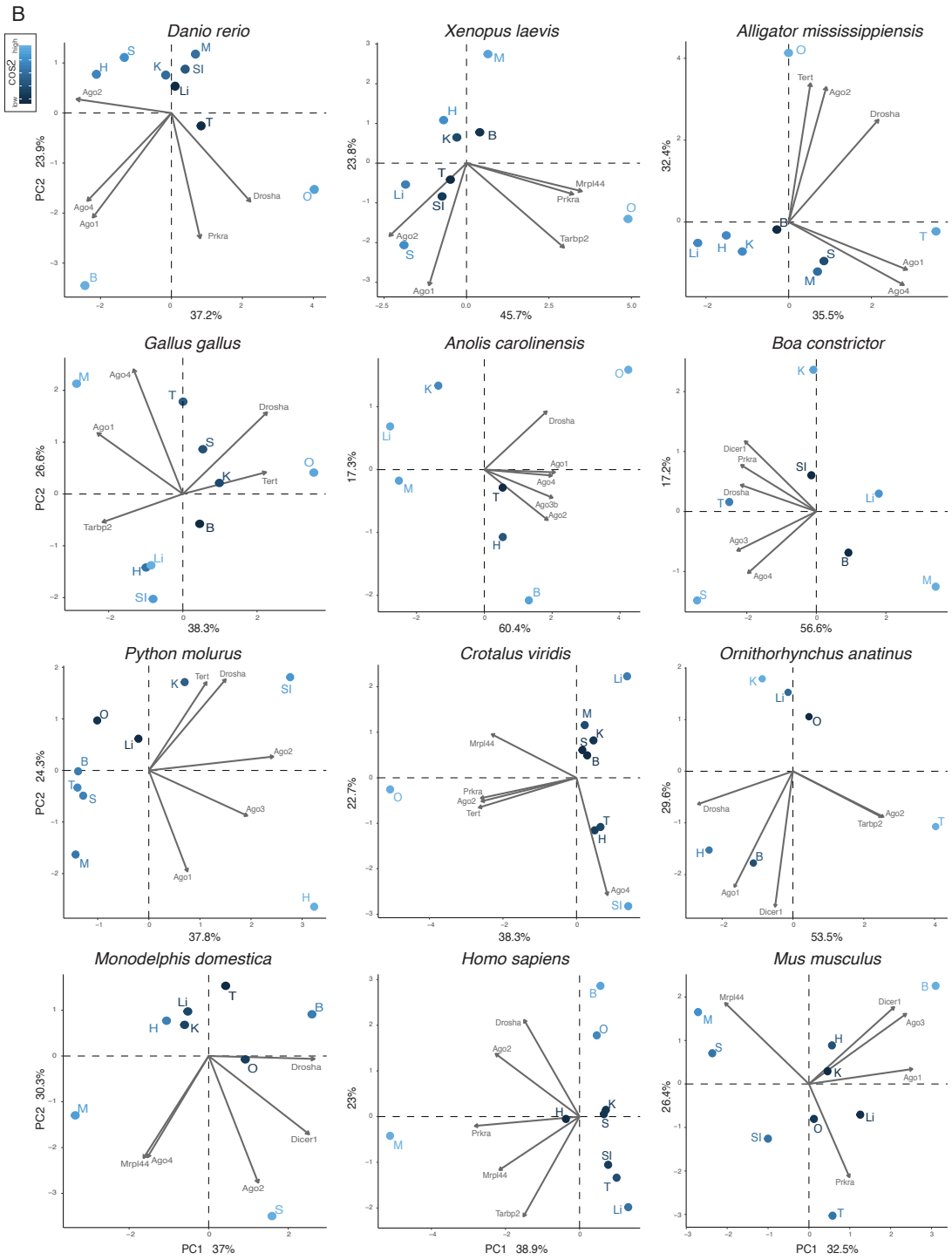


Figure supplement 2B. Principal Component Analyses (PCA) for genes taking part in the small interfering RNA (siRNA) pathway across tissues in vertebrate species. PCA plots show tissue clustering based on variance stabilized transformation (vst) of gene expression estimates assessed in DeSeq2. Tissues are colored according to their contributions (cos2 = quality of the sample on the factor map). Arrows represent the eigen vectors for the 5 most contributing variables in the variables factor map. In contrast to PIWI pathway gene PCAs, variance in gene expression levels between somatic tissues is greater than between germline and somatic tissues.

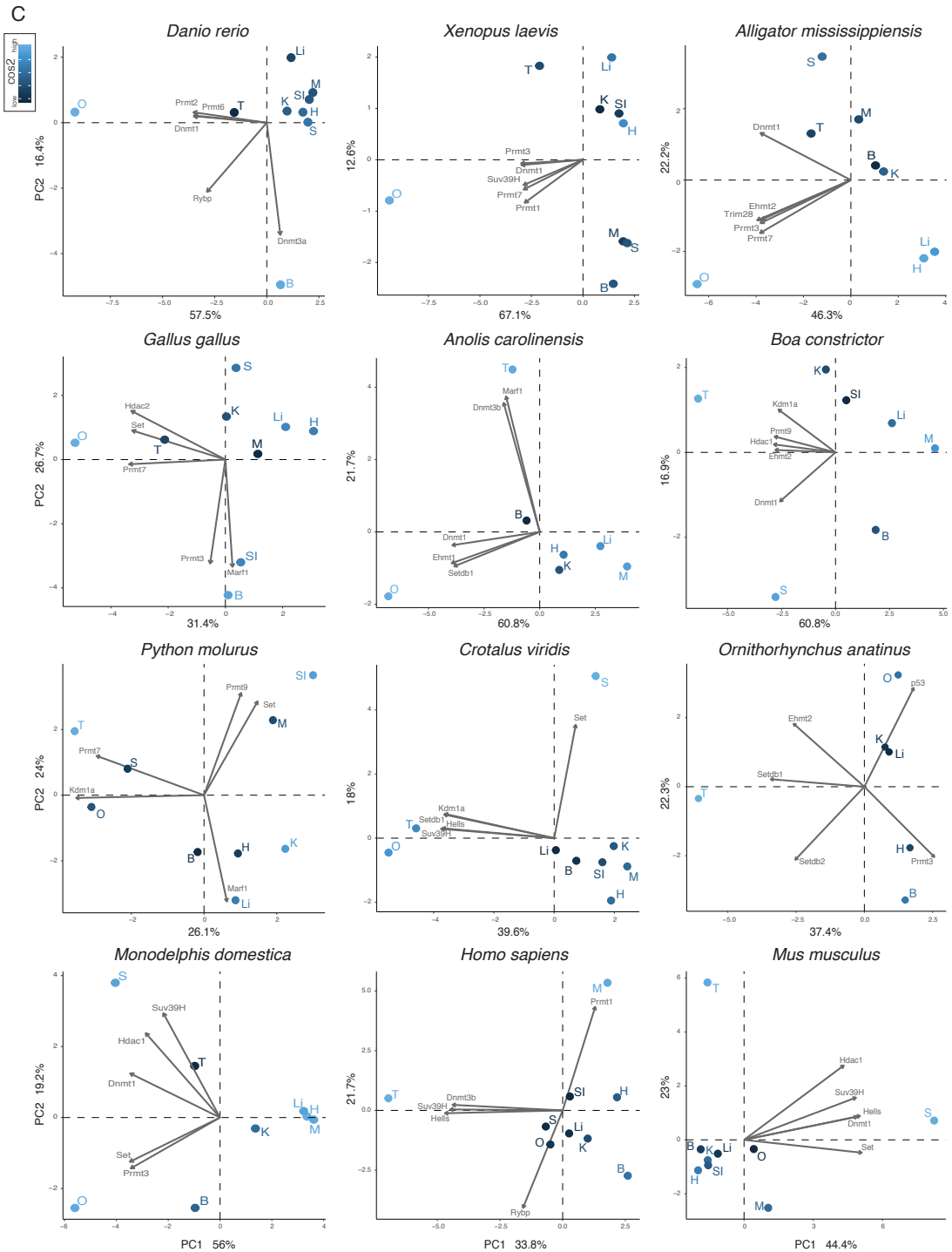


Figure supplement 2C. Principal Component Analyses (PCA) for genes taking part in negative transcriptional regulation across tissues in vertebrate species. PCA plots show tissue clustering based on variance stabilized transformation (vst) of gene expression estimates assessed in DeSeq2. Tissues are colored according to their contributions (cos2 = quality of the sample on the factor map). Arrows represent the eigen vectors for the 5 most contributing variables in the variables factor map. Whereas variance in gene expression levels between somatic tissues varies between species, the ovary is characterized by the highest variance in most vertebrates to the exception of eutheria mammals.

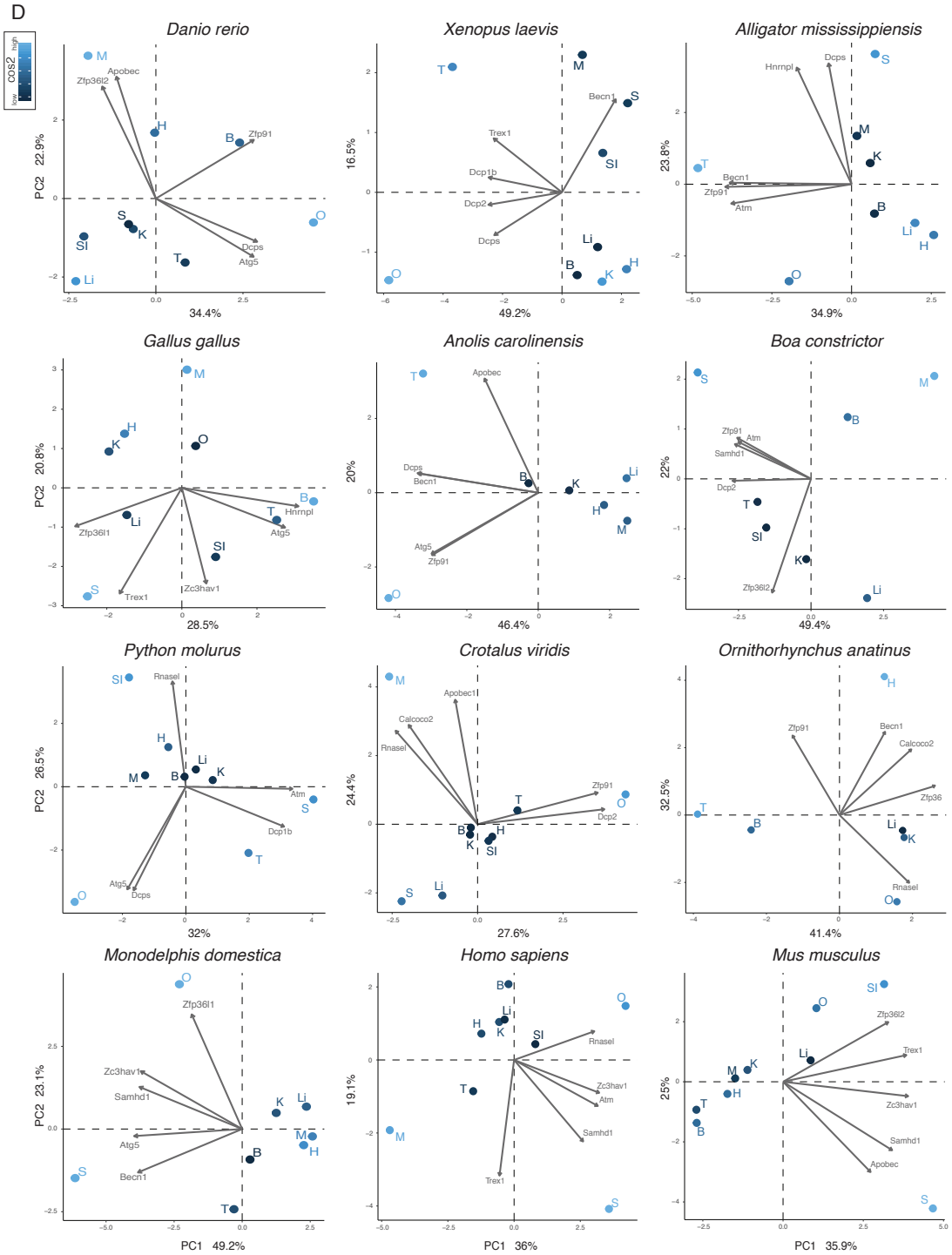


Figure supplement 2D. Principal Component Analyses (PCA) for genes taking part in negative post-transcriptional regulation across tissues in vertebrate species. PCA plots show tissue clustering based on variance stabilized transformation (vst) of gene expression estimates assessed in DeSeq2. Tissues are colored according to their contributions (cos2 = quality of the sample on the factor map). Arrows represent the eigen vectors for the 5 most contributing variables in the variables factor map.

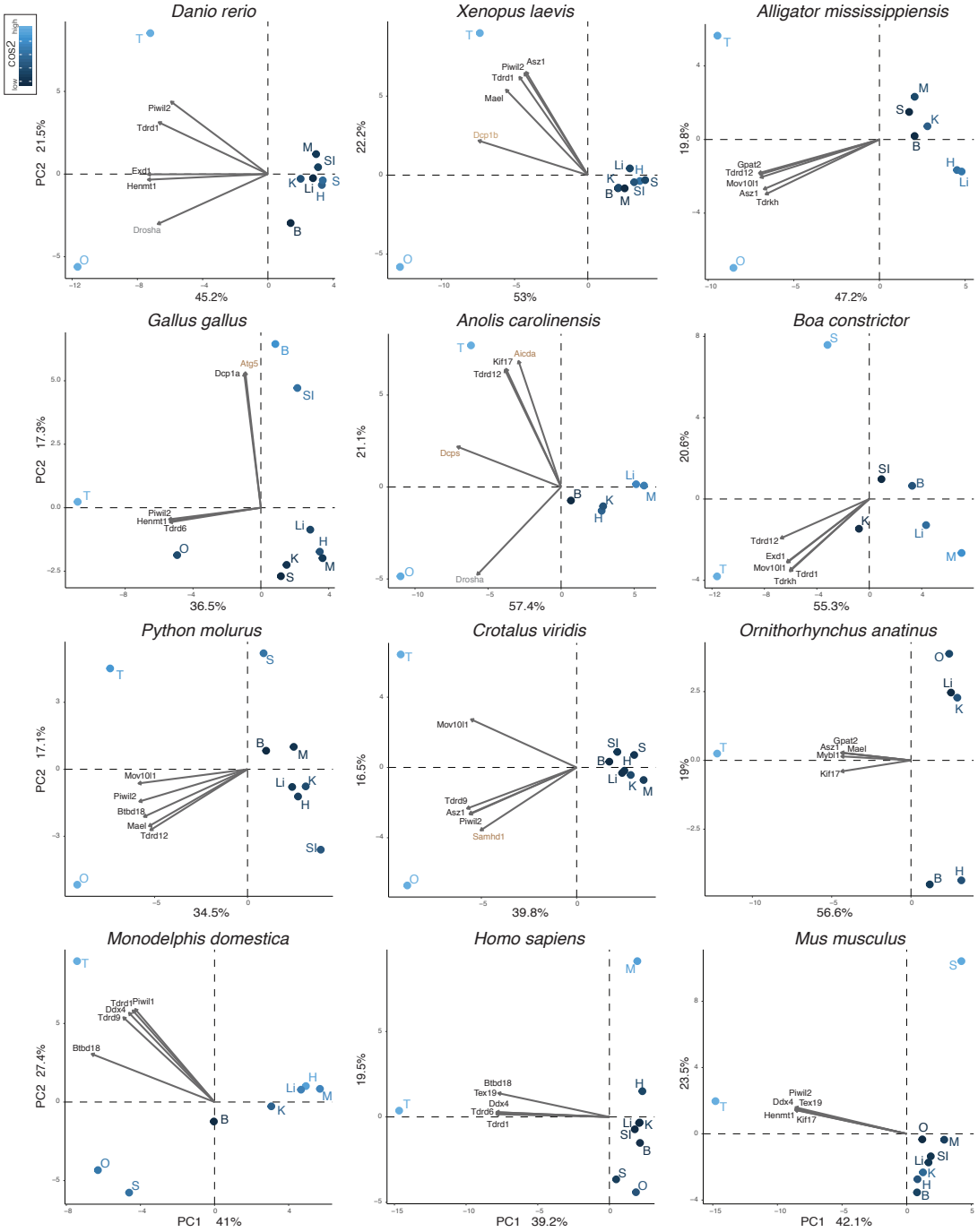


Figure supplement 3. Principal Component Analyses (PCA) for genes taking part in negative TE regulation across tissues in vertebrate species. PCA plots show tissue clustering based on variance stabilized transformation (vst) of gene expression estimates assessed in DeSeq2. Tissues are colored according to their contributions (cos2 = quality of the sample on the factor map). Arrows represent the eigen vectors for the 5 most contributing variables in the variables factor map. The evolutionary transition to mammals is mirrored by a dicrease in variance and contribution of the ovary compared to other somatic tissues, whereas in non mammal species both germline tissues cluster independently with high support. Genes belonging to the PIWI:piRNA pathway (black) appear to be major factors driving tissue separation relatively to genes bleonging to the siRNA pathway (gray), trnascriptional (brown), and post-transcriptional mechanisms (light brown).

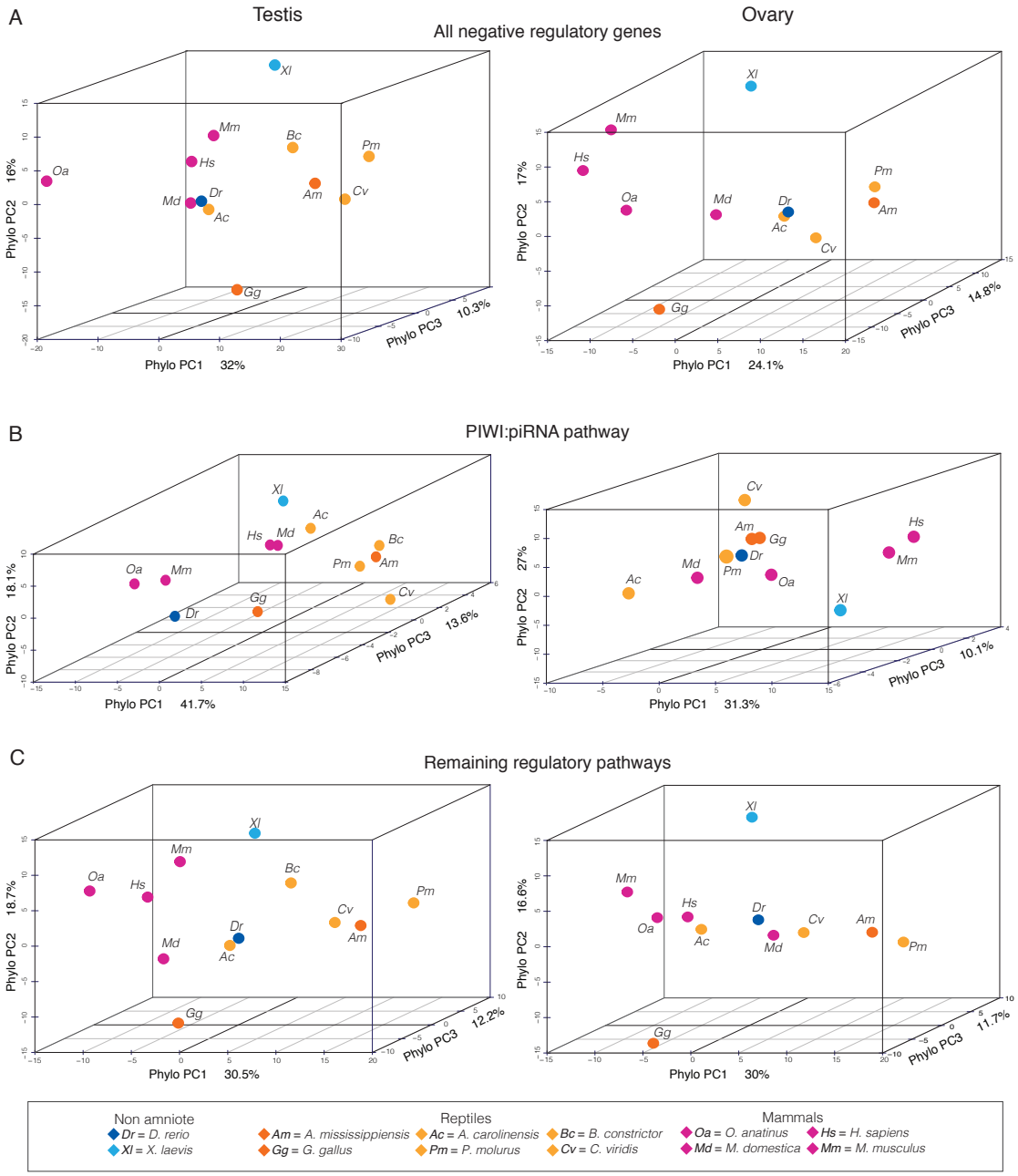


Figure supplement 4. Phylogenetic PCAs for expression levels of genes involved in negative regulation of TEs in germline tissues. A-C) Phylogenetic PCAs show clustering of vertebrate species according to variance in vst (variance-stabilized transormation) gene expression levels in the testis (left) and in the ovary (right). A) When all genes are considered, mammals (eutheria in particular) form a distinct cluster for the ovary, whereas clustering structure is less pronounced in the testis. A similar trend is observed for the genes involved in the PIWI:piRNA pathway (B), but exclusively for eutheria mammals in the ovary. C) Remaining regulatory pathways include genes that take part in the siRNA pathway and in transcriptional and post-transcriptional regulation of TEs. Clustering of mammals is present iin both germline tissues, but more evident in the male than in the female germline. Overall, squamates tend to cluster together in all PCAs, except for the chicken that, like the xenopus, tends to cluster individually in A and C. Results support our hypothesis of differential regulation of TE expression in germline tissues between mammals and other vertebrate species.

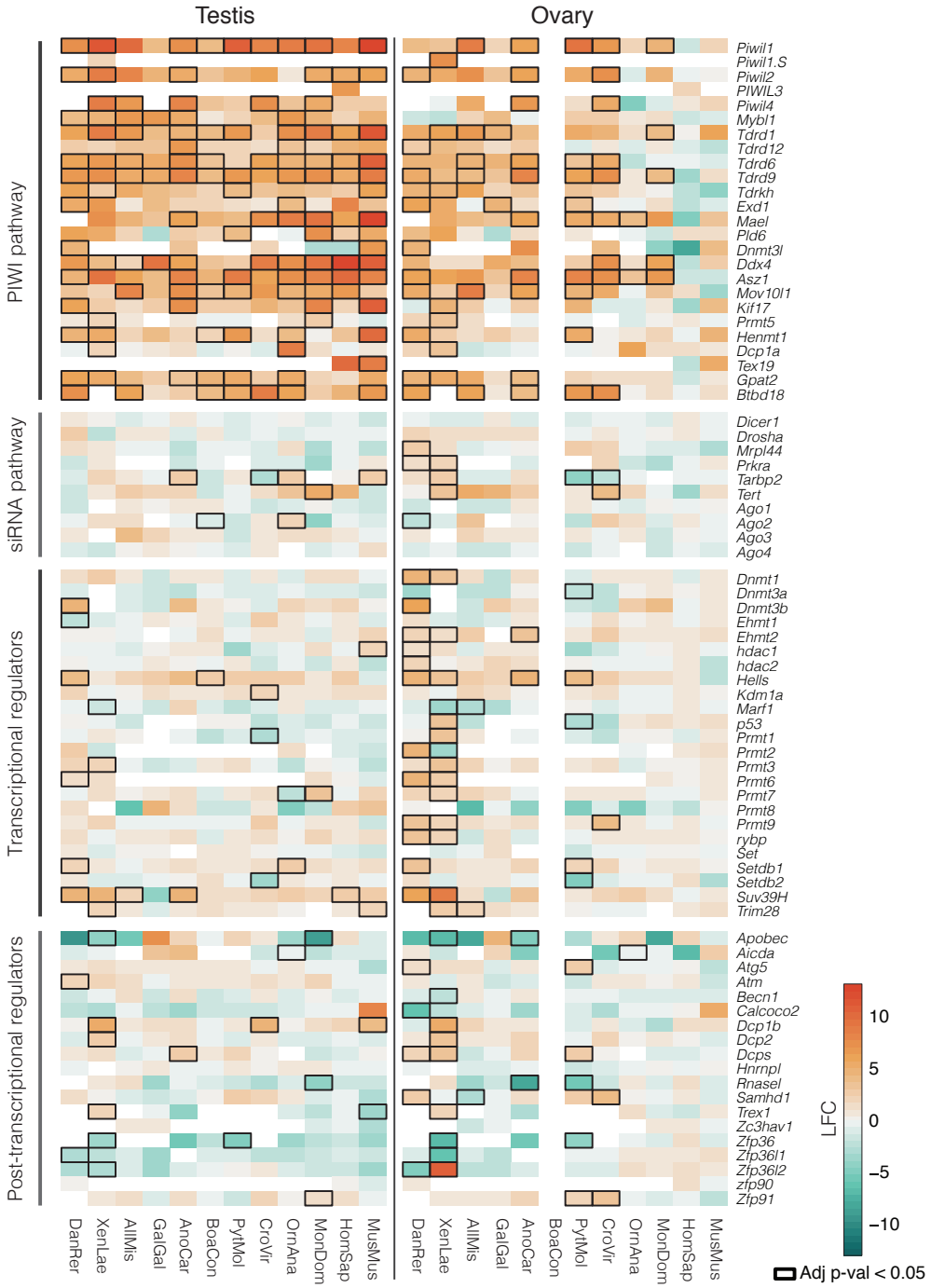


Figure supplement 5. Germline differential expression of genes involved in negative TE regulative pathways. Heatmap shows log fold changes (LFC) in expression levels of main negative TE regulators in germline tissues (testis on the left and ovary on the right) compared to their somatic expression. Warmer colors represent enrichment in the germline, and colder colors represent lower expression. Significant differential expression (DE) is reported as black cell outlines. DE analyses were performed in DEseq2. In agreement with the literature, we detected a significant enrichment for genes involved in the PIWI:piRNA pathway in the male germline across all vertebrate species, but mainly across non mammal species in the female germline.

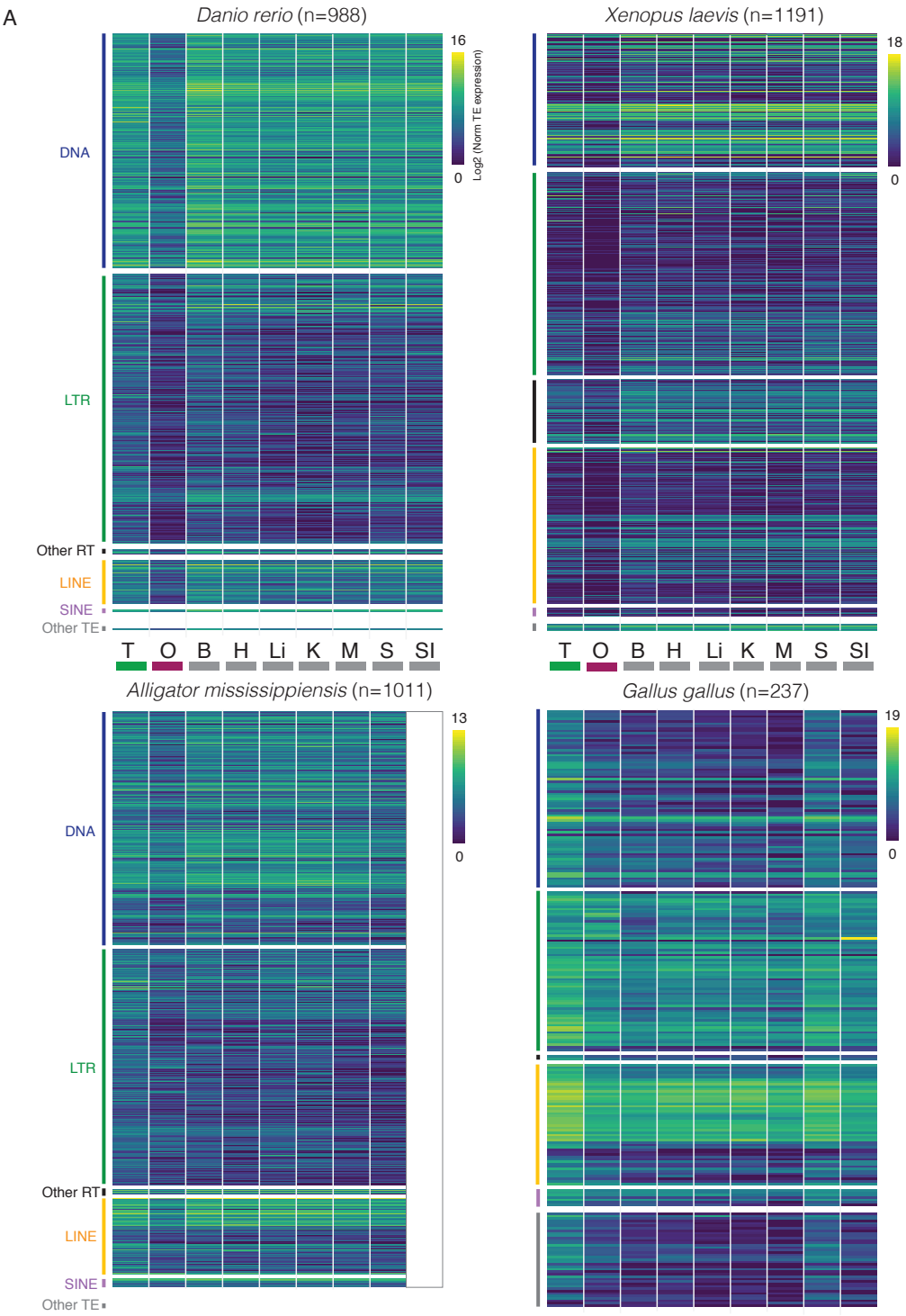


Figure supplement 6A. Total-TE derived transcript expression across tissues. Heatmaps show individual TE expression levels across tissues for lower vertebrates and archosauria reptiles. Empty columns represent tissues with unavailable transcriptome data. Hearmaps also reflect the relative abundance in terms of number of individual elements belonging to each of the major TE classes. T = testis; O = ovary; B = brain; H = heart; Li = liver; K = kidney; M = muscle; S = spleen; SI = small intestine. LTR = Long Terminal Repeats; Other RT = other retrotransposons (PLE and DIRS); LINE = Long Interspersed Nuclear Elements; SINE = Short Interspersed Nuclear Elements. To the exception of the chicken, TE expression levels are consistent across tissues for the other vertbrate species.

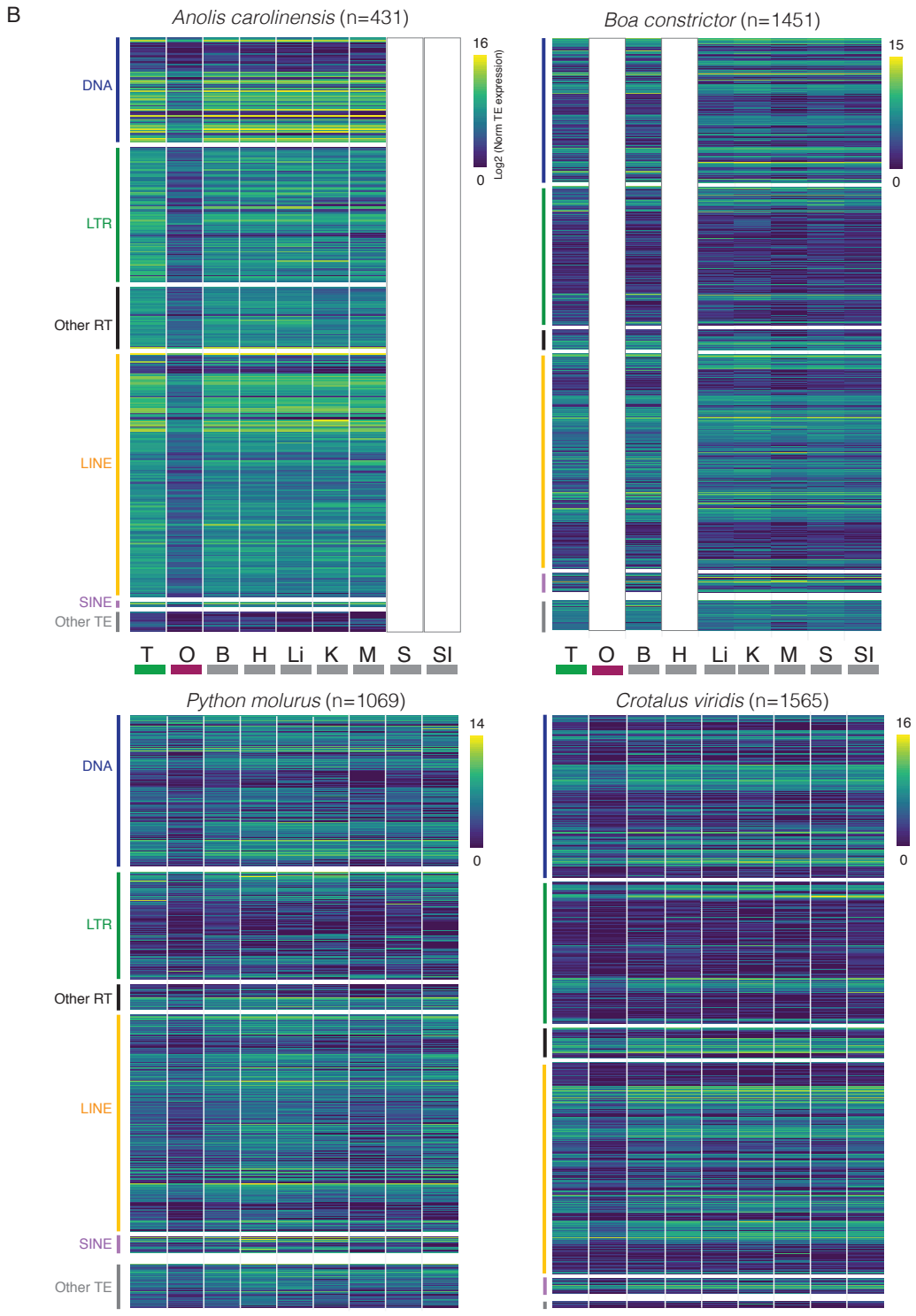


Figure supplement 6B. Total-TE derived transcript expression across tissues. Heatmaps show little variability in individual TE expression levels across tissues in squamate reptile species, although some tissues like the ovary in the green anole lizard (*Anolis carolinensis*) and in the prairie rattlesnake (*Crotalus viridis*), and the muscle in *Boa constrictor* are characterized by overall lower expression levels.

Figure supplement 6C. Total-TE derived transcript expression across tissues. Heatmaps show that individual TE expression levels across tissues are more variable in mammal species compared to other vertebrates, and highlight how pervasive transcription of TEs represents a common feature of healthy tissue transcriptomes.

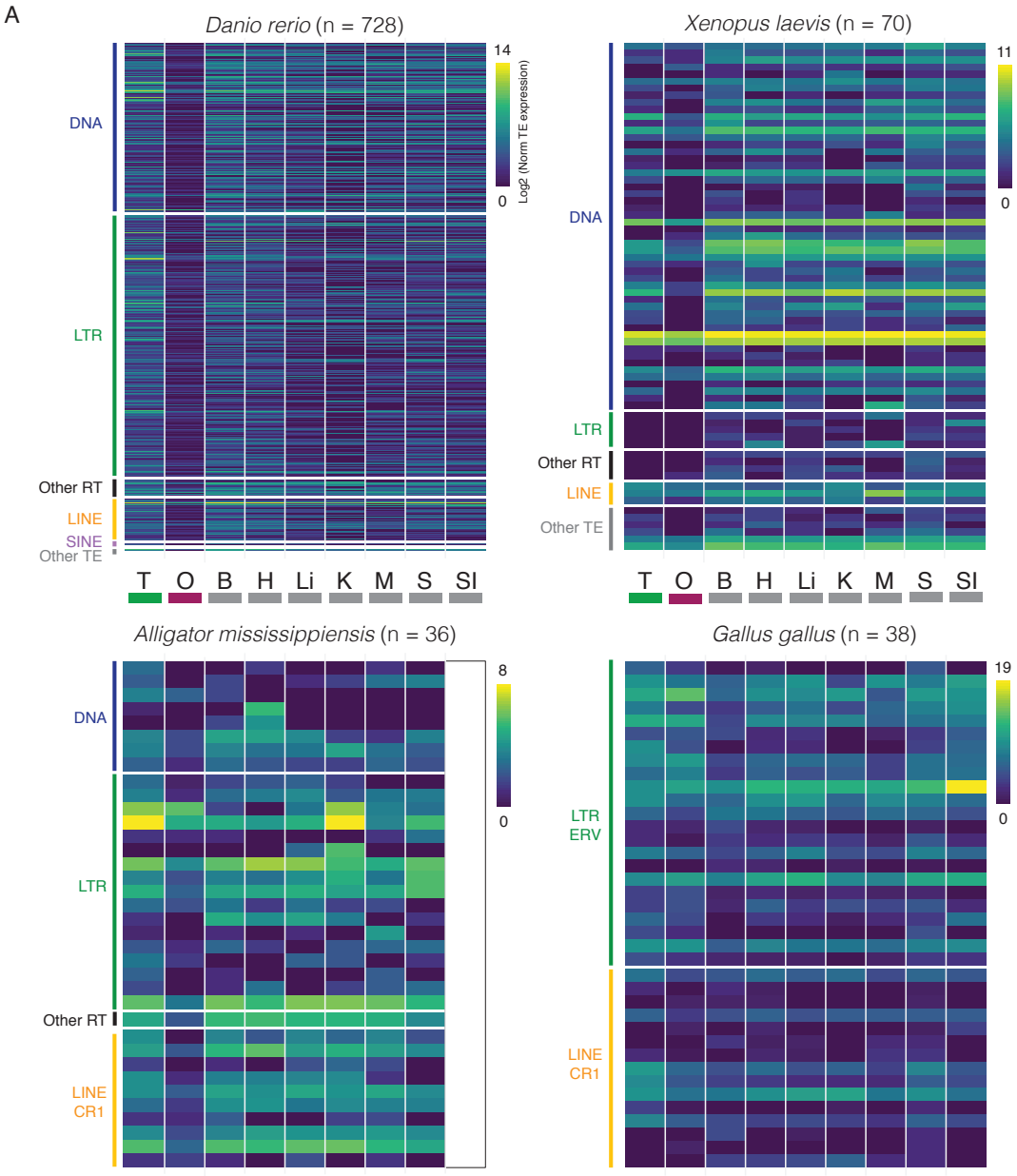


Figure supplement 7A. Potentially active TE-derived transcript expression across tissues. Heatmaps show individual TE expression levels across tissues for lower vertebrates and archosauria reptiles. Empty columns represent tissues with unavailable transcriptome data. Hearmaps also reflect the relative abundance in terms of number of individual elements belonging to each of the major TE classes. T = testis; O = ovary; B = brain; H = heart; Li = liver; K = kidney; M = muscle; S = spleen; SI = small intestine. LTR = Long Terminal Repeats; Other RT = other retrotransposons (PLE and DIRS); LINE = Long Interspersed Nuclear Elements; SINE = Short Interspersed Nuclear Elements. Across vertebrates, the zebrafish (Danio rerio) is one of the few species to have a remarkably high incidence of potentially active TEs (n= 728; generated from genomic loci with a Kimura 2D distance from the consensus less than 2%) that include all major TE families, supporting high genomic turnover of TEs in the species. On the other hand, most vertebrate species are characgterized by a very small number of TEs capable of generating transcripts.

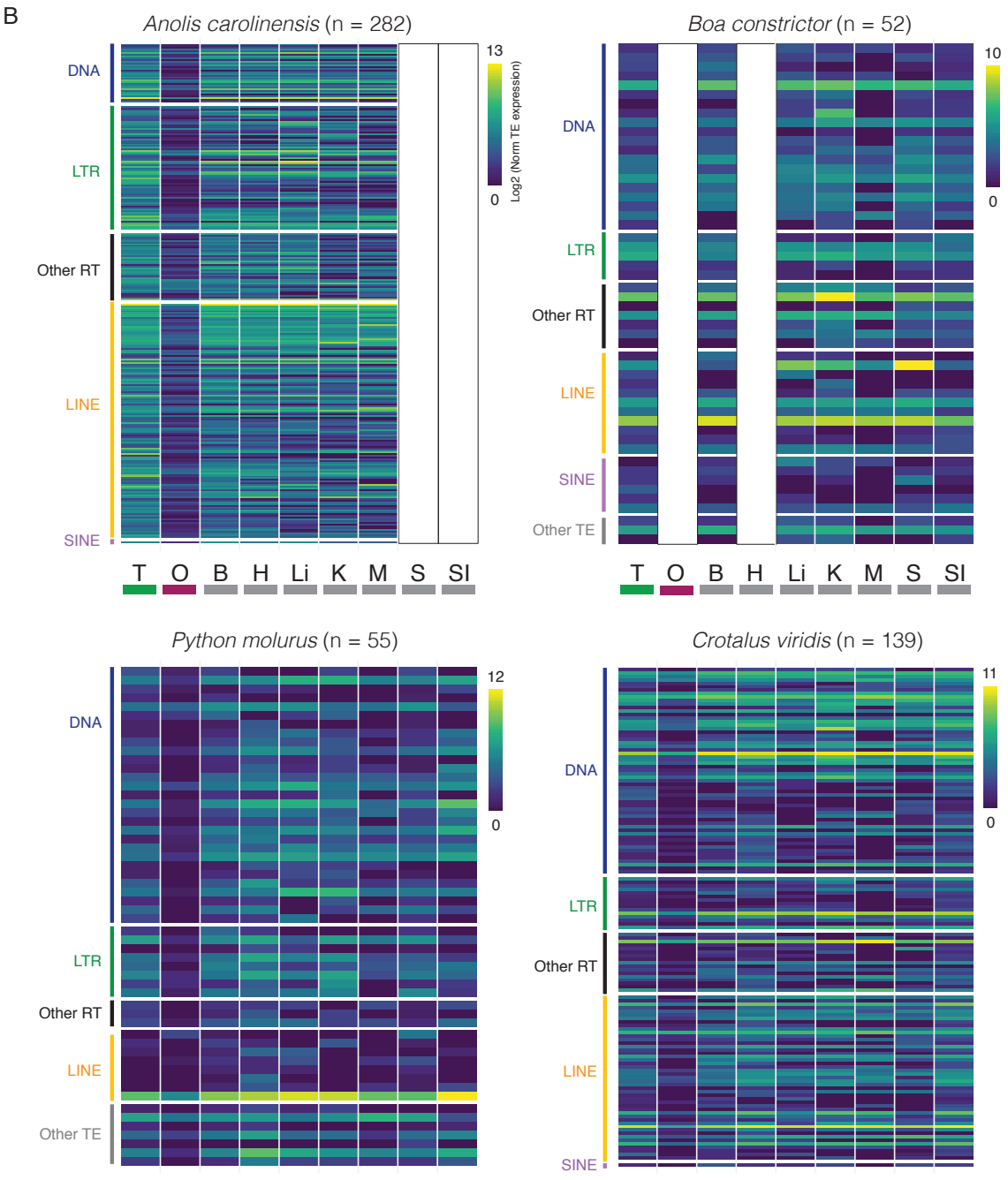


Figure supplement 7B. Recent-TE derived transcript expression across tissues in squamate reptiles. Heatmaps show individual TE expression levels across tissues of potentially active TEs that recently amplified in the genome across squamates. Among squamates, the two non-colubroid snake species (*Boa constrictor* and *Python molurus*) show only a small number of TEs (although belonging to several subfamilies) being capable of originating transcript, in contrast to the green anole lizard and the prairie rattlesnake that show more highly dynamic TE trasncriptomes.

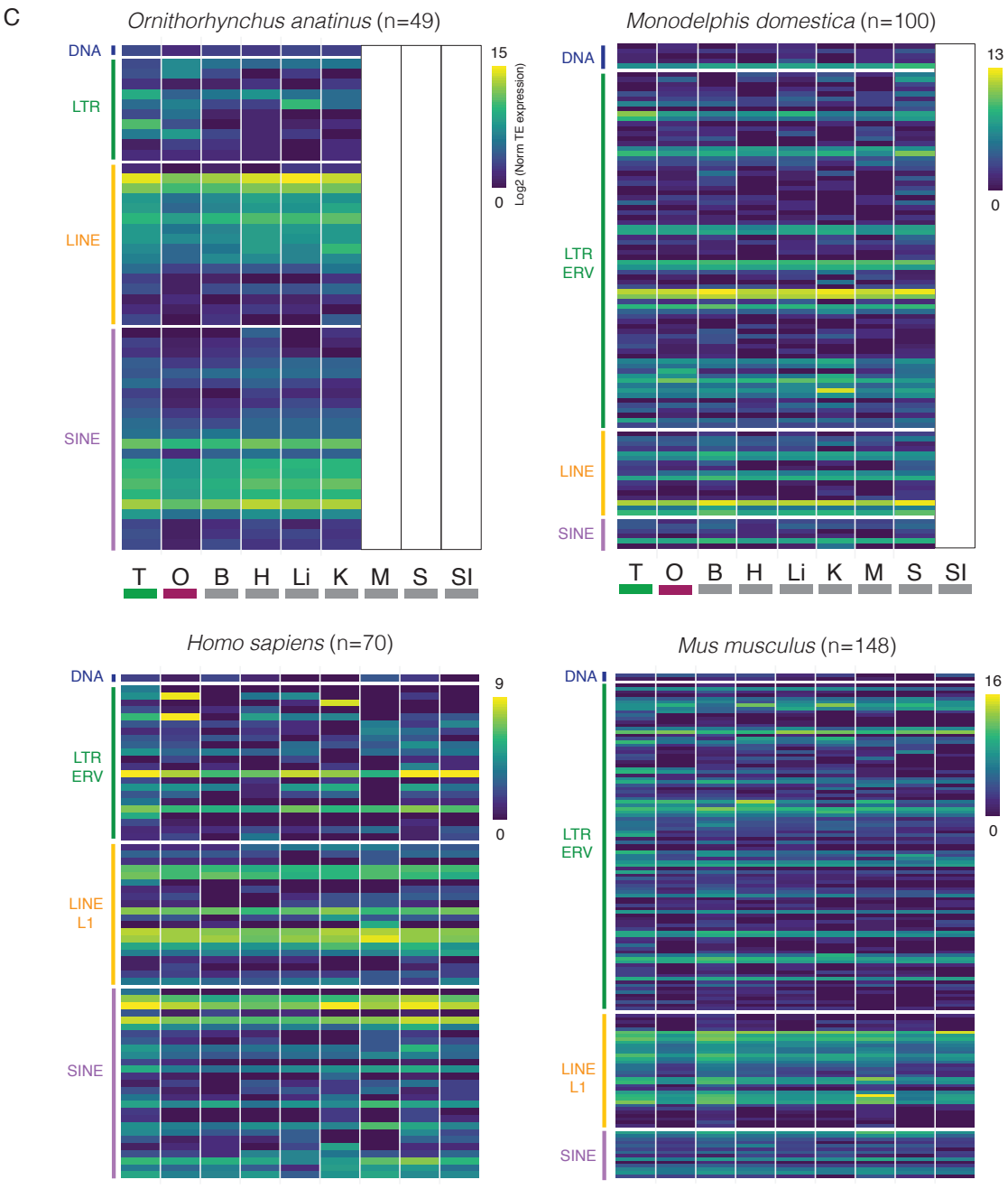


Figure supplement 7C. Potentially active TE-derived transcript expression across tissues in mammals. Heatmaps show individual TE expression levels across tissues of potentially active TEs across the mammalian radiation. Compared to other vertebrate species, mammals are characterized by a compelling small number of families (mostly L1 and SINEs in theria, and L2 LINEs in the platypus, and ERV LTRs) capable to originate transcripts. Surprisingly, we found also few DNA elements in our selected subsample of young TE inserts, which might either reflect incorrect mapping/attribution, or the result of pervasive transcription.

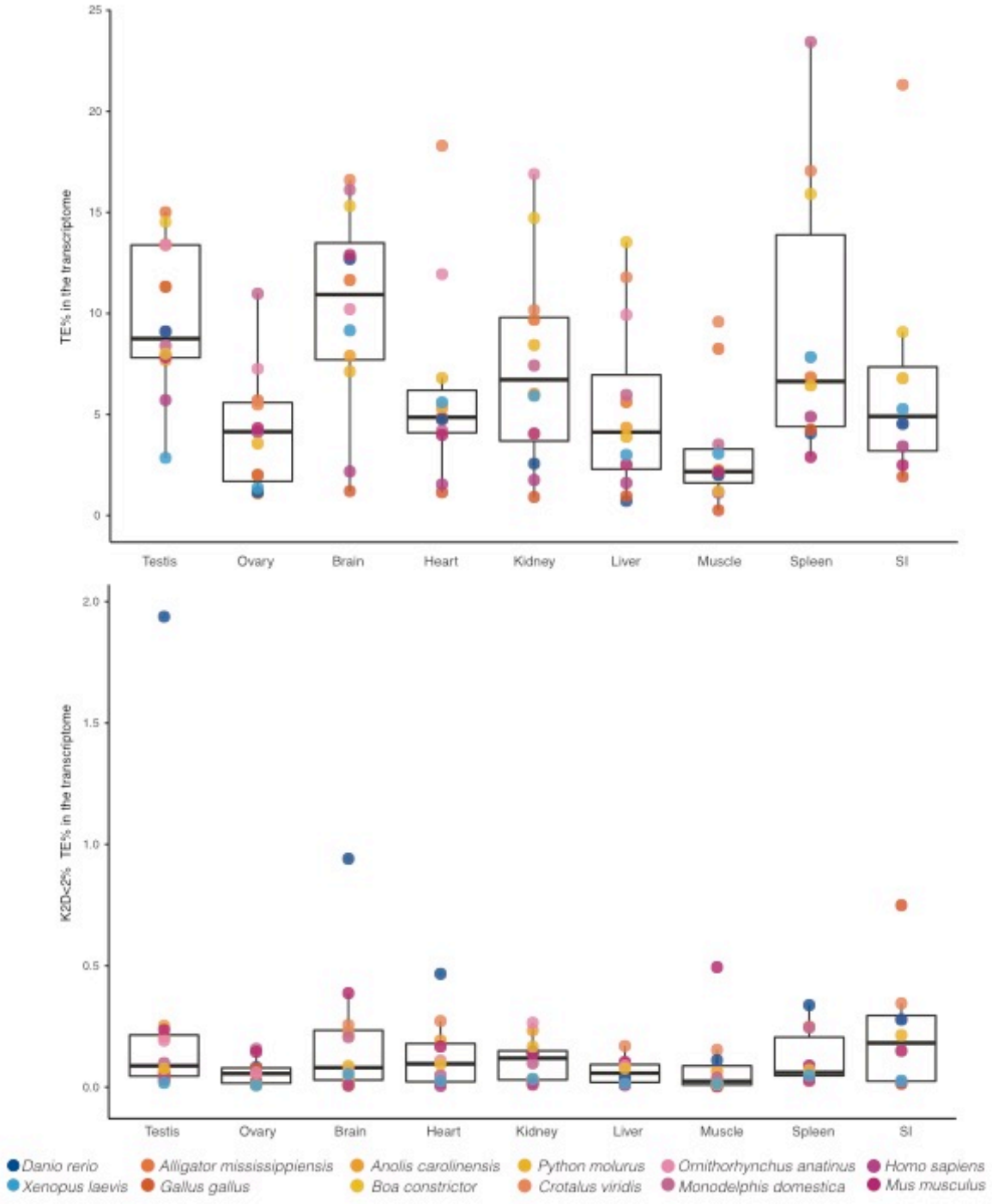


Figure supplement 8. Contribution of transposable elements (TEs) to the transcriptome of germine and somatic tissues. Box plots depicit the variation across species in the percent of the transcriptome (following within species normalization across tissues) corresponding to annotated TE-derived transcripts (total-TE expression; top) and to transcripts derived from recently inserted TE copies (recent-TEs; bottom). Analysis of TE transcripts between tissues for 12 vertebrate species highlights remarkable variance in TE expression levels both across species and among tissues per species. However, we find variability to be much lower when only recent-TEs are analyzed,which suggests widespread transcription of TE relics in vertebrate tissues.

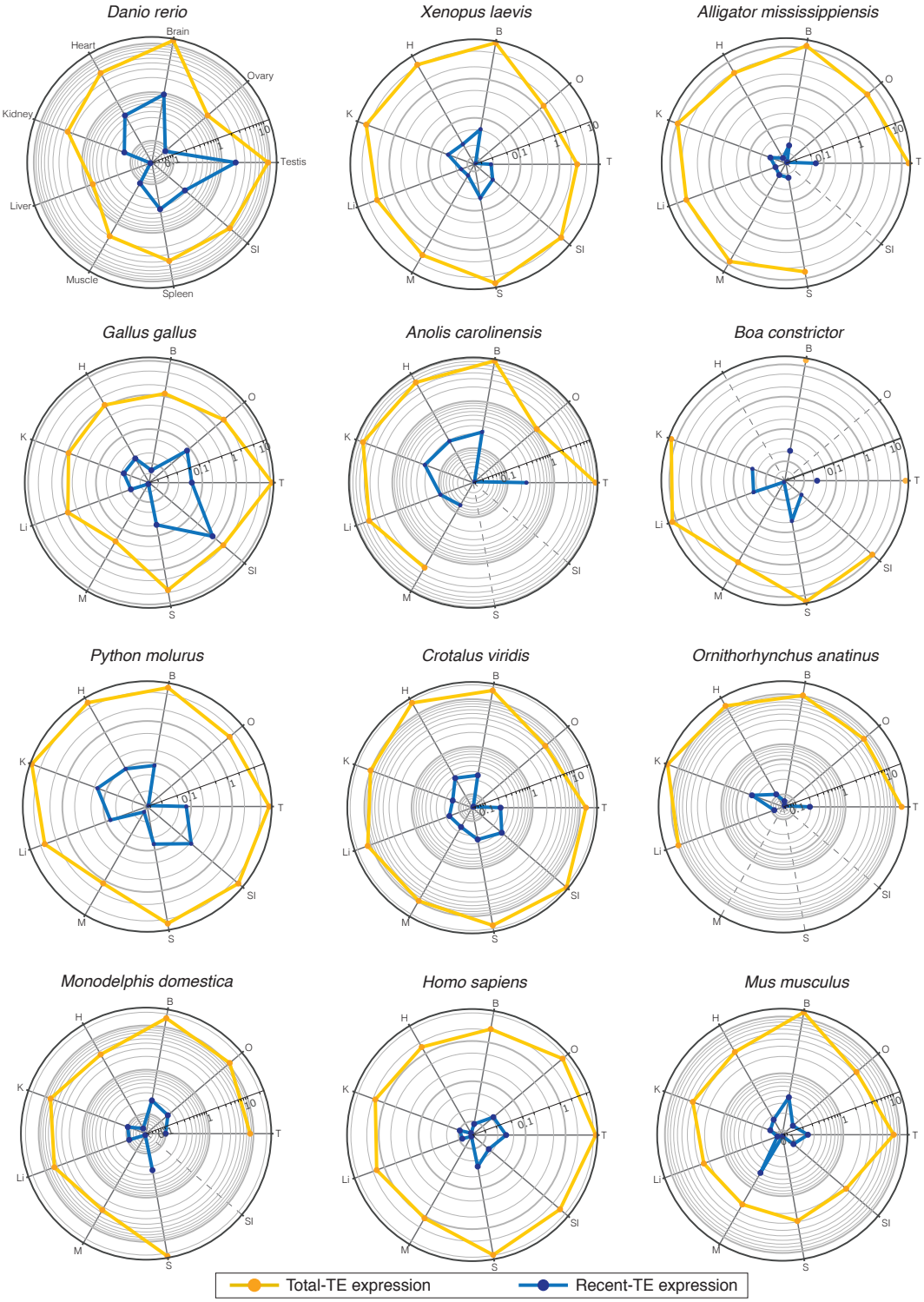


Figure supplement 9. Transcriptome estimates of total and recent TE expression levels. Radar plots show the percentage of the transcriptome (on a log10 scale) made up by total-TE transcripts (total-TEs; yellow) and by recently inserted TE copies (recent-TEs; blue) across somatic and germline tissues. We found abundant TE transcription to be a common physiological feature across all vertebrate tissues, whereas recent-TEs represent a marginal fraction of the transcriptomes.

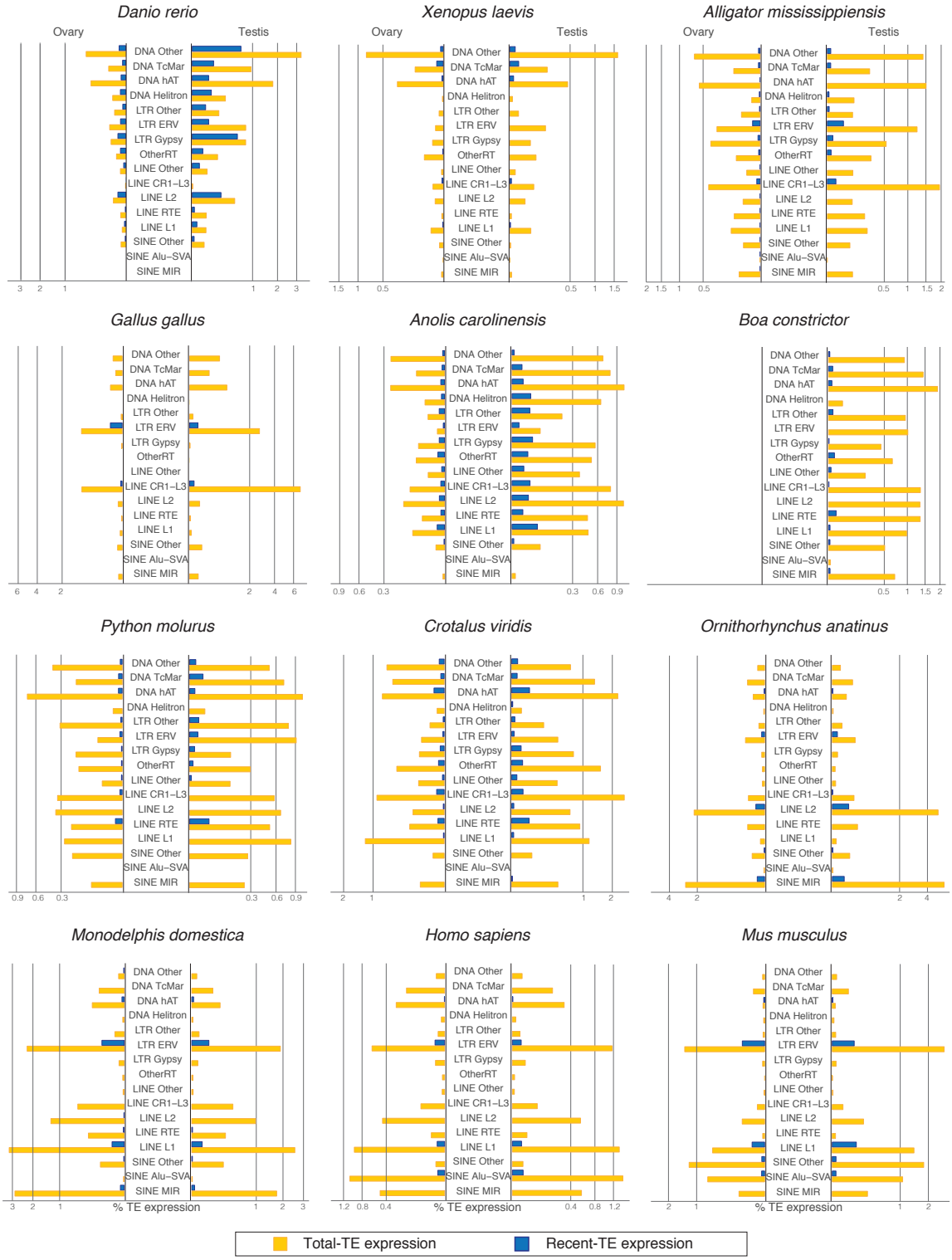


Figure supplement 10. Expression level estimates of total and potentially transpositionally-competent major TE families expression levels in germline tissues. For each species, barplots show the difference between total (yellow) and recently inserted TE copies (blue) expression level estimates in ovary (left) and testis (right) transcriptomes. Expression levels are reported as transcriptome percent and plotted in logaritmic scale.

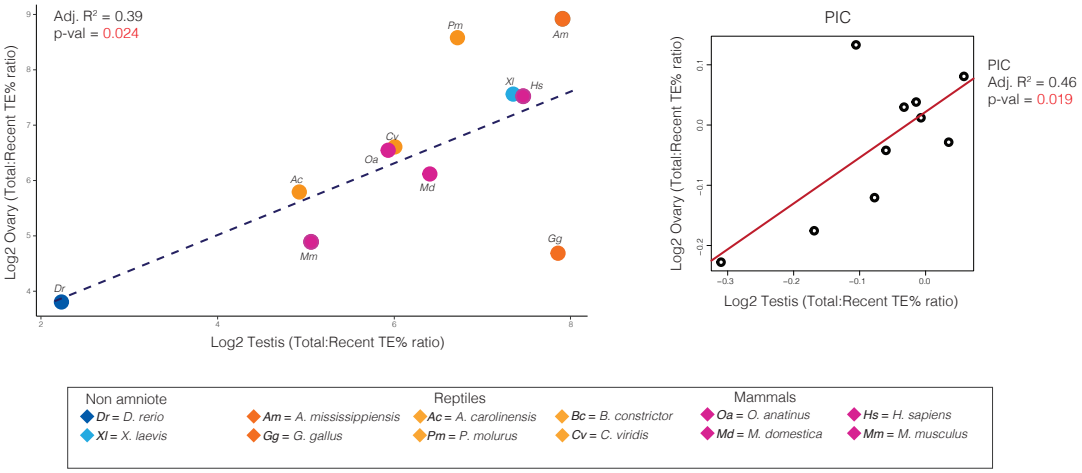


Figure supplement 11. Relationships between total and recent TE expression levels. Scatterplot and phylogenetically independent contrast (PIC) show a positive exponential correlation between the fold change in percentages of total-TE transcripts and recent-TE transcripts between testis and ovary across vertebrate species.

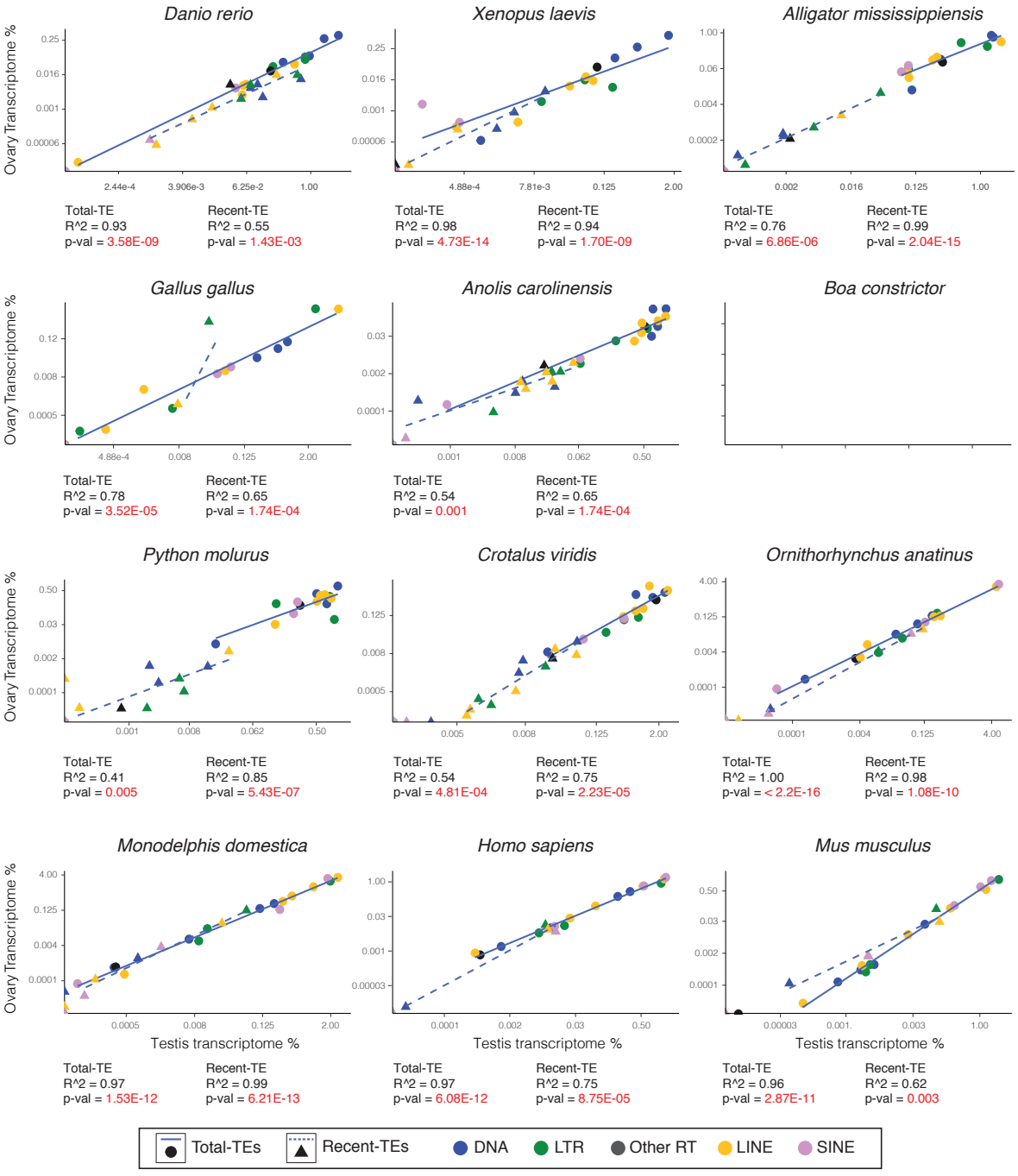


Figure supplement 12. Regression analyses of TE expression levels in germline tissues across vetrebrate species. Scatterplots show positive correlative trends of major TE subfamilies relative composition between the male and female germline in the total-TE (circles and solid lines) and recent-TE (triangles and dashed lines) datasets. Regression analyses were performed on transcriptome percent values. Scatterplots axes have been log2 transformed for display purposes. Reptile species display a stronger association (higher R^2 values) between germline tissue expression levels of TE subfamilies when only recent-TEs are considered compared to other vertebrate species, that show the opposite trend instead.

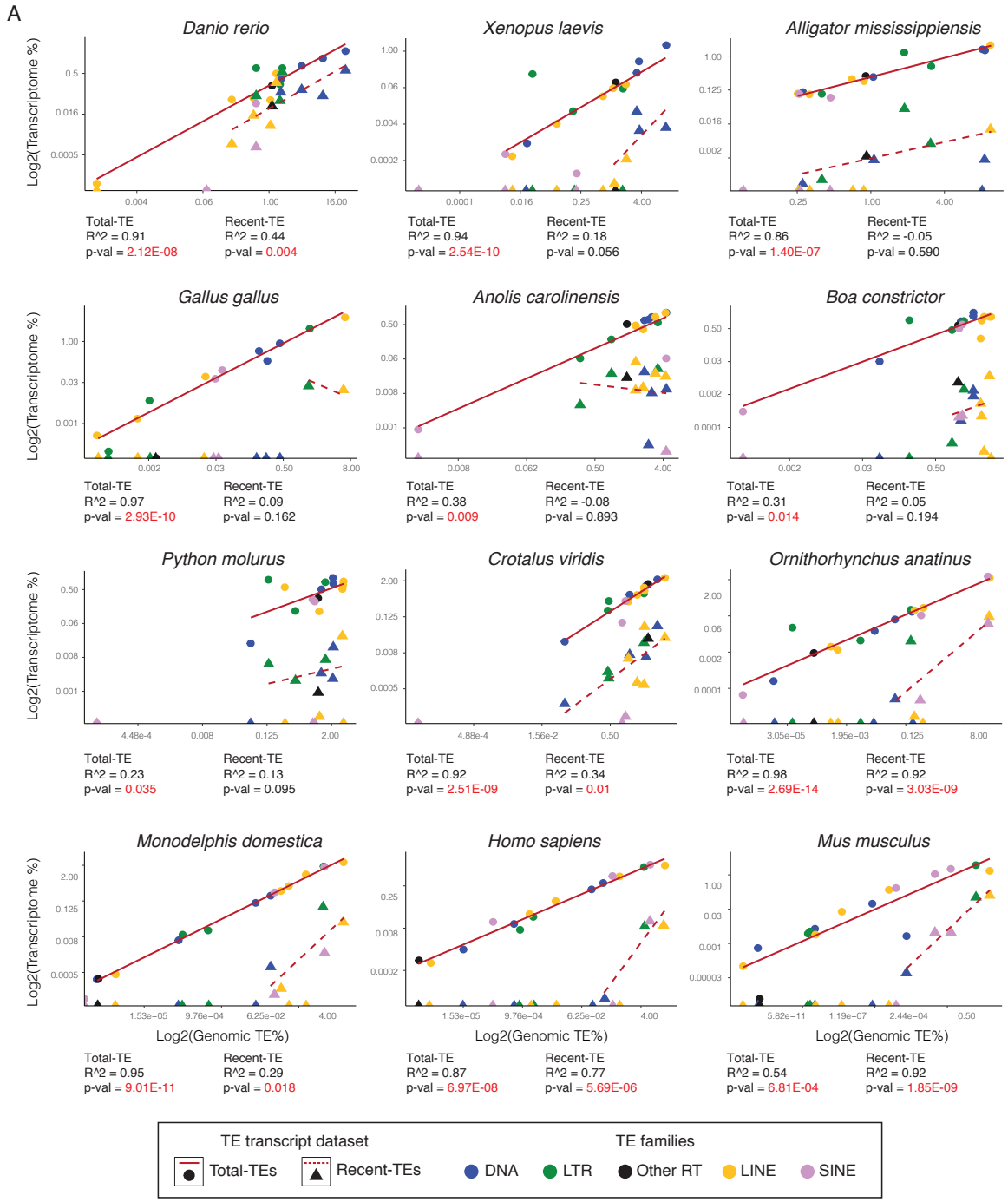


Figure supplement 13A. Regression analyses of TE expression levels and genomic TE content in germline tissues across vetrebrate species. A) Scatterplots show positive correlative trends of major TE families relative composition in the male germline between total transcriptome and genomic content (total-TEs; circles and solid lines), supporting a pervasive model of TE transcription. In contrast, linear regressions between relative composition of the total TE genomic content and recent-TE transcripts (triangles and dashed lines) show absence of a relationship in most non-mammal species, and a positive, although weaker correlation across mammals (and zebrafish and prairie rattlesnake). Regression analyses were performed on transcriptome percent and genomic percent values. Scatterplot axes have been log2 transformed for display purposes.

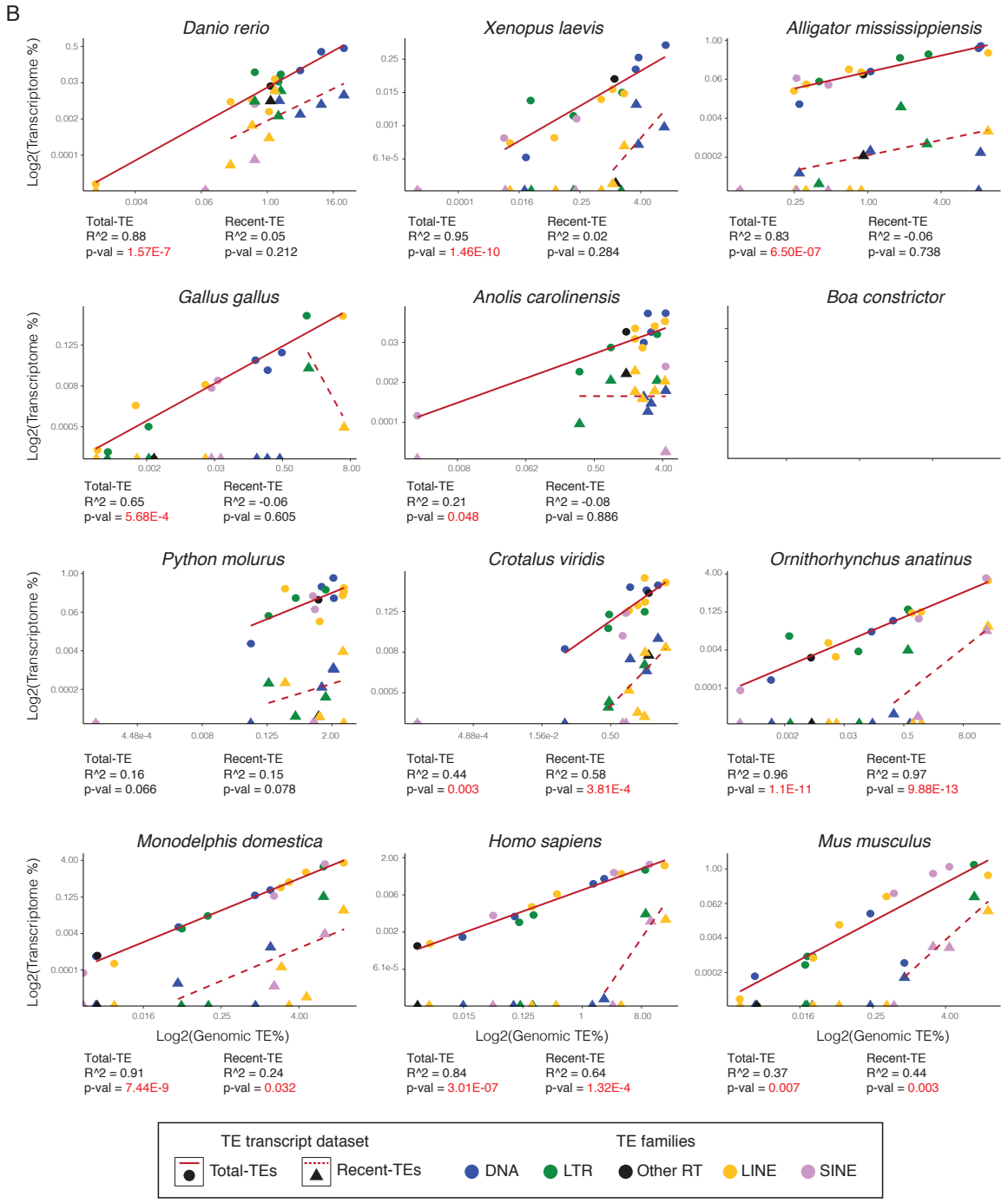


Figure supplement 13B. Regression analyses of expression levels and genomic TE content in germline tissues across vetrebrate species. B) Scatterplots show positive correlative trends of major TE families relative composition in the female germline between total transcriptome and genomic content (total-TEs; circles and solid lines), supporting a pervasive model of TE transcription. In contrast, linear regressions between relative composition of the total TE genomic content and recent-TE transcripts (triangles and dashed lines) show absence of a relationship in most non-mammal species, and a positive, generally weaker, correlation across mammals (and prairie rattlesnake). Regression analyses were performed on transcriptome percent and genomic percent values. Scatterplot axes have been log2 transformed for display purposes.

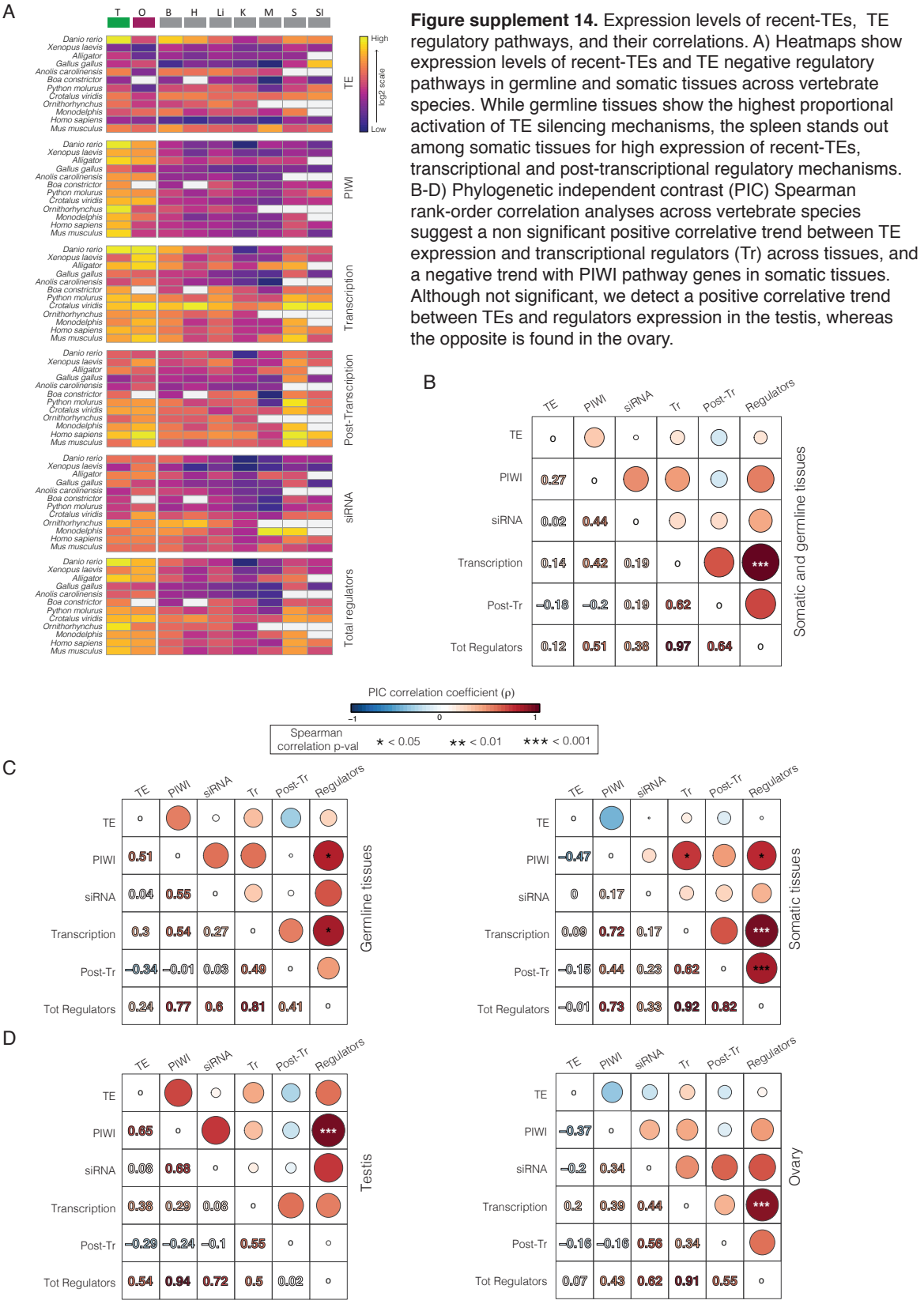


Figure supplement 14. Expression levels of recent-TEs, TE regulatory pathways, and their correlations. A) Heatmaps show expression levels of recent-TEs and TE negative regulatory pathways in germline and somatic tissues across vertebrate species. While germline tissues show the highest proportional activation of TE silencing mechanisms, the spleen stands out among somatic tissues for high expression of recent-TEs, transcriptional and post-transcriptional regulatory mechanisms. B-D) Phylogenetic independent contrast (PIC) Spearman rank-order correlation analyses across vertebrate species suggest a non significant positive correlative trend between TE expression and transcriptional regulators (Tr) across tissues, and a negative trend with PIWI pathway genes in somatic tissues. Although not significant, we detect a positive correlative trend between TEs and regulators expression in the testis, whereas the opposite is found in the ovary.

APPENDIX A

CHAPTER 2 SUPPLEMENTARY DATA

Supplementary Data 1. Genomic transposable element (TE) content in bird and mammal genomes.

Birds

Species	Common name	TE content	Source
Picoides pubescens	Downy woodpecker	22.55%	Kapusta et al, 2017[1]
Merops nubicus	Carmine bee-eater	8.48%	Kapusta et al, 2017
Apaloderma vittatum	Bar-tailed trogon	9.02%	Kapusta et al, 2017
Aptenodytes forsteri	Emperor penguin	6.16%	Kapusta et al, 2017
Pygoscelis adeliae	Adelie penguin	6.76%	Kapusta et al, 2017
Egretta garzetta	Little egret	7.69%	Kapusta et al, 2017
Calypte anna	Anna's hummingbird	8.68%	Kapusta et al, 2017
Chaetura pelagica	Chimney swift	9.60%	Kapusta et al, 2017
Antrostomus carolinensis	Chuck-Will's-widow	8.65%	Kapusta et al, 2017
Gallus gallus	Chicken	9.88%	Kapusta et al, 2017
Tinamus guttatus	White-throated tinamou	4.64%	Kapusta et al, 2017
Struthio camelus	Common ostrich	4.97%	Kapusta et al, 2017
Geospiza fortis	Medium ground-finch	8.53%	Kapusta et al, 2017
Corvus brachyrhynchos	American crow	8.72%	Kapusta et al, 2017
Manacus vitellinus	Golden-collared manakin	7.29%	Kapusta et al, 2017
Melopsittacus undulatus	Budgerigar	9.22%	Kapusta et al, 2017
Nestor notabilis	Kea	7.40%	Kapusta et al, 2017
Falco peregrinus	Peregrine falcon	6.05%	Kapusta et al, 2017
Colius striatus	Speckled mousebird	9.77%	Kapusta et al, 2017
Haliaeetus albicilla	White-tailed eagle	6.21%	Kapusta et al, 2017
Cathartes aura	Turkey vulture	5.24%	Kapusta et al, 2017
Chlamydotis macqueenii	MacQueen's bustard	6.86%	Kapusta et al, 2017
Tauraco erythrolophus	Red-crested turaco	9.40%	Kapusta et al, 2017
Cuculus canorus	Common cuckoo	10.04%	Kapusta et al, 2017

WITHOUT WOODPECKER	TE content	**WITH WOODPECKER**	TE content
Min	4.64%	Min	4.64%
Max	10.04%	Max	22.55%
Avg	7.79%	Avg	8.41%
Fold-variation	2.16	Fold-variation	4.86

[1] Kapusta A, Suh A, Feschotte C 2017. *PNAS* 114: E1460-E1469.

Mammals

Species	Common name	TE content	Source
Homo sapiens	Humans	48.49%	RepMasker genome database[2]
Pan troglodytes	Chimp	48.77%	RepMasker genome database
Gorilla gorilla	Gorilla	46.12%	RepMasker genome database
Pongo pygmaeus abelii	Orangutan	48.79%	Locke et al, 2011[3]
Macaca mulatta	Macaca	47.33%	RepMasker genome database
Callithrix jacchus	Marmoset	47.57%	Worley et al, 2014[4]
Tarsius syrichta	Tarsier	41.87%	Schmitz et al, 2016[5]
Rattus norvegicus	Rat	39.18%	RepMasker genome database
Mus musculus	Mouse	41.73%	RepMasker genome database
Cavia porcellus	Guinea Pig	37.06%	RepMasker genome database
Oryctolagus cuniculus	Rabbit	43.13%	RepMasker genome database
Myotis lucifugus	Microbat	35.51%	RepMasker genome database
Pteropus vampyrus	Megabat	33.40%	RepMasker genome database
Equus caballus	Horse	46.00%	Wade et al, 2009[6]
Bos taurus	Cow	47.98%	RepMasker genome database
Tursiops truncatus	Dolphin	41.24%	RepMasker genome database
Sus scrofa	Pig	43.10%	RepMasker genome database
Felis catus	Cat	41.48%	RepMasker genome database
Ailuropoda melanoleuca	Panda	39.20%	RepMasker genome database
Canis lupus familiaris	Dog	39.66%	RepMasker genome database
Erinaceus europeus	Hedgehog	42.84%	RepMasker genome database
Loxodonta africana	Elephant	56.38%	RepMasker genome database
Procavia capensis	Rock hyrax	50.53%	RepMasker genome database
Macropus eugenii	Tammar wallabi	52.80%	Renfree et al, 2011[7]
Monodelphis domestica	Opossum	52.20%	Mikkelsen et al, 2007[8]
Ornithorhynchus anatinus	Platypus	44.96%	Warren et al, 2008[9]

ALL MAMMALS	TE content
Min	33.40%
Max	56.38%
Avg	44.51%
Fold-variation	1.69

[2] Smit AFA, Hubley R, Green P. Last accessed 2017. http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html
[3] Locke DP, et al. 2011. *Nature* 469: 529-533. doi: 10.1038/nature09687
[4] Worley KC, et al. 2014. *Nature Genetics* 46: 850-857. doi: 10.1038/ng.3042
[5] Schmitz J, et al. 2016. *Nature Communications* 7. doi: 10.1038/ncomms12997
[6] Wade CM, et al. 2009. *Science* 326: 865-867. doi: 10.1126/science.1178158
[7] Renfree MB, et al. 2011. *Genome Biology* 12. doi: 10.1186/gb-2011-12-12-414
[8] Mikkelsen TS, et al. 2007. *Nature* 447: 167-U161. doi: 10.1038/nature05805
[9] Warren WC, et al. 2008. *Nature* 455: 256-256. doi: 10.1038/nature07253

Supplementary Data 2. Flow cytometry estimates of squamate, bird and mammal genome size (C-value). Dataset includes flow cytometry estimates for 86 squmate reptile, 170 mammal and 140 bird species or subspecies (Animal Genome Size database: http://www.genomesize.com)

Squamate

Family	Species	Common Name	C-value
Agamidae	*Amphibolurus longirostris*	Long-nosed water dragon	2.00
Agamidae	*Laudakia bochariensis*	Unknown	1.97
Agamidae	*Laudakia caucasia*	Rock agama	1.87
Agamidae	*Laudakia himalayana*	Rock agama	1.91
Agamidae	*Phrynocephalus helioscopus*	Toad-headed agama	2.08
Agamidae	*Phrynocephalus versicolor*	Toad-headed agama	1.95
Agamidae	*Pogona vitticeps*	Bearded dragon	1.81
Agamidae	*Trapelus sanguinolentus*	Steppe agama	1.72
Anguidae	*Anguis fragilis*	Slow worm	2.00
Anguidae	*Pseudopus apodus*	Armored glass lizard	1.90
Boidae	*Eryx jaculus*	Javelin sand boa	1.73
Boidae	*Python curtus*	Blood python	1.83
Colubridae	*Chironius fuscus*	N/A	2.24
Colubridae	*Coluber najadum*	Dahl's whip snake	1.77
Colubridae	*Coluber nummifer*	Desert whip snake	1.73
Colubridae	*Coluber ravergieri*	Ravergier's whip snake	1.71
Colubridae	*Coluber schmidti*	Whip snake	1.65
Colubridae	*Eirenis collaris*	Collared dwarf racer	1.80
Colubridae	*Eirenis punctatolineatus*	Dotted dwarf racer	1.86
Colubridae	*Elaphe quatuorlineata*	Bulgarian ratsnake	1.83
Colubridae	*Liophis miliaris*	Swampsnake	2.01
Colubridae	*Natrix natrix*	European grass snake	1.99
Colubridae	*Natrix tessellata*	Dice snake	1.91
Colubridae	*Nerodia rhombifera*	Broad-banded water snake	2.03
Colubridae	*Nerodia sipedon*	Northern water snake	1.90
Colubridae	*Nerodia sipedon*	Northern water snake	2.00
Colubridae	*Oxyrhopus petola*	Red-banded snake	1.54
Colubridae	*Tantilla melanocephala*	Black-headed snake	2.25
Colubridae	*Telescopus fallax*	Cat snake	1.87
Colubridae	*Thamnophis sirtalis*	Common garter snake	1.91
Elapidae	*Furina ornata*	Orange-naped snake	2.00
Elapidae	*Micrurus lemniscatus*	South American coral snake	1.85
Elapidae	*Simoselaps anomalus*	Eastern brown snake	1.81
Elapidae	*Simoselaps fasciolatus*	Narrow-banded snake	1.85
Elapidae	*Simoselaps incinctus*	Burowing snake	1.85
Eublepharidae	*Coleonyx brevis*	Texas banded gecko	2.01
Eublepharidae	*Coleonyx elegans*	Banded gecko	1.56
Eublepharidae	*Coleonyx mitratus*	Banded gecko	1.76
Eublepharidae	*Coleonyx variegatus*	Western banded gecko	1.92
Eublepharidae	*Eublepharis angramainyu*	Eyelid gecko	2.00
Eublepharidae	*Eublepharis cf. fuscus*	Eyelid gecko	1.88
Eublepharidae	*Eublepharis macularius*	Leopard gecko	1.86
Eublepharidae	*Goniurosaurus araneus*	Eyelid gecko	1.87
Eublepharidae	*Goniurosaurus lichtenfelderi*	Eyelid gecko	1.85
Eublepharidae	*Goniurosaurus luii*	Eyelid gecko	1.87
Eublepharidae	*Hemitheconyx caudicinctus*	Eyelid gecko	1.81
Eublepharidae	*Holodactylus africanus*	Eyelid gecko	1.76
Gekkonidae	*Gekko sinensis*	Gecko	2.83
Iguanidae	*Cyclura cornuta*	Horned ground iguana	1.80
Lacertidae	*Eremias grammica*	Reticulate racerunner	1.90
Lacertidae	*Eremias multiocellata*	Multi-cellated racerunner	1.73
Lacertidae	*Lacerta agilis*	Sand lizard	1.60
Lacertidae	*Lacerta viridis*	Green lacerta lizard	1.66
Lacertidae	*Lacerta vivipara*	European common lizard	1.64
Lacertidae	*Ophisops elegans*	Snake-eyed lizard	1.57
Lacertidae	*Podarcis muralis*	European wall lizard	1.70
Polychrotidae	*Anolis carolinensis*	Green anole	2.20
Polychrotidae	*Anolis cf. nitens, sp.1*	Anole	2.29
Polychrotidae	*Anolis cf. nitens, sp.2*	Anole	2.49
Scincidae	*Asymblepharus alaicus*	Skink	1.93
Scincidae	*Carlia triacantha*	Desert rainbow skink	1.41
Scincidae	*Cryptoblepharus plagiocephalus*	Snake-eyed skink	1.51
Scincidae	*Ctenotus alacer*	Skink	1.57
Scincidae	*Ctenotus leonhardii*	Leonhard's skink	1.65
Scincidae	*Ctenotus quattuordecimlineatus*	Fourteen-lined skink	1.59
Scincidae	*Ctenotus saxatilis*	Skink	1.50
Scincidae	*Egernia inornata*	Desert skink	1.69
Scincidae	*Eremiascincus fasciolatus*	Narrow-banded sand swimmer	1.61
Scincidae	*Lerista desertorum*	Skink	1.64
Scincidae	*Lerista frosti*	Skink	2.60
Scincidae	*Lerista labialis*	Southern sandslider	1.73
Scincidae	*Mabuya mabouya*	American shiny skink	1.27
Scincidae	*Menetia greyii*	Common dwarf skink	1.71
Scincidae	*Morethia ruficauda*	Southwestern mulch skink	1.64
Scincidae	*Tiliqua scincoides*	Eastern blue-tongued lizard	1.82
Teiidae	*Kentropyx calcarata*	N/A	1.55

Tropiduridae	*Liolaemus sp.*	Snow swift	2.05
Tropiduridae	*Tropidurus umbra*	Green tree climber	2.11
Typhlopidae	*Ramphotyphlops braminus*	Flowerpot blindsnake	2.98
Typhlopidae	*Typhlops vermicularis*	European blindsnake	1.96
Varanidae	*Varanus komodoensis*	Komodo dragon	1.93
Varanidae	*Varanus niloticus*	Nile monitor	2.19
Varanidae	*Varanus salvadorii*	Crocodile monitor	2.29
Viperidae	*Bothrops atrox*	Fer-de-lance	1.82
Viperidae	*Crotalus horridus*	Timber rattlesnake	1.75
Viperidae	*Vipera berus*	Adder	1.88
Viperidae	*Vipera eriwanensis*	Steppe adder	1.79

Mammals

Family	Species	Common Name	C-value
Bovidae	*Addax nasomaculatus*	Addax	3.98
Bovidae	*Aepyceros melampus*	Impala	4.69
Bovidae	*Bos javanicus javanicus*	Banteng	3.75
Bovidae	*Bos taurus*	Domestic cattle	3.60
Bovidae	*Bos taurus*	Domestic cattle	3.65
Bovidae	*Bos taurus*	Domestic cattle	3.70
Bovidae	*Capra falconeri*	Markhor	3.23
Bovidae	*Gazella dama ruficollis*	Dama gazelle	3.48
Bovidae	*Gazella granti*	Grant's gazelle	3.30
Bovidae	*Litocranius walleri*	Gerenuk	3.66
Bovidae	*Ovis aries aries*	Sheep	3.41
Bovidae	*Tragelaphus angasii*	Lowland nyala	3.94
Camelidae	*Camelus bactrianus*	Bactrian camel	2.41
Camelidae	*Camelus dromedarius*	Dromedary camel	2.62
Cervidae	*Muntiacus muntjak*	Barking deer	3.44
Cervidae	*Muntiacus muntjak vaginalis*	Indian muntjac	2.22
Cervidae	*Muntiacus reevesi*	Reeves's (Chinese) muntjac	2.85
Cervidae	*Rangifer tarandus*	Reindeer, caribou	3.41
Giraffidae	*Giraffa camelopardalis*	Giraffe	2.85
Giraffidae	*Giraffa camelopardalis*	Reticulated giraffe	2.69
Suidae	*Sus scrofa domesticus*	Domestic pig	2.99
Suidae	*Sus scrofa domesticus*	Domestic pig	3.00
Suidae	*Sus scrofa domesticus*	Domestic pig	3.21
Suidae	*Sus scrofa scrofa*	Vietnamese pot-bellied pig	2.81
Canidae	*Canis familiaris*	Domestic dog	2.80
Canidae	*Canis familiaris*	Domestic dog	2.85
Canidae	*Canis familiaris*	Domestic dog	2.88
Canidae	*Canis latrans*	Coyote	2.82
Canidae	*Canis lupus*	Timber wolf	2.81
Canidae	*Canis rufus*	Red wolf	3.04
Canidae	*Lycaon pictus*	African hunting dog	2.73
Canidae	*Nyctereutes procyonoides procyonoides*	Chinese raccoon dog	3.27
Canidae	*Nyctereutes procyonoides viverrinus*	Japanese raccoon dog	3.19
Canidae	*Urocyon cinereoargenteus*	Gray fox	3.07
Canidae	*Vulpes vulpes*	Red fox	2.85
Felidae	*Felis catus*	Domestic cat	2.91
Felidae	*Felis catus*	Domestic cat	3.10
Felidae	*Felis lynx*	Lynx	2.92
Felidae	*Felis silvestris*	Wildcat	2.92
Felidae	*Felis silvestris*	Wildcat	3.00
Felidae	*Neofelis nebulosa*	Clouded leopard	2.77
Felidae	*Panthera leo*	African lion	2.95
Felidae	*Panthera tigris*	Tiger	2.71
Felidae	*Panthera tigris tigris*	Bengal tiger	2.90
Mustelidae	*Mustela putorius*	Domestic ferret	2.81
Otariidae	*Zalophus californianus*	California sea lion	3.15
Phocidae	*Phoca largha*	Spotted seal / Largha seal	2.94
Procyonidae	*Procyon lotor*	Raccoon	2.85
Ursidae	*Ursus arctos*	Brown bear	2.75
Ursidae	*Ursus hibetanus*	Himalayan black bear	2.75
Delphinidae	*Sousa chinensis chinensis*	Chinese white dolphin	3.46
Delphinidae	*Tursiops truncatus*	Bottlenose dolphin	3.27
Lipotidae	*Lipotes vexillifer*	Chinese river dolphin / Baiji	3.91
Monodontidae	*Delphinapterus leucas*	Beluga whale	3.29
Phocoenidae	*Neophocaena phocaenoides asiaorientalis*	Yangtze finless porpoise	3.46
Mormoopidae	*Pteronotus parnellii*	Parnell's moustached bat	2.67
Mormoopidae	*Pteronotus personatus*	Moustached bat	2.84
Noctilionidae	*Noctilio leporinus*	Greater bulldog bat	2.63
Phyllostomidae	*Artibeus jamaicensis*	Fruit bat	2.74
Phyllostomidae	*Artibeus lituratus*	Big fruit bat	2.70
Phyllostomidae	*Carollia brevicauda*	Short-tailed fruit bat	2.93
Phyllostomidae	*Carollia perspicillata*	Short-tailed fruit bat	3.06
Phyllostomidae	*Dermanura phaeotis*	Dwarf fruit bat	2.85
Phyllostomidae	*Dermanura tolteca*	Dwarf fruit bat	2.71
Phyllostomidae	*Dermanura watsoni*	Dwarf fruit bat	2.73
Phyllostomidae	*Glossophaga soricina*	Pallas's long-tongued bat	2.78
Phyllostomidae	*Lonchorhina aurita*	Tomes's sword-nosed bat	2.56
Phyllostomidae	*Macrophyllum macrophyllum*	Long-legged bat	3.29
Phyllostomidae	*Mimon cozumelae*	Cozumel spear-nosed bat	2.47
Phyllostomidae	*Phylloderma stenops*	Peters's spear-nosed bat	2.45

Phyllostomidae	*Phyllostomus discolor*	Pale spear-nosed bat	2.52
Phyllostomidae	*Sturnira lilium*	Common yellow-shouldered bat	2.84
Phyllostomidae	*Tonatia bidens*	Round-eared bat	2.35
Phyllostomidae	*Tonatia evotis*	Davis's round-eared bat	2.51
Phyllostomidae	*Trachops cirrhosus*	Frog-eating bat	2.41
Phyllostomidae	*Uroderma bilobatum*	Tent-building bat	2.67
Phyllostomidae	*Vampyressa pusilla*	Yellow-eared bat	2.73
Phyllostomidae	*Vampyrodes caraccioli*	Great stripe-faced bat	2.49
Vespertilionidae	*Eptesicus furinalis*	Big brown bat	2.43
Vespertilionidae	*Lasiuris borealis*	Red bat	2.56
Vespertilionidae	*Lasiuris ega*	Yellow bat	2.93
Vespertilionidae	*Lasiuris intermedius*	Hairy-tailed bat	2.91
Vespertilionidae	*Myotis keaysi*	Little brown bat	2.65
Vespertilionidae	*Rhogeessa tumida*	Little yellow bat	2.80
Didelphidae	*Didelphis virginiana*	Virginia opossum	4.15
Macropodidae	*Macropus parma*	Parma wallaby	4.02
Macropodidae	*Macropus rufogrigeus*	Bennett's wallaby	5.58
Soricidae	*Sorex araneus*	Common shrew	2.91
Talpidae	*Talpa occidentalis*	Mole	2.50
Leporidae	*Lepus timidus*	Mountain hare	3.25
Leporidae	*Oryctolagus cuniculus*	Rabbit	3.26
Leporidae	*Oryctolagus cuniculus*	Rabbit	3.42
Equidae	*Equus caballus*	Horse	3.15
Equidae	*Equus caballus*	Horse	3.21
Rhinocerotidae	*Diceros bicornis*	Black rhinoceros	3.34
Tapiridae	*Tapirus bairdii*	Baird's tapir	2.54
Tapiridae	*Tapirus indicus*	Malayan tapir	2.75
Cercopithecidae	*Chlorocebus sabaeus*	Green monkey	4.04
Hominidae	*Gorilla gorilla*	Gorilla	4.16
Hominidae	*Gorilla gorilla*	Western lowland gorilla	3.52
Hominidae	*Pan troglodytes*	Chimpanzee	3.46
Hominidae	*Pan troglodytes*	Chimpanzee	3.76
Hominidae	*Pongo pygmaeus*	Orangutan	3.60
Lemuridae	*Eulemur coronatus*	Crowned lemur	3.47
Lemuridae	*Eulemur coronatus x Eulemur macaco*	Lemur (hybrid)	3.05
Lemuridae	*Eulemur fulvus albocollaris*	Brown lemur	2.73
Lemuridae	*Eulemur fulvus mayottensis*	Brown lemur	2.86
Lemuridae	*Eulemur macaco*	Black lemur	2.68
Lemuridae	*Eulemur macaco*	Black lemur	2.74
Lemuridae	*Eulemur rubriventer*	Red-bellied lemur	2.59
Elephantidae	*Elephas maximus*	Asian elephant	4.03
Elephantidae	*Loxodonta africana*	African elephant	4.11
Abrocomidae	*Abrocoma bennetti*	Chincilla rat	3.60
Bathyergidae	*Bathyergus suillus*	Cape dune mole-rat	2.90
Bathyergidae	*Cryptomys damarensis*	Damaraland mole-rat	3.60
Bathyergidae	*Cryptomys hottentotus*	African mole-rat	3.40
Bathyergidae	*Georhychus capensis*	African mole-rat	3.20
Bathyergidae	*Heliophobius argenteocinereus*	African silvery mole-rat	3.05
Bathyergidae	*Heterocephalus glaber*	Naked mole-rat	2.90
Caviidae	*Cavia porcellus*	Guinea pig	3.92
Caviidae	*Cavia porcellus*	Guinea pig	4.10
Caviidae	*Cavia tschudii*	Wild guinea pig	4.55
Caviidae	*Dolichotis patagonum*	Patagonian cavy	3.70
Caviidae	*Dolichotis salinicola*	Mara	3.85
Caviidae	*Galea musteloides*	Cui	3.95
Caviidae	*Microcavia australis*	Rock cavy	3.10
Chinchillidae	*Lagostomus maximus*	Plains viscacha	3.30
Ctenomyidae	*Ctenomys boliviensis*	Tuco-tuco	3.77
Ctenomyidae	*Ctenomys boliviensis*	Tuco-tuco	4.30
Ctenomyidae	*Ctenomys conoveri*	Tuco-tuco	3.90
Ctenomyidae	*Ctenomys fochi*	Tuco-tuco	4.55
Ctenomyidae	*Ctenomys frater*	Tuco-tuco	3.77
Ctenomyidae	*Ctenomys leucodon*	Tuco-tuco	3.88
Ctenomyidae	*Ctenomys lewisi*	Tuco-tuco	3.56
Ctenomyidae	*Ctenomys mendocinus*	Tuco-tuco	4.90
Ctenomyidae	*Ctenomys opimus*	Tuco-tuco	2.80
Ctenomyidae	*Ctenomys opimus*	Tuco-tuco	3.16
Ctenomyidae	*Ctenomys opimus*	Tuco-tuco	4.80
Ctenomyidae	*Ctenomys porteousi*	Tuco-tuco	3.20
Ctenomyidae	*Ctenomys steinbachi*	Tuco-tuco	3.73
Echimyidae	*Proechimys semispinosus*	Central American spiny rat	3.85
Geomyidae	*Geomys attwateri*	Eastern pocket gopher	3.31
Geomyidae	*Geomys breviceps*	Eastern pocket gopher	3.35
Geomyidae	*Geomys bursarius major*	Eastern pocket gopher	3.67
Geomyidae	*Geomys knoxjonesi*	Eastern pocket gopher	3.67
Geomyidae	*Thomomys bottae actuosus*	Western pocket gopher	4.21
Geomyidae	*Thomomys bottae alienus*	Western pocket gopher	4.71
Geomyidae	*Thomomys bottae bottae*	Western pocket gopher	5.30
Geomyidae	*Thomomys bottae fulvus*	Western pocket gopher	4.58
Geomyidae	*Thomomys bottae grahamensis*	Western pocket gopher	4.86
Geomyidae	*Thomomys bottae ingens*	Western pocket gopher	5.26
Geomyidae	*Thomomys bottae internatus*	Western pocket gopher	4.30
Geomyidae	*Thomomys bottae laticeps*	Western pocket gopher	5.01
Geomyidae	*Thomomys bottae leucodon*	Western pocket gopher	4.98
Geomyidae	*Thomomys bottae mewa*	Western pocket gopher	5.22
Geomyidae	*Thomomys bottae minor*	Western pocket gopher	5.19

Geomyidae	*Thomomys bottae morulus*	Western pocket gopher	4.38
Geomyidae	*Thomomys bottae planorum*	Western pocket gopher	5.37
Geomyidae	*Thomomys bottae ruidosae*	Western pocket gopher	4.67
Geomyidae	*Thomomys bottae sylvifugus*	Western pocket gopher	5.59
Geomyidae	*Thomomys monticola*	Western pocket gopher	2.17
Geomyidae	*Thomomys talpoides fossor*	Western pocket gopher	2.57
Geomyidae	*Thomomys townsendii relictus*	Western pocket gopher	5.20
Geomyidae	*Thomomys townsendii similis*	Western pocket gopher	5.38
Geomyidae	*Thomomys townsendii townsendii*	Western pocket gopher	5.60
Geomyidae	*Thomomys umbrinus intermedius*	Western pocket gopher	4.38
Hystricidae	*Hystrix africaenaustralis*	Short-tailed porcupine	2.90
Muridae	*Apodemus sylvaticus*	Wood mouse	3.29
Muridae	*Calomyscus mystax*	Mouse-like hamster	3.11
Muridae	*Cricetus cricetus*	Common hamster	3.57
Muridae	*Graomys centralis*	N/A	3.69
Muridae	*Graomys griseoflavus*	N/A	2.81
Muridae	*Mesocricetus auratus*	Golden hamster	3.43
Muridae	*Mus musculus*	House mouse	3.25
Muridae	*Mus musculus*	House mouse	3.26
Muridae	*Mus musculus*	House mouse	3.35
Muridae	*Ondatra zibethicus*	Muskrat	2.78
Muridae	*Rattus norvegicus*	Brown rat	3.05
Muridae	*Rattus norvegicus*	Brown rat	3.36
Muridae	*Rattus rattus*	Black rat	3.03
Myocastoridae	*Myocastor coypus*	Nutria, coypu	3.60
Octodontidae	*Aconaemys fuscus*	Viscacha rat	3.75
Octodontidae	*Aconaemys porteri*	Viscacha rat	3.70
Octodontidae	*Aconaemys sagei*	Viscacha rat	3.70
Octodontidae	*Octodon bridgesi*	Viscacha rat	3.85
Octodontidae	*Octodon degus*	Degu	4.30
Octodontidae	*Octodon degus*	Degu	4.32
Octodontidae	*Octodon lunatus*	Viscacha rat	4.40
Octodontidae	*Octodontomys gliroides*	Chozchoz	4.10
Octodontidae	*Octomys mimax*	Viscacha rat	3.80
Octodontidae	*Octomys mimax*	Viscacha rat	4.00
Octodontidae	*Spalacopus cyanus*	Coruro	3.54
Octodontidae	*Spalacopus cyanus*	Coruro	4.20
Octodontidae	*Tympanoctomys barrerae*	Red viscacha rat	8.40

Birds

Family	Species	Common Name	C-value
Anatidae	*Anas capensis*	Cape teal	1.41
Anatidae	*Anas castanea*	Chestnut-breasted teal	1.36
Anatidae	*Anas platyrhynchos*	Mallard	1.44
Anatidae	*Anas platyrhynchos*	Mallard	1.46
Anatidae	*Anas platyrhynchos*	Mallard	1.49
Anatidae	*Anas platyrhynchos*	Mallard	1.54
Anatidae	*Anser rossii*	Ross's goose	1.38
Anatidae	*Cairina scutulata*	White-winged wood duck	1.49
Anatidae	*Cygnus atratus*	Black swan	1.45
Anatidae	*Cygnus buccinator*	Trumpeter swan	1.54
Anatidae	*Cygnus melanocoryphus*	Black-necked swan	1.53
Anatidae	*Cygnus olor*	Mute swan	1.48
Anatidae	*Dendrocygna viduata*	White-faced whistling duck	1.37
Anatidae	*Mergus cucullatus*	Hooded merganser	1.23
Anatidae	*Mergus cucullatus*	Hooded merganser	1.28
Anatidae	*Mergus merganser*	American merganser	2.00
Anatidae	*Netta rufina*	Red-crested pochard	1.38
Dromaiidae	*Dromaius novaehollandiae*	Emu	1.63
Recurvirostridae	*Recurvirostra avosetta*	Pied avocet	1.59
Ciconiidae	*Anastomus lamelligerus*	African open-billed stork	1.62
Ciconiidae	*Ciconia abdimii*	Adbim's stork	1.53
Ciconiidae	*Ciconia ciconia*	White stork	1.58
Ciconiidae	*Ciconia episcopus*	Woolly-necked stork	1.61
Ciconiidae	*Ciconia maguari*	Maguari stork	1.54
Ciconiidae	*Ciconia stormi*	Storm's stork	1.68
Ciconiidae	*Leptoptilos crumeniferus*	Marabou stork	1.55
Phoenicopteridae	*Phoenicopterus ruber*	Greater flamingo	1.52
Columbidae	*Caloenas nicobarica*	Nicobar pigeon	1.54
Columbidae	*Columba livia*	Rock pigeon	1.46
Columbidae	*Columba livia*	Rock pigeon	1.59
Columbidae	*Ducula bicolor*	Pied imperial pigeon	1.60
Columbidae	*Goura cristata*	Western crowned pigeon	1.39
Columbidae	*Goura victoria*	Victoria crowned pigeon	1.38
Columbidae	*Ptilinopus perlatus*	Pink-spotted dove	1.48
Columbidae	*Ptilinopus pulchellus*	Crimson-capped fruit dove	1.47
Columbidae	*Ptilinopus regina*	Rose-crowned fruit dove	1.37
Columbidae	*Trugon terrestris*	Thick-billed ground pigeon	1.70
Bucerotidae	*Bucorvus abyssinicus*	Abyssinian ground hornbill	1.43
Accipitridae	*Aegypius monachus gingintanus*	Cinereous vulture	1.59
Accipitridae	*Aguila chrysaetos*	Golden eagle	1.48
Accipitridae	*Buteo buteo*	Eurasian buzzard	1.53
Accipitridae	*Circus aeruginosus*	Marsh harrier	1.43
Accipitridae	*Circus cyaneus*	Northern harrier	1.42

Accipitridae	*Circus pygargus*	Montagu's harrier	1.46
Accipitridae	*Elanus caerulens*	Black-winged kite	1.55
Accipitridae	*Gyps bengalensis*	Asian white-backed vulture	1.39
Accipitridae	*Haliaeetus leucocephalus*	Bald eagle	1.43
Accipitridae	*Harpia harpyja*	Harpy eagle	1.58
Accipitridae	*Milvus migrans*	Black kite	1.47
Accipitridae	*Neophron percnopterus*	Egyptian vulture	1.58
Accipitridae	*Neophron percnopterus*	Egyptian vulture	1.60
Accipitridae	*Pernis apivorus*	Honey buzzard	1.53
Falconidae	*Falco peregrinus*	Peregrine falcon	1.45
Falconidae	*Falco sparverius*	American kestrel	1.43
Falconidae	*Falco tinnunculus*	Eurasian kestrel	1.54
Falconidae	*Falco vespertinus*	Red-footed falcon	1.47
Phasianidae	*Argusianus argus*	Great argus	1.52
Phasianidae	*Chrysolophus pictus*	Golden pheasant	1.21
Phasianidae	*Coturnix coturnix*	Quail	1.35
Phasianidae	*Coturnix japonica*	Japanese quail	1.41
Phasianidae	*Guttera edouardi*	Crested (Kenyan) guineafowl	1.54
Phasianidae	*Meleagris gallopavo*	Turkey	1.31
Phasianidae	*Numida meleagris*	Guineafowl	1.31
Phasianidae	*Pavo muticus*	Green peafowl	1.34
Phasianidae	*Phasianus colchicus*	Common pheasant	1.26
Cariamidae	*Cariama cristata*	Red-legged seriema	1.50
Gruidae	*Anthropoides paradisea*	Stanley crane	1.52
Gruidae	*Balearica pavonina*	Black (W. African) crowned crane	1.48
Gruidae	*Balearica pavonina*	Black (W. African) crowned crane	1.52
Gruidae	*Balearica regulorum*	Gray (E. African) crowned crane	1.44
Gruidae	*Balearica regulorum*	Gray (E. African) crowned crane	1.51
Gruidae	*Grus vipio*	Japanese white-naped crane	1.63
Otididae	*Ardeotis kori*	Kori bustard	1.42
Rallidae	*Gallinula chloropus*	Moorhen	1.53
Rallidae	*Gallirallus philippensis*	Buff-banded rail	1.59
Corvidae	*Aphelocoma coerulescens coerulescens*	Florida scrub jay	1.56
Corvidae	*Corvus corone*	Carrion crow	1.46
Corvidae	*Cyanocitta cristata*	Blue jay	1.43
Emberizidae	*Pipilo erythrophthalmus*	Rufous-sided towhee	1.55
Emberizidae	*Zonotrichia albicollis*	White-throated sparrow	1.37
Estrildidae	*Lonchura striata*	White-backed munia	1.42
Estrildidae	*Taeniopygia guttata*	Zebra finch	1.25
Eurylaimidae	*Calyptomena viridis*	Green broadbill	1.31
Icteridae	*Euphagus cyanocephalus*	Brewer's blackbird	1.46
Icteridae	*Passerina amoena*	Lazuli bunting	1.39
Oriolidae	*Oriolus chinensis*	Black-naped oriole	1.43
Ploceidae	*Passer domesticus*	House sparrow	1.51
Ploceidae	*Passer domesticus*	House sparrow	1.57
Sturnidae	*Streptocitta albicollis*	White-collared mynah	1.58
Sylviidae	*Regulus calendula*	Ruby-crowned kinglet	1.31
Turdidae	*Hylocichla mustelina*	Wood thrush	1.38
Turdidae	*Hylocichla mustelina*	Wood thrush	1.50
Capitonidae	*Psilopogon pyrolophus*	Fire-tufted barbet	1.69
Ramphastidae	*Megalaima mystacophanos*	Red-throated barbet	2.02
Ramphastidae	*Megalaima oorti*	Black-browed barbet	1.63
Psittacidae	*Agapornis personata*	Masked lovebird	1.35
Psittacidae	*Agapornis roseicollis*	Peach-faced lovebird	1.31
Psittacidae	*Amazona aestiva*	Blue-fronted parrot	1.61
Psittacidae	*Amazona albifrons*	White-fronted parrot	1.47
Psittacidae	*Amazona amazonica*	Amazon parrot	1.60
Psittacidae	*Amazona amazonica*	Orange-winged parrot	1.41
Psittacidae	*Amazona autumnalis*	Red-lored parrot	1.52
Psittacidae	*Amazona barbadensis*	Yellow-shouldered parrot	1.60
Psittacidae	*Amazona brasiliensis*	Red-tailed parrot	1.62
Psittacidae	*Amazona farinosa guatemalae*	Mealy parrot	1.58
Psittacidae	*Amazona leucocephala*	Cuban parrot	1.58
Psittacidae	*Amazona ochrocephala*	Yellow-crowned parrot	1.90
Psittacidae	*Amazona ochrocephala ochrocephala*	Yellow-headed parrot	1.70
Psittacidae	*Amazona ochrocephala oratrix*	Yellow-headed parrot	1.74
Psittacidae	*Amazona ochrocephala tresmariae*	Yellow-headed parrot	2.08
Psittacidae	*Amazona pretrei*	Red-spectacled parrot	1.41
Psittacidae	*Amazona tucumana*	Tucuman parrot	1.49
Psittacidae	*Amazona viridigenalis*	Red-crowned parrot	1.54
Psittacidae	*Anodorhynchus hyacinthinus*	Hyacinth macaw	1.35
Psittacidae	*Anodorhynchus hyacinthinus*	Hyacinth macaw	1.43
Psittacidae	*Ara ambiguus*	Great green macaw	1.37
Psittacidae	*Ara ararauna*	Blue-and-yellow macaw	1.40
Psittacidae	*Ara ararauna*	Blue-and-yellow macaw	1.45
Psittacidae	*Ara ararauna*	Blue-and-yellow macaw	1.48
Psittacidae	*Ara caninde*	Wagler's (Caninde) macaw	1.52
Psittacidae	*Ara chloroptera*	Green-winged macaw	1.48
Psittacidae	*Ara chloroptera*	Green-winged macaw	1.48
Psittacidae	*Ara macao*	Scarlet macaw	1.34
Psittacidae	*Ara macao*	Scarlet macaw	1.42
Psittacidae	*Ara militaris*	Military macaw	1.37
Psittacidae	*Ara militaris*	Military macaw	1.38
Psittacidae	*Ara rubrogenys*	Red-fronted macaw	1.39
Psittacidae	*Ara rubrogenys*	Red-fronted macaw	1.42
Psittacidae	*Aratinga solstitalis*	Sun conure	1.36

Psittacidae	*Cacatua alba*	White cockatoo	1.27
Psittacidae	*Cacatua galerita*	Sulpher-crested cockatoo	1.58
Psittacidae	*Cacatua moluccensis*	Salmon-crested cockatoo	1.52
Psittacidae	*Cacatua moluccensis*	Salmon-crested cockatoo	1.54
Psittacidae	*Cacatua sanguinea*	Little corella	1.71
Psittacidae	*Cacatua sulphurea abbotti*	Cockatoo	1.22
Psittacidae	*Cacatua sulphurea citrinocristata*	Citron-crested cockatoo	1.33
Psittacidae	*Cacatua sulphurea sulphurea*	Yellow-crested cockatoo	1.48
Psittacidae	*Cyanoliseus patagonus*	Burrowing parakeet	1.31
Psittacidae	*Deroptyus accitrinus*	Hawk-headed parrot	1.64
Psittacidae	*Eclectus roratus*	Eclectus parrot	1.59
Psittacidae	*Eclectus roratus*	Eclectus parrot	1.67
Psittacidae	*Enicognathus leptorhynchus*	Slender-billed parakeet	1.28
Psittacidae	*Eolophus roseicapilla*	Galah	1.24
Psittacidae	*Eolophus roseicapilla*	Galah	1.43
Psittacidae	*Lorius garrulus*	Chattering lory	1.36
Psittacidae	*Melopsittacus undulatus*	Budgerigar	1.33
Psittacidae	*Nymphicus hollandicus*	Cockateil	1.45
Psittacidae	*Nymphicus hollandicus*	Cockateil	1.46
Psittacidae	*Poicephalus senegalus*	Senegal parrot	1.37
Psittacidae	*Probosciger aterrimus*	Palm cockatoo	1.64
Psittacidae	*Psittacula krameri*	Ring-necked parakeet	1.37
Psittacidae	*Psittaculirostris edwardsii*	Edward's fig parrot	1.52
Psittacidae	*Psittacus erithacus*	Gray parrot	1.46
Psittacidae	*Psittacus erithacus*	Grey parrot	1.40
Psittacidae	*Trichoglossus haematodus*	Rainbow lory	1.33
Spheniscidae	*Spheniscus demersus*	Jackass penguin	1.63
Strigidae	*Athene noctua*	Little owl	1.67
Strigidae	*Bubo bubo*	Northern eagle owl	2.03
Strigidae	*Nyctea scandiaca*	Snowy owl	1.68
Strigidae	*Otus asio*	Eastern screech owl	1.47
Strigidae	*Otus asio*	Eastern screech owl	1.48
Strigidae	*Strix aluco*	Tawny owl	1.59
Strigidae	*Strix nebulosa*	Great grey owl	1.65
Tytonidae	*Tyto alba*	Common barn owl	1.73

Supplementary Data 3. Genome sampling statistics for 66 squamate species.

Illumina MiSeq shotgun sequencing samples

Species	Common name	Totat Combined Reads	Total Raw bp	Total Nuclear bp	GC content (%)
*Coleonyx elegans**	Yucatán banded gecko	3,479,532	869,883,000	661,966,988	45.64
Gekko gekko	Tokay gecko	2,074,340	518,585,000	399,222,362	46.57
Zonosaurus madagascariensis	Madagascar plated lizard	1,570,986	392,746,500	285,011,068	43.21
Platysaurus intermedius	Common flat lizard	3,498,226	1,084,450,060	713,972,746	44.22
Lepidophyma mayae	Mayan tropical Night Lizard	2,048,596	512,149,000	376,651,498	46.27
Lepidophyma flavimaculatum	Yellow-spotted night lizard	2,910,410	727,602,500	540,463,724	45.94
Plestiodon fasciatus	Five-lined skink	3,072,974	768,243,500	594,792,793	46.76
Tribolonotus gracilis	Red-eyed croc skink	2,604,866	807,508,460	579,616,674	46.07
Lamprolepis smaragdina	Emerald tree skink	2,808,668	870,687,080	376,649,200	43.9
Aspidoscelis scalaris	Plateau spotted whiptail	250,110	77,534,100	54,953,808	43.38
Proctoporus pachyurus	Tschudi's lightbulb lizard	1,504,862	466,507,220	333,261,942	40.2
*Bipes canaliculauts**	Four-toed worm lizard	5,322,166	1,330,541,500	1,031,694,072	43.31
Varanus exanthematicus	Savannah monitor	2,905,680	900,760,800	642,548,170	44.41
Abronia graminea	Green arboreal alligator lizard	1,344,546	336,136,500	263,697,926	45.36
Abronia matudai	Matuda's arboreal Alligator Lizard	1,583,430	395,857,500	305,932,955	45.76
*Ophisaurus attenuatus**	Slender glass lizard	3,307,234	823,808,500	619,395,214	45.41
Trioceros melleri	Meller's chameleon	2,343,340	726,435,400	504,624,342	44.18
Uromastyx geyri	Saharan spiny-tailed lizard	1,998,784	619,623,040	447,386,744	42.31
Dendragama boulengeri	Boulenger's tree agama	2,134,810	661,791,100	480,620,070	42.8
Lophocalotes ludekingi	Crested lizard	1,694,118	525,176,580	372,355,916	43.89
Gonocephalus grandis	Great anglehead lizard	2,037,934	631,759,540	325,284,517	43.75
Bronchocela jubata	Maned forest lizard	2,303,186	713,987,660	516,377,597	42.74
Calotes sp.		3,712,834	928,208,500	611,431,640	45.09
Sceloporus poinsettii	Crevice spiny lizard	2,523,062	630,765,500	403,601,134	42.95
Sceloporus teapensis	Teapa scaly lizard	2,446,190	611,547,500	465,440,852	42.91
Phrynosoma cornutum	Texas horned lizard	1,953,066	488,266,500	352,479,216	42.92
Crotaphytus collaris	Eastern collared lizard	2,839,282	709,820,500	560,017,110	42.72
Norops humillis	Ground anole lizard	2,365,870	591,467,500	420,458,457	45.04
Oplurus quadrimaculatus	Madagascar spiny-tailed iguana	3,392,902	848,225,500	658,997,359	40.57
Eryx jaculus	Javelin sand boa	2,310,286	577,571,500	410,158,507	42.9
Acrochordus granulatus	Little file snake	2,698,706	836,598,860	606,704,306	39.26
Pareas carinata	Keeled slug-eating snake	1,553,330	388,332,500	300,894,252	43.35
*Sistrurus catenatus**	Massasauga	3,800,896	950,224,000	731,856,378	42.09
Bothriechis schlegelii	Eyelash viper	1,463,136	453,572,160	305,424,903	40.57
Cerrophidion godmani	Godman's montane pitviper	2,154,462	538,615,500	377,329,574	42.26
Bothrops asper	Terciopelo	4,029,527	1,020,000,000	673,181,829	44.9
*Tropidolaemus subannulatus**	North Philippine temple pitviper	5,315,414	1,328,853,500	914,955,964	42.4
*Cerastes cerastes**	Horned desert viper	5,176,826	1,294,206,500	921,873,870	42.26
Ahaetulla prasina	Asian vine snake	3,487,320	871,830,000	653,137,137	42.17
Thelotornis kirtlandii	Twig snake	2,416,080	604,020,000	456,106,084	43.21
Coluber constrictor	Black racer	2,382,600	595,650,000	323,175,136	41.57
Pantherophis emoryi	Great plains rat snake	2,860,290	715,072,500	496,928,698	44.04
Coniophanes piceivittis	Black striped snake	1,905,662	476,415,500	346,481,831	41.88
Coniophanes fissidens	Yellowbelly snake	1,431,564	357,891,000	259,876,691	47.81
Cerberus rhynchops	Dog-faced water snake	2,368,578	734,259,180	544,544,296	41.24

* = merging of two independent replicates

454 shotgun sequencing samples	Common name	Total number of reads	Total Raw bp	Total Nuclear bp	GC content (%)
Leptotyphlops dulcis	Texas blind snake	71,058	11,828,885	11,823,143	43.18
Typhlops reticulatus	Reticulate worm snake	50,087	6,741,155	6,720,475	46.13
Anilius scytale	American pipe snake	50,319	7,542,192	7,508,176	43.23
Casarea dussumieri	Round island boa	470,682	76,243,119	76,218,678	43.4
Loxocemus bicolor	Mexican burrowing python	40,583	6,172,347	6,163,619	42.59
Crotalus atrox	Western diamondback rattlesnake	63,094	19,098,306	18,965,550	38.79
Agkistrodon contortryx	Copperhead	280,303	60,344,580	60,175,941	42.49
Sibon nebulatus	Slug-eating snake	43,542	12,772,185	10,989,600	41.01
Micrurus fulvius	Eastern coral snake	26,831	7,735,311	6,769,294	39.35

Whole assembled genomes	Common name	Number of Scaffolds	Total bp	Total unambiguous nt	GC content (%)
Gekko japonicus	Schlegel's Japanese gecko	335,469	2,490,274,461	2,402,030,469	45.47
Ophisaurus gracilis	Burmese glass lizard	6,715	1,781,357,942	1,729,274,821	43.71
Pogona vitticeps	Central bearded dragon	100,000	1,659,313,787	1,592,670,006	41.84
Anolis carolinensis	Green anole lizard	6,457	1,799,143,587	1,701,422,805	40.32
Boa constrictor	Boa constrictor	111,002	1,387,463,918	1,387,241,914	40.26
Python molurus bivittatus	Burmese python	39,112	1,435,034,535	1,385,275,938	39.61
Crotalus mitchellii	Speckled rattlesnake	478,598	1,139,346,324	1,129,318,242	38.65
Crotalus viridis	Prairie rattlesnake	56,243	1,213,434,727	1,191,577,356	38.61
Deinagkistrodon acutus	Five-pacer viper	162,571	1,506,308,921	1,417,468,057	39.82
Pantherophis guttatus	Corn snake	883,920	1,404,220,341	1,358,371,200	39.58
Thamnophis sirtalis	Common garter snake	7,930	1,424,897,867	1,122,631,552	39.92
Ophiophagus hannah	King cobra	296,399	1,594,076,454	1,379,208,606	39.51

Supplementary Data 4. Repeat element landscape composition for 66 sampled squamate genomes estimated using RepeatMasker.

	Total Repeats	Total SSRs	Total TEs	SINEs	LINEs					PLEs	DIRS	LTR	DNA transposons			Unclassified
Species					CR1-L3	L2	BovB	L1	Others				hAT	Tc1	Others	
Gekkota																
Coleonyx elegans	24.37	0.78	23.67	3.48	4.49	4.66	1.49	1.6	0.33	0.74	0.73	1.76	1.04	0.41	0.47	2.31
Gekko gekko	45.24	0.42	44.8	5.95	7.62	4.15	2.49	2.79	0.31	1.18	0.99	5.19	1.71	0.51	0.97	10.62
Gekko japonicus	42.12	1.07	41.27	6.9	6.38	3.48	2.29	2.62	0.28	0.69	1.98	3.31	1.76	0.52	0.69	10.53
Scincoidea																
Zonosaurus madagascariensis	36.66	1.98	34.3	2.1	4.92	4.86	2.18	0.98	0.38	0.79	0.24	1.74	4.61	0.72	2.36	8.71
Platysaurus intermedius	40.02	1.81	37.8	2.73	6.24	5.42	2.23	1.33	0.65	0.92	0.23	1.28	6.09	1.44	2.6	6.76
Lepidophyma flavimaculatum	44.2	0.92	44.01	1.92	7.55	6.22	1.38	1.42	0.48	2.04	0.92	4.11	3.97	1.72	2.72	9.55
Lepidophyma mayae	44.87	0.89	43.37	1.84	7.3	5.98	1.26	1.4	0.42	2.22	0.87	4.05	3.75	1.54	2.8	9.31
Lamprolepis smaragdina	35.03	1.19	34.66	3.24	4.02	3.16	2.29	0.79	0.79	0.25	0.42	2.17	5.77	1.62	3.31	4.36
Tribolonotus gracilis	36.97	2.27	34.3	1.73	4.8	2.8	0.4	1.39	0.48	0.26	0.3	2.05	9.97	1.71	2.96	5.55
Plestiodon fasciatus	36.15	0.17	34.74	2.55	3.98	1.14	1.15	0.68	0.35	0.31	0.17	1.92	13.58	1.71	2.35	7.45
Lacertoidea																
Aspidoscelis scalaris	36.72	3.15	32.84	2.57	5.31	3.39	0.82	1.17	1.43	0.26	0.33	1.84	4.28	1.07	1.4	8.97
Proctoporus pachyurus	33.31	3.27	29.41	3.71	4.31	2.85	1.08	1.25	1.16	0.02	0.58	1.5	2.52	1.41	2.31	6.74
Bipes canaliculauts	41.83	0.94	40.5	4.47	3.54	4.87	10.34	1.07	0.35	0.08	0.13	1.77	3.06	1.94	1.59	7.34
Anguimorpha																
Abronia graminea	49.82	1.57	46.79	2.82	8.13	7.16	1.58	1.78	0.81	1.7	0.81	5.71	3.8	0.52	6.11	6.85
Abronia matudai	48.99	1.55	46.68	2.63	8.02	7.45	1.42	1.61	0.77	1.53	0.83	5.83	3.85	0.67	5.96	6.57
Ophisaurus attenuatus	48.9	1.11	34.27	3.05	6.63	4.77	5.63	1.76	0.79	0.61	0.22	2.02	3.4	0.79	2.04	6.41
Ophisaurus gracilis	44.78	1.89	42.26	2.21	5.59	6.54	0.43	1.16	1.31	0.77	0.54	6.49	5.66	0.14	5.59	6.34
Varanus exanthematicus	35.68	1.52	47.63	2.57	8.06	6.81	1.53	0.89	0.42	2.12	0.82	6.38	3.86	0.58	5.93	4.28
Iguania																
Bronchocela jubata	41.15	1.22	30.22	3.53	4.06	3.31	2.51	1.58	1.21	0.51	1.75	2	2.54	0.81	1.88	6
Gonocephalus grandis	41.56	1.11	31.57	3.38	2.63	2.73	5.32	1.29	1.08	0.62	1.98	2.23	3.08	1.79	1.55	5.38
Calotes sp.	41.98	1.55	39.76	4.55	1.9	3.59	4.33	1.32	1.07	2.25	0.83	4.2	5.09	1.4	1.95	5.19
Dendragama boulengeri	39.7	2.46	37.12	2.56	4.99	2.66	1.94	1.76	1.2	0.27	0.85	2.67	6.75	3.09	2.79	5.66
Lophocalotes ludekingi	42.94	2.6	39.66	2.44	5.73	2.71	1.79	2.03	1.12	0.73	0.79	3.74	6.94	2.79	2.76	6.2
Pogona vitticeps	32.4	2.57	39.05	2.49	5.34	5.28	2.04	0.86	1.32	0.67	0.83	3.32	5.74	2.98	2.66	4.32
Uromastyx geyri	31.35	2.41	38.87	2.55	4.81	5.14	1.91	1.31	0.98	0.57	0.83	3.34	5.74	3.68	2.73	5.33
Trioceros melleri	41.52	2.76	38.19	2.49	5.58	2.65	1.79	1.46	2.86	0.45	1.21	3.02	7.36	2.61	2.52	5.3
Crotaphytus collaris	37.3	2.1	42.95	3.89	5.14	5.77	2.7	1.86	0.36	1.73	2.01	2.72	3.82	2.63	4.06	2.65
Phrynosoma cornutum	49.56	1.89	43.39	4.05	5.42	5.99	2.85	1.71	0.46	1.47	2.03	2.84	3.83	2.7	4.1	6.98
Sceloporus poinsettii	45.49	1.81	35.41	2.66	6.31	3.94	4.55	1.61	0.54	1.26	1.18	2.37	2.94	2.14	2.79	6.1
Sceloporus teapensis	45.36	1.33	48.19	4.16	5.35	8.11	4.05	1.69	0.63	0.71	1.79	3.83	4.07	2.27	4.88	6.29
Anolis carolinensis	51.4	1.74	43.55	1.4	9.91	2.45	0.87	2	2.43	0.25	2.58	3.67	2.53	1.25	4.15	5.7
Norops humillis	57.6	3.32	53.55	3.74	3.97	9.24	3.32	2.15	1.61	0.86	0.75	5.89	5.58	2.84	6.43	9.17
Oplurus quadrimaculatus	45.3	2.26	43.55	5.13	3.72	4.43	2.46	3.73	0.83	0.46	2.08	5.83	6.1	5.17	6.33	10.22

	Total Repeats	Total SSRs	Total TEs	SINEs	LINEs					PLEs	DIRS	LTR	DNA transposons			Unclassified
Species					CR1-L3	L2	BovB	L1	Others				hAT	Tc1	Others	
Non colubroid snakes																
Leptotyphlops dulcis	50.4	1.18	48.67	2.42	10.41	7.27	0.73	0.51	2.84	1.52	0. 10	3.35	3.2	2.31	2.31	11.7
Typhlops reticulatus	46.77	1.84	44.33	1.43	7.82	2.22	0.96	0.56	2.13	0.71	0.01	3.58	3.38	2.03	1.54	16.13
Anilius scytale	40.8	1.46	39.11	2.12	3.87	4.17	3.03	1.64	1.99	2.07	0.04	3.19	2.53	3.56	2.07	8.82
Boa constrictor	34.9	2.34	32.21	2.58	2.91	4.06	3.88	2.43	1.45	1.05	0.05	2.5	1.91	2.07	1.83	5.43
Eryx jaculus	44.44	2.1	42.11	3.53	2.74	6.29	4.93	2.35	2.62	1.13	0.19	3.61	2.8	3.19	2.27	6.41
Loxocemus bicolor	41.28	1.24	40.14	2.2	2.5	3.77	5.23	1.79	3.14	2.53	0.05	2.53	1.48	2.82	1.86	9.93
Python molurus	31.02	1.99	28.7	1.6	2.1	3.4	3.05	2.9	1.71	0.93	0.05	1.9	1.56	2.11	1.31	5.98
Casarea dussumieri	42.57	11.46	30.87	2.62	2.74	3.22	3.93	1.31	2.44	0.99	0.2	2.75	2.08	4.12	1.92	2.55
Acrochordus granulatus	43.75	1.13	42.29	3.22	1.44	4.01	12.42	2	2.62	1.99	0.06	1.81	2.69	4.36	1.47	3.38
Colubroidea																
Pareas carinata	53.37	4.08	48.19	2.15	3.04	7.87	1.33	3.77	3.16	1.11	0.18	5.21	2.65	3.45	1.47	12.84
Agkistrodon contortrix	55.25	4.8	49.64	2.28	9.81	1.69	2.38	3.01	2.19	2.3	2.21	7.43	7.45	4.27	2.59	1.89
Crotalus atrox	47.31	2.55	44.65	2.13	7.11	1.78	3.43	4.47	3.63	2.09	1.19	6.28	4.68	3.86	2.12	1.67
Crotalus mitchellii	35.27	2.03	32.98	2.42	4.41	1.92	2.13	2	1.63	1.61	0.86	3.44	3.78	4.22	1.87	2.5
Crotalus viridis	40.86	2.26	38.1	1.83	6.47	1.08	2.51	2.45	1.71	1.49	1.11	4.25	4.44	4.22	1.26	5.03
Sistrurus catenatus	48.09	4.29	42.91	2.23	8.61	1.9	3.05	1.71	3.88	1.48	1.21	5.19	4.07	3.54	1.96	3.9
Bothriechis schlegelli	51.1	4.27	46.31	2.34	7.05	2.88	2.61	4.25	2.17	2.41	1.81	6.92	5.28	3.86	2.52	1.98
Bothrops asper	60.74	6.44	53.09	2.54	7.66	6.51	2.35	3.47	2.02	2.07	2.19	5.99	7.76	4.34	2.79	3.48
Cerrophidion godmani	53.47	3.67	49.4	3.17	9.09	2.97	3.23	3.01	2.57	2.23	2	6.62	5.92	4.26	2.45	1.84
Tropidolaemus subannulatus	48.15	3.54	44.03	2.01	7.91	6.34	3.24	1.68	4.16	1.14	1.73	4.05	2.96	3.6	1.94	3.26
Deinagkistrodon acutus	48.86	4.33	43.51	2.13	8.18	2.26	2.52	2.31	3.64	1.53	1.95	4.34	5.02	4.64	1.54	3.44
Cerastes cerastes	54.43	2.54	51.13	5.02	3.86	1.86	6.6	4.09	8.33	1.65	0.34	6.71	4.23	4.53	2.07	1.88
Cerberus rhynchops	48.4	3.32	44.04	2.41	2.48	4.53	1.97	3.26	7.5	2.43	1.01	6.06	4	3.85	1.96	2.29
Ahaetulla prasina	58.38	6.5	51.39	3.97	5.1	3.65	1.93	3.42	9.97	1.73	0.51	4.54	5.08	4.25	2.11	4.94
Coluber constrictor	53.34	4.91	47.91	4.49	3.37	2.68	1.98	2.99	7	1.61	0.88	4.28	5.8	5.5	2.02	5.16
Pantherophis emoryi	58.28	4.75	52.67	4.8	4.97	7.4	2.11	3.39	2.88	1.64	1.04	5.61	6.49	5.43	2.36	4.47
Pantherophis guttatus	43.38	2.6	40.04	3.53	4.43	2.75	1.97	2.54	2.35	1.66	0.74	3.38	5.31	5.1	2.02	3.99
Thelotornis kirtlandii	57.19	4.6	51.97	5.89	2.64	5.17	1.7	3.08	8.24	1.49	1.06	4.37	5.91	4.45	2.1	6.01
Coniophanes fissidens	73.02	14.16	56.34	4.26	6.21	7.16	3.02	1.35	9.26	1.79	1.37	4.45	7.42	3.56	2.71	4.8
Coniophanes piceivittis	59.35	7.92	50.19	3.66	5.78	3.81	2.57	2.47	9.92	1.26	0.9	3.37	5.43	3.82	2.72	4.56
Sibon nebulatus	57.57	3.1	53.34	3.55	9.62	4.53	4.53	2.59	6.18	1.33	0.68	4.62	5.87	3.63	2.65	3.4
Thamnophis sirtalis	39.01	2.83	35.69	3.9	3.22	1.01	1.49	1.44	3.19	1.11	0.87	2.6	4.31	4.53	1.64	6.37
Micrurus fulvius	48.17	3.16	44.04	3.44	3.89	4.31	1.63	3.11	6.81	1.74	0.83	4.28	3.8	4	1.86	4.09
Ophiophagus hannah	41.32	2.99	37.87	3.09	5.05	2.78	1.46	2.43	2.31	1.46	0.6	2.79	3.2	4.5	1.14	6.74

Supplementary Data 5. Microsatellite density estimates across 66 squamate species.
Loci/Mbp density estimates for each species and averages for each major squamate lineages are reported for 2-6mer SSR loci and for the total microsatellite content as estimated in PAL-finder.

Family	Species	Loci/Mbp					
		2mer	3mer	4mer	5mer	6mer	Total
Gekkota	*Coleonyx elegans*	60.74	74.26	101.78	16.99	8.03	261.81
	Gekko gecko	60.05	126.30	124.66	24.31	14.24	349.55
	Gekko japonicus	90.57	165.78	177.66	31.15	14.43	479.59
Scincoidea	*Zonosaurus madagascariensis*	127.90	119.72	237.00	53.94	13.69	552.26
	Platysaurus intermedius	117.33	118.62	240.50	44.73	24.33	545.51
	Lepidophyma mayae	73.98	93.79	158.52	17.91	8.42	352.62
	Lepidophyma flavimaculatum	67.68	94.62	147.62	20.85	14.73	345.51
	Plestiodon fasciatus	132.02	114.51	140.51	20.15	15.19	422.38
	Tribolonotus gracilis	92.73	68.03	410.13	17.38	6.00	594.27
	Lamprolepis smaragdina	90.93	58.62	129.30	15.96	5.77	300.57
Lacertoidea	*Aspidoscelis scalaris*	64.24	129.55	219.04	42.05	27.50	482.37
	Proctoporus pachyurus	131.82	289.99	420.64	41.18	39.82	923.45
	Bipes canaliculauts	72.19	153.76	160.87	21.77	9.90	418.49
Anguimorpha	*Varanus exanthematicus*	62.52	82.56	202.55	38.79	13.04	399.46
	Abronia graminea	153.52	132.45	193.16	32.89	13.55	525.57
	Abronia matudai	180.73	139.75	204.96	34.08	15.59	575.10
	Ophisaurus gracilis	226.01	217.15	273.37	49.39	14.83	780.74
	Ophisaurus attenuatus	158.55	139.50	196.17	34.13	14.52	542.86
Iguania	*Trioceros melleri*	96.03	133.95	207.10	44.01	16.38	497.47
	Uromastyx geyri	54.62	75.74	188.27	39.98	24.67	383.27
	Pogona vitticeps	128.42	118.32	262.22	43.20	16.20	568.36
	Dendragama boulengeri	170.73	133.56	323.78	68.98	27.46	724.50
	Lophocalotes ludekingi	247.77	136.06	311.22	66.56	38.82	800.42
	Gonocephalus grandis	159.05	149.36	331.85	92.62	26.01	758.90
	Bronchocela jubata	142.27	142.40	323.73	74.97	40.73	724.10
	Calotes sp.	163.70	135.20	282.73	69.74	72.78	724.15
	Sceloporus poinsettii	110.58	152.23	146.72	26.22	15.50	451.25
	Sceloporus teapensis	116.24	176.74	168.17	30.49	43.60	535.25
	Phrynosoma cornutum	91.00	151.78	135.07	37.53	25.55	440.93
	Crotaphytus collaris	106.82	108.04	142.68	27.81	52.23	437.58
	Norops humillis	120.45	147.18	390.59	21.78	106.67	786.66
	Oplurus quadrimaculatus	84.96	95.34	139.71	34.98	80.34	435.32
	Anolis carolinensis	98.73	262.53	198.52	28.87	12.55	601.20
Lizards	*Avg*	116.81	134.47	220.93	38.35	26.46	537.01
	SD	47.18	49.47	86.64	18.73	22.88	163.65
Scolecophidia	*Typhlops reticulatus*	87.04	123.61	202.43	89.22	19.79	522.10
	Leptotyphlops dulcis	91.95	78.11	209.03	38.49	9.29	426.87
Aniliidae	*Anilius scytale*	86.75	72.29	173.65	44.84	15.48	393.01
Bolyeriidae	*Casarea dussumieri*	69.15	83.15	158.52	32.66	20.20	363.69
Boidae	*Boa constrictor*	99.44	101.90	273.48	60.20	23.37	558.38
	Eryx jaculus	90.40	86.60	263.67	60.34	27.60	528.61
Pythonidae	*Loxocemus bicolor*	97.18	88.24	167.35	29.47	9.82	392.06
	Python molurus	104.56	98.35	265.60	56.50	20.26	545.28
Acrochordidae	*Acrochordus granulatus*	46.49	54.10	170.69	27.78	4.45	303.51
Non Colubroid Snakes	*Avg*	85.89	87.37	209.38	48.83	16.70	448.17
	SD	17.86	19.63	46.63	19.86	7.51	92.34
Pareatidae	*Pareas carinata*	119.13	82.34	250.90	56.98	254.72	764.08
Viperidae	*Crotalus atrox*	172.99	182.16	468.04	229.77	29.53	1,082.49
	Crotalus mitchellii	119.69	145.90	305.20	84.33	11.98	667.10
	Crotalus viridis	120.93	133.34	274.90	77.99	14.60	621.75
	Sistrurus catenatus	127.79	142.26	304.50	89.10	15.62	679.27
	Agkistrodon contortrix	114.42	119.77	290.91	83.51	20.86	629.47
	Bothriechis schlegelii	220.69	120.22	277.69	94.97	20.99	734.56
	Cerrophidion godmani	197.06	208.69	566.12	150.22	50.86	1,172.96
	Bothrops asper	145.99	129.41	300.70	92.16	24.31	692.59
	Tropidolaemus subannulatus	111.60	127.92	314.37	112.63	24.94	691.46
	Deinagkistrodon acutus	164.46	168.82	337.81	121.22	27.88	820.19
	Cerastes cerastes	113.28	116.80	221.45	83.48	79.64	614.65
Homalopsidae	*Cerberus rhynchops*	111.98	83.22	293.11	103.16	30.87	622.35
Colubridae	*Ahaetulla prasina*	109.48	93.44	605.35	299.73	40.96	1,148.96
	Thelotornis kirtlandii	104.86	87.67	315.11	114.07	34.92	656.63
	Coluber constrictor	117.05	102.27	345.94	118.46	95.00	778.71
	Pantherophis emoryi	119.35	101.06	321.05	98.40	28.27	668.14
	Pantherophis guttatus	262.82	97.05	457.04	96.51	33.56	946.96
	Coniophanes piceivittis	186.74	111.12	1,487.92	302.70	756.33	2,844.81

	Coniophanes fissidens	156.51	121.38	1,034.89	242.55	59.14	1,614.47
	Sibon nebulatus	141.50	158.88	342.60	126.30	38.85	808.13
	Thamnophis sirtalis	157.18	104.55	346.69	89.36	44.19	741.96
Elapidae	*Micrurus fulvius*	117.29	97.65	348.49	100.90	20.98	685.30
	Ophiophagus hannah	148.52	111.18	388.35	134.71	39.29	822.06
Colubroid Snakes	*Avg*	144.22	122.80	424.96	129.30	74.93	896.21
	SD	40.07	32.01	281.44	68.07	153.11	476.79

bp/Mbp density estimates for each species and averages for each major squamate lineages are reported for 2-6mer SSR loci and for the total microsatellite content as estimated in PAL-finder.

Family	Species	bp/Mbp					
		2mer	3mer	4mer	5mer	6mer	Total
Gekkota	*Coleonyx elegans*	1,114.14	1,050.10	1,384.51	268.86	262.07	4,079.68
	Gekko gecko	1,019.60	2,096.10	2,037.50	445.64	391.78	5,990.62
	Gekko japonicus	1,605.92	2,860.14	2,851.52	528.24	293.13	8,138.95
Scincoidea	*Zonosaurus madagascariensis*	2,730.18	2,205.11	4,343.97	1,104.45	304.26	10,687.97
	Platysaurus intermedius	2,413.34	2,115.69	4,118.43	838.48	571.78	10,057.72
	Lepidophyma mayae	1,324.20	1,436.78	2,862.16	285.00	219.64	6,127.78
	Lepidophyma flavimaculatum	1,181.99	1,425.63	2,584.84	335.54	468.32	5,996.32
	Plestiodon fasciatus	3,106.69	1,798.60	2,062.05	319.51	408.29	7,695.14
	Tribolonotus gracilis	1,794.01	980.24	6,548.24	275.37	148.33	9,746.19
	Lamprolepis smaragdina	2,037.48	870.48	2,126.13	255.26	134.49	5,423.84
Lacertoidea	*Aspidoscelis scalaris*	1,064.75	2,435.32	4,438.56	824.42	682.94	9,445.99
	Proctoporus pachyurus	2,379.77	5,419.57	8,125.43	666.04	864.74	17,455.55
	Bipes canaliculauts	1,125.34	2,938.42	2,128.01	343.39	223.50	6,758.66
Anguimorpha	*Varanus exanthematicus*	1,170.44	1,195.59	3,399.05	786.41	324.79	6,876.28
	Abronia graminea	3,277.05	2,440.76	2,846.73	524.62	366.19	9,455.35
	Abronia matudai	4,100.37	2,657.95	3,053.95	546.49	442.53	10,801.29
	Ophisaurus gracilis	5,044.67	4,183.94	4,204.94	819.03	294.90	14,547.48
	Ophisaurus attenuatus	3,389.86	2,614.25	2,911.14	547.18	410.67	9,873.10
Iguania	*Trioceros melleri*	1,678.75	2,147.09	3,110.97	735.92	356.77	8,029.50
	Uromastyx geyri	927.47	1,242.06	3,908.75	735.26	500.60	7,314.14
	Pogona vitticeps	2,658.91	2,267.70	6,288.45	791.74	371.95	12,378.75
	Dendragama boulengeri	4,004.01	3,248.34	8,535.32	1,319.15	783.04	17,889.86
	Lophocalotes ludekingi	5,005.01	3,013.50	7,487.30	1,192.42	1,067.49	17,765.72
	Gonocephalus grandis	3,355.00	3,461.93	6,999.16	1,652.20	632.88	16,101.17
	Bronchocela jubata	2,870.52	3,513.88	8,099.10	1,481.35	1,262.89	17,227.74
	Calotes sp.	3,375.49	2,733.65	6,037.49	1,283.08	2,485.13	15,914.84
	Sceloporus poinsettii	2,073.60	2,595.75	2,331.35	427.40	385.69	7,813.79
	Sceloporus teapensis	2,166.43	2,992.91	2,766.99	506.99	1,148.50	9,581.82
	Phrynosoma cornutum	1,549.70	2,528.93	2,105.26	655.68	600.78	7,440.35
	Crotaphytus collaris	1,942.55	1,681.50	2,142.32	478.61	1,284.38	7,529.36
	Norops humillis	2,694.40	2,421.44	5,824.09	344.45	3,787.82	15,072.20
	Oplurus quadrimaculatus	1,626.74	1,663.24	2,448.39	684.41	1,936.76	8,359.54
	Anolis carolinensis	2,106.54	6,417.96	3,234.03	485.52	246.52	12,490.57
Lizards	*Avg*	2,361.06	2,504.68	4,040.79	681.46	717.08	10,305.07
	SD	1,125.08	1,189.03	2,092.09	368.77	755.53	4,043.20
Scolecophidia	*Typhlops reticulatus*	3,012.85	2,190.20	4,055.37	1,836.49	434.72	11,529.63
	Leptotyphlops dulcis	2,434.79	1,307.36	3,517.76	650.31	181.44	8,091.66
Aniliidae	*Anilius scytale*	3,636.21	1,260.51	3,081.53	949.29	425.87	9,353.41
Bolyeriidae	*Casarea dussumieri*	1,724.03	1,291.51	2,913.26	613.81	471.26	7,013.87
Boidae	*Boa constrictor*	1,796.65	1,734.44	4,997.37	1,505.89	599.08	10,633.43
	Eryx jaculus	1,699.63	1,602.77	5,004.01	1,336.99	702.94	10,346.34
Pythonidae	*Loxocemus bicolor*	2,444.30	1,273.02	2,771.64	598.18	236.82	7,323.96
	Python molurus	1,943.90	1,898.98	5,732.26	1,443.77	576.67	11,595.58
Acrochordidae	*Acrochordus granulatus*	676.54	710.77	2,719.06	479.13	111.16	4,696.66
Non Colubroid Snakes	*Avg*	2,152.10	1,474.40	3,865.81	1,045.98	415.55	8,953.84
	SD	854.67	434.10	1,132.00	494.20	201.62	2,340.41
Pareatidae	*Pareas carinata*	2,177.37	1,677.42	4,984.17	1,292.58	8,450.71	18,582.25
Viperidae	*Crotalus atrox*	3,624.36	3,976.01	11,194.38	7,920.39	778.91	27,494.04
	Crotalus mitchellii	2,993.67	3,238.20	5,129.87	2,015.72	251.14	13,628.59
	Crotalus viridis	2,127.67	2,031.70	4,049.18	1,355.21	278.24	9,842.00
	Sistrurus catenatus	2,433.35	2,416.90	5,362.27	1,914.18	306.33	12,433.02
	Agkistrodon contortrix	2,261.05	2,256.41	6,496.20	2,485.73	524.34	14,023.73
	Bothriechis schlegelii	3,988.82	2,291.48	5,492.52	2,507.65	451.06	14,731.54
	Cerrophidion godmani	3,926.70	4,362.98	12,309.03	4,088.25	1,292.33	25,979.28
	Bothrops asper	3,026.83	2,307.28	6,090.44	2,400.94	644.69	14,470.18
	Tropidolaemus subannulatus	2,101.76	2,501.76	6,089.14	2,756.60	634.61	14,083.88

	Deinagkistrodon acutus	2,992.90	3,850.58	6,977.02	3,908.50	958.95	18,687.95
	Cerastes cerastes	2,094.13	2,049.97	4,480.10	2,118.70	2,247.12	12,990.02
Homalopsidae	*Cerberus rhynchops*	2,335.53	1,554.65	6,152.50	2,804.99	674.45	13,522.11
Colubridae	*Ahaetulla prasina*	2,014.28	1,604.15	15,379.75	8,330.12	966.61	28,294.90
	Thelotornis kirtlandii	2,012.97	1,431.24	6,421.86	3,512.91	1,020.67	14,399.64
	Coluber constrictor	2,110.81	1,655.42	6,382.96	3,110.94	2,865.28	16,125.41
	Pantherophis emoryi	2,125.79	1,609.56	5,484.17	2,125.89	597.30	11,942.71
	Pantherophis guttatus	4,729.97	1,617.06	11,192.23	2,368.74	885.64	20,793.64
	Coniophanes piceivittis	3,639.39	1,847.73	33,660.57	6,550.23	22,239.56	67,937.48
	Coniophanes fissidens	3,025.58	2,331.15	23,568.76	7,268.74	1,694.86	37,889.08
	Sibon nebulatus	3,851.10	2,965.17	6,115.96	2,902.29	818.41	16,652.93
	Thamnophis sirtalis	2,501.36	1,348.27	6,718.71	1,624.80	1,189.20	13,382.34
Elapidae	*Micrurus fulvius*	2,784.34	1,569.29	7,335.48	2,000.21	495.47	14,184.79
	Ophiophagus hannah	2,655.03	1,771.45	6,195.46	2,570.23	837.63	14,029.79
Colubroid Snakes	*Avg*	2,813.95	2,261.08	8,885.95	3,330.61	2,129.31	19,420.89
	SD	774.93	844.72	6,815.44	2,052.19	4,590.44	12,215.45

Supplementary Data 6. Statistics of the microsatellite landscape across lineages of squamate reptiles.

Bp/Mbp microsatellite density statistics

Taxonomic Group	2mer bp/Mbp	3mer bp/Mbp	4mer bp/Mbp	5mer bp/Mbp	6mer bp/Mbp	Tot SSR
Average						
Lizards	2361.06	2504.68	4040.79	681.46	717.08	10305.07
Non Colubroid snakes	2152.10	1474.40	3865.81	1045.98	415.55	8953.84
Colubroid snakes	2813.95	2261.08	8885.95	3330.61	2129.31	19420.89
Standard Deviation						
Lizards	1125.08	1189.04	2092.09	368.77	755.53	4043.20
Non Colubroid snakes	854.67	434.10	1132.00	494.20	201.62	2340.41
Colubroid snakes	774.93	844.72	6815.44	2052.19	4590.44	12215.45
Fold Difference						
Colubroidea-Lizard	1.19	0.90	2.20	4.89	2.97	1.88
Colubroidea -Non Colubroid snakes	1.31	1.53	2.30	3.18	5.12	2.17
Lizard-Non Colubroid snakes	1.10	1.70	1.05	0.65	1.73	1.15
Colubroidea-other	1.21	0.99	2.22	4.38	3.26	1.94
Kruskal–Wallis test						
chi-squared	5.57	9.64	22.52	44.68	11.16	25.23
p-val	0.062	0.008	1.29E-05	1.99E-10	0.004	3.33E-06

Dunn test for multiple comparisons (p-values adjusted with the Benjamini-Hochberg method)

	2mer	Z	P.unadj	**P.adj**	**p-val**
	Colubroidea - Lizard	2.12	0.03	0.102	
	Colubroidea - Non colubroid snakes	1.79	0.07	0.109	
	Lizard - Non colubroid snakes	0.35	0.72	0.724	
Kruskal–Wallis test	Colubroidea - Other squamates				0.019
	3mer				
	Colubroidea - Lizard	-0.80	0.42	0.421	
	Colubroidea - Non colubroid snakes	2.43	0.02	0.023	
	Lizard - Non colubroid snakes	3.10	0.00	0.006	
Kruskal–Wallis test	Colubroidea - Other squamates				0.894
	4mer				
	Colubroidea - Lizard	4.51	6.56E-06	1.97E-05	
	Colubroidea - Non colubroid snakes	3.15	0.002	0.002	
	Lizard - Non colubroid snakes	0.06	0.950	0.950	
Kruskal–Wallis test	Colubroidea - Other squamates				2.08E-06
	5mer				
	Colubroidea - Lizard	6.66	2.80E-11	8.41E-11	
	Colubroidea - Non colubroid snakes	3.20	0.001	0.002	
	Lizard - Non colubroid snakes	-1.42	0.156	0.156	
Kruskal–Wallis test	Colubroidea - Other squamates				6.51E-11
	6mer				
	Colubroidea - Lizard	2.79	0.005	0.008	
	Colubroidea - Non colubroid snakes	2.80	0.005	0.015	
	Lizard - Non colubroid snakes	0.92	0.359	0.359	
Kruskal–Wallis test	Colubroidea - Other squamates				0.001
	Tot SSR				
	Colubroidea - Lizard	4.49	7.13E-06	2.14E-05	
	Colubroidea - Non colubroid snakes	3.85	1.18E-04	1.77E-04	
	Lizard - Non colubroid snakes	0.80	0.424	0.424	
Kruskal–Wallis test	Colubroidea - Other squamates				7.10E-07

Loci/Mbp microsatellite density statistics

Taxonomic Group	2mer loci/Mbp	3mer loci/Mbp	4mer loci/Mbp	5mer loci/Mbp	6mer loci/Mbp	Tot SSR
Average						
Lizards	116.81	134.47	220.93	38.34	26.46	537.01
Non Colubroid snakes	85.88	87.37	209.38	48.83	16.69	448.17
Colubroid snakes	144.22	122.80	424.96	129.30	74.93	896.21
Standard Deviation						
Lizards	47.18	49.47	86.64	18.73	22.88	163.65
Non Colubroid snakes	17.86	19.63	46.63	19.86	7.51	92.35
Colubroid snakes	40.07	32.01	281.44	68.07	153.11	476.79
Fold Difference						
Colubroidea-Lizard	1.23	0.91	1.92	3.37	2.83	1.67
Colubroidea -Non Colubroid snakes	1.68	1.41	2.03	2.65	4.49	2.00
Lizard-Non Colubroid snakes	1.36	1.54	1.06	0.79	1.58	1.20
Colubroidea-other	1.31	0.99	1.95	3.19	3.08	1.73
Kruskal–Wallis test						
chi-squared	16.442	11.155	26.949	42.67	13.465	29.65
p-val	2.69E-4	0.004	1.41E-06	5.43E-10	0.001	3.65E-07

Dunn test for multiple comparisons (p-values adjusted with the Benjamini-Hochberg method)

	2mer	Z	P.unadj	**P.adj**	**p-val**
	Colubroidea - Lizard	2.58	0.01	0.015	
	Colubroidea - Non colubroid snakes	3.90	0.00	2.93E-04	
	Lizard - Non colubroid snakes	2.21	0.03	0.027	
Kruskal–Wallis test	Colubroidea - Other squamates				0.001
	3mer				
	Colubroidea - Lizard	-0.88	0.38	0.381	
	Colubroidea - Non colubroid snakes	2.61	0.01	0.014	
	Lizard - Non colubroid snakes	3.34	0.00	0.003	
Kruskal–Wallis test	Colubroidea - Other squamates				0.894
	4mer				
	Colubroidea - Lizard	4.87	0.00	3.32E-06	
	Colubroidea - Non colubroid snakes	3.58	0.00	0.001	
	Lizard - Non colubroid snakes	0.25	0.80	0.804	
Kruskal–Wallis test	Colubroidea - Other squamates				2.16E-07
	5mer				
	Colubroidea - Lizard	6.46	0.00	3.12E-10	
	Colubroidea - Non colubroid snakes	3.45	0.00	0.001	
	Lizard - Non colubroid snakes	-1.02	0.31	0.306	
Kruskal–Wallis test	Colubroidea - Other squamates				1.11E-10
	6mer				
	Colubroidea - Lizard	3.07	0.00	0.006	
	Colubroidea - Non colubroid snakes	3.06	0.00	0.003	
	Lizard - Non colubroid snakes	0.99	0.32	0.320	
Kruskal–Wallis test	Colubroidea - Other squamates				4.12E-04
	Tot SSR				
	Colubroidea - Lizard	4.61	0.00	1.24E-05	
	Colubroidea - Non colubroid snakes	4.50	0.00	1.04E-05	
	Lizard - Non colubroid snakes	1.39	0.17	0.165	
Kruskal–Wallis test	Colubroidea - Other squamates				1.40E-07

AATAG	Fold-change		ATAG	Fold-change	
	Bp/Mbp	Loci/Mbp		Bp/Mbp	Loci/Mbp
Colubroidea - Lizard	103.37	75.93		8.35	9.29
Colubroidea - Non colubroid snakes	56.35	70.65		5.29	6.04
Lizard - Non colubroid snakes	1.83	1.07		1.58	1.54
Colubroidea - Other squamates	87.69	74.73		7.43	8.33

Supplementary Data 7. Statistical analysis of AATAG microsatellite loci seeding by transposable elements for 8 squamate genomes.

P-values for a one-tailed Fisher's exact test (α= 0.01) according to the null hypothesis of higher transposable element (TE) content in 5 randomly extracted genomic background than at the 5'-end of genomic AATAG loci.

	Anolis carolinensis	*Boa constictor*	*Python molurus*	*Crotalus viridis*	*Crotalus mitchellii*	*Deinagkistrodon acutus*	*Thamnophis sirtalis*	*Ophiophagus hannah*
SINEs	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CR1/L3	**2e-03**	0.73	0.06	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**
L2	0.19	0.86	0.37	1.00	1.00	1.00	1.00	1.00
Rex	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**	**< 2.2e-16**
BovB	0.02	0.42	**0.00**	1.00	1.00	1.00	0.99	1.00
DIRS	1.00	0.32	0.98	1.00	**< 2.2e-16**	1.00	1.00	0.89
PLEs	0.99	0.99	0.97	1.00	1.00	1.00	1.00	1.00
LTRs	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DNA transposons	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Joint (Genomic AATAG-TE) and conditional (TE|AATAG) probabilities of co-occurrence of an AATAG locus and a transposable element (TE) in whole genome and AATAG adjacent context.

	Anolis carolinensis	*Boa constictor*	*Python molurus*	*Crotalus viridis*	*Crotalus mitchellii*	*Deinagkistrodon acutus*	*Thamnophis sirtalis*	*Ophiophagus hannah*
Genomic (AATAG-CR1)	5.08E-05	1.58E-04	1.50E-04	2.71E-03	2.16E-03	1.98E-03	2.58E-03	6.40E-05
P(CR1|AATAG)	5.58E-05	1.34E-04	1.75E-04	6.76E-03	3.70E-03	7.06E-03	6.15E-03	8.12E-03
Genomic (AATAG-Rex)	5.16E-07	2.51E-06	1.37E-06	6.32E-05	4.73E-05	1.88E-04	1.32E-04	1.09E-05
P(Rex|AATAG)	8.43E-05	2.00E-04	7.70E-05	1.53E-03	1.34E-03	1.61E-03	3.20E-03	4.96E-03

Supplementary Data 8. Adult body mass measurements for 66 sampled squamate species as from Feldman *et al*., 2016 (Body sizes and diversification rates of lizards, snakes, amphisbaenians and the tuatara. Global Ecology and Biogeography).

Species	Adult Maximum Mass (log10(g))
Gekkota	
Coleonyx elegans	1.568
Gekko gekko	2.112
Gekko japonicus	1.083
Scincoidea	
Zonosaurus madagascariensis	1.838
Platysaurus intermedius	1.2
Lepidophyma mayae	1.863
Lepidophyma flavimaculatum	1.161
Plestiodon fasciatus	1.541
Tribolonotus gracilis	1.374
Lamprolepis smaragdina	1.121
Lacertoidea	
Aspidoscelis scalaris	1.774
Proctoporus pachyurus	0.645
Bipes canaliculauts	1.147
Anguimorpha	
Varanus exanthematicus	1.283
Abronia graminea	1.339
Abronia matudai	2.166
Ophisaurus gracilis	2.129
Ophisaurus attenuatus	4
Iguania	
Trioceros melleri	1.603
Uromastyx benti	2.354
Pogona vitticeps	2.071
Dendragama boulengeri	1.136
Lophocalotes ludekingi	1.412
Gonocephalus grandis	2.158
Bronchocela jubata	2.071
Calotes versicolor	2.594
Sceloporus poinsettii	1.794
Sceloporus teapensis	1.804
Phrynosoma cornutum	1.865
Crotaphytus collaris	1.084
Oplurus quadrimaculatus	0.942
Norops humillis	0.424
Anolis carolinensis	2.077
Non Colubroid Snakes	
Typhlops reticulatus	2.024
Leptotyphlops dulcis	0.687
Anilius scytale	2.532
Boa constrictor	4.548
Eryx jaculus	2.405
Loxocemus bicolor	3.196
Python molurus	5.079
Casarea dussumieri	3.171
Acrochordus granulatus	2.478
Colubroidea	
Pareas carinata	1.735
Agkistrodon contortrix	3.096
Crotalus atrox	3.79
Crotalus mitchellii	3.115
Crotalus viridis	3.35
Sistrurus catenatus	2.754
Bothriechis schlegelli	2.466
Bothrops asper	3.875
Cerrophidion godmani	2.469
Tropidolaemus subannulatus	2.669
Deinagkistrodon acutus	3.221
Cerastes cerastes	2.57
Cerberus rhynchops	3.199
Ahaetulla prasina	2.817
Coluber constrictor	2.783
Pantherophis emoryi	2.538
Pantherophis guttatus	2.735
Thelotornis kirtlandii	2.587
Coniophanes fissidens	1.974
Coniophanes piceivittis	1.608
Sibon nebulatus	2.241
Thamnophis sirtalis	2.936
Micrurus fulvius	2.607
Ophiophagus Hannah	4.249

Supplementary Data 11. BEAST 2 priors used to estimate divergence times. Clade ages for the Amniotes, Reptiles, Lepidosaurs, and Archosaurs were obtained from Benton and Donogue 2007, except for Alethinophidian snakes, Iguania, and Gekkota (Pyron et al. 2015).

Clade	Distribution	Mean	Sigma	Offset
Reptile-Mammal Split	Normal	321.3	4.64	0
Archosaur-Lepidosaur Split	Normal	277.8	11.2	0
Squamate-Rhynchocephalia	Log Normal	1	1	222.8
Crocodile-Avian Split	Normal	244.7	2.9	0
Gekkota	Normal	86.5	10	0
Iguania	Normal	146.4	5	0
Alethinophidia	Normal	102.75	4.625	0

Supplementary Data 12. Generation time assessments used for PSMC.

Species	Generation time	Source
Ophisaurus gracilis	2 years	Lindemann; Animal diversity web 2009
Pogona vitticeps	1-2 years	Pest risk assessment: Central bearded dragon (*Pogona vitticeps*). DPIPWE 2011
Boa constrictor	3 years	Lindemann; Animal diversity web 2009
Python molurus	2-3 years	Lindemann; Animal diversity web 2009
Crotalus mitchellii	3 years	Klauber; *Rattlesnakes: Their Habits, Life Histories, and Influence on Mankind*. University
Crotalus viridis	2-3 years	Lindemann; Animal diversity web 2009
Deinagkistrodon acutus	3-4 years	Lindemann; Animal diversity web 2009
Thamnophis sirtalis	1-2 years	Lindemann; Animal diversity web 2009

APPENDIX B

CHAPTER 3 SUPPLEMENTARY TABLES

Supplementary Table S1. Sequencing libraries used in the prairie rattlesnake genome assembly. Where noted, various libraries were used in the previous assembly (CroVir2.0), published in Pasquesi et al. (2018).

Library	Read Type	Number of Reads	Assembly Version
50bp short read	single end	9,536,384	CroVir2.0
100bp short read	paired end	449775645	CroVir2.0, CroVir3.0
150bp short read	paired end	41,211,014	CroVir2.0
150bp long insert mate pair (3-5Kb)	paired end	188,532,564	CroVir2.0
150bp long insert mate pair (6-8Kb)	paired end	189,928,342	CroVir2.0
PacBio long reads	-	1,027,365	CroVir2.0
Chicago long range proximity ligation library 1 (150bp)	paired end	251,689,106	CroVir3.0
Chicago long range proximity ligation library 2 (150bp)	paired end	206,176,028	CroVir3.0
Hi-C library 1 (150bp)	paired end	230,083,402	CroVir3.0
Hi-C library 2 (150bp)	paired end	160,673,944	CroVir3.0

Supplementary Table S2. Basic information about assembly versions for the prairie rattlesnake genome.

	Input Assembly (CroVir2.0)	Chicago Assembly	HiRise (Chicago + Hi-C) Assembly
Longest Scaffold (bp)	1,184,546	11,576,738	311,712,589
Number of Scaffolds	47,782	8,183	7,034
Number of Scaffolds > 1Kb	47,658	8,059	6,910
Contig N50 (Kb)	15.81	14.91	14.96
Scaffold N50 (Kb)	139	2,472	179,898
Number of Gaps	112,369	158,269	159,024
Percent of Genome in Gaps	5.84%	6.15%	6.16%

Supplemental Table S3. RNA-seq libraries used for transcriptome assembly. Raw reads for each library are available on the NCBI Short Read Archive, accession PRJNA477004.

Sample ID	Tissue	Raw Reads	Quality Trimmed Reads
CroVirPan	pancreas	28,126,703	27,073,946
CroVirTon	tongue	24,451,116	23,561,349
CroVirVG1	venom gland	41,744,110	40,147,306
CroVirVG3	venom gland	29,216,664	28,035,353
Cvv01	liver	7,833,506	7,365,740
Cvv02	liver	7,451,792	7,064,234
Cvv11	liver	9,218,939	8,441,587
Cvv20	kidney	6,958,120	6,580,387
Cvv22	kidney	8,116,679	7,601,517
Cvv23	kidney	7,193,762	6,785,947
Cvv25	skin	7,849,895	7,303,441
Cvv26	pancreas	8,886,612	8,160,214
Cvv27	venom gland	3,098,151	2,928,974
Cvv28	lung	6,613,196	6,024,613
Cvv29	testes	5,055,189	4,745,375
Cvv30	accessory venom gland	3,261,326	3,053,142
Cvv31	shaker muscle	4,290,989	3,996,274
Cvv32	pancreas	4,836,715	4,566,165
Cvv33	brain	3,815,570	3,569,113
Cvv34	stomach	5,297,110	4,993,142
Cvv35	ovaries	3,737,870	3,528,104
Cvv36	rictal gland	6,654,626	6,070,883
Cvv37	spleen	7,776,020	6,975,210
Cvv38	blood	2,550,433	2,364,162

Supplemental Table S4. BUSCO results for assembly versions of the prairie rattlesnake genome. Proportions of each category are in parentheses.

BUSCO category	CroVir2.0	CroVir3.0 (current)
Complete	3,277 (83.0 %)	3,372 (85.3 %)
Complete and single-copy	3,253 (82.4%)	3,347 (84.7%)
Complete and duplicated	24 (0.6%)	25 (0.6%)
Fragmented	364 (9.2%)	298 (7.5%)
Missing	309 (7.8%)	280 (7.2%)
Total searched	3,950	3,950

Supplementary Table S5. Genome-wide annotated repeat proportions identified using RepeatMasker.

	# elements	length masked (bp)	% of sequence	% element masked
Total masked	2966274	489373735	38.91	100.00
Total interspersed repeats	2348232	463237605	36.83	79.16
Retroelements	1139213	295244109	22.81	38.41
SINEs	173332	22894322	1.82	5.84
Squam1/Sauria	19230	3376458	0.27	0.65
Other SINEs	126898	15602678	1.24	4.28
LINEs	621859	170275973	13.54	20.96
CR1-Like	359387	91177000	7.25	12.12
CR1/L3	288888	74285822	5.91	9.74
L2	53219	12036490	0.96	1.79
Rex	19032	5339363	0.42	0.64
R1/LOA/Jockey	3272	854611	0.07	0.11
R2/R4/NeSL	35256	9045775	0.72	1.19
RTE/Bov-B	101958	32795496	2.61	3.44
L1/CIN4	78926	28358227	2.25	2.66
Other LINEs	154019	16472232	0.64	5.19
Other nonLTR	10119	1572442	0.13	0.34
DIRS	28657	13553057	1.08	0.97
PLEs	120162	19278497	1.53	4.05
LTR elements	156427	54116761	4.30	5.27
BEL/Pao	4007	1927682	0.15	0.14
Ty1/Copia	9160	3340874	0.27	0.31
Gypsy	77793	35080772	2.79	2.62
Retroviral	16727	5393228	0.43	0.56
Other LTR	48740	8374205	0.67	1.64
DNA transposons	850487	125287793	9.96	28.67
hobo-Activator	428247	60243144	4.79	14.44
Tc1-IS630-Pogo	283367	48888185	3.89	9.55
En-Spm	12485	1964905	0.16	0.42
MuDR-IS905	1300	383077	0.03	0.04
PiggyBac	131	22504	0.00	0.00
Tourist/Harbinger	80904	7193605	0.57	2.73
P elements	155	45074	0.00	0.01
Rolling-circles	3736	635885	0.05	0.13
SPIN	253	26640	0.00	0.01
Other DNA	39909	5884774	0.47	1.35
Unclassified	358532	48493199	3.86	12.09
Total interspersed repeats	2348232	463237605	36.83	79.16
Small RNA	2054	174940	0.01	0.07
Satellites	4952	1104344	0.09	0.17
Simple repeats	540288	28572170	2.27	18.21
Low complexity	70748	4755565	0.38	2.39

Supplemental Table S6. Mapping of cDNA markers from Matsubara et al. 2006 to the Prairie Rattlesnake genome. Locations of best BLAST hits of each cDNA marker to the genome are reported. Markers that mapped with exceptional similarity to multiple locations in the genome are denoted with a '#', and markers that did not map to the chromosome as predicted by Matsubara et al. (2006) are denoted with a '*'. Details for these markers are provided in Supplemental Table S7 and Supplemental Fig. S2.

Marker	Accession	Chromosome	Scaffold	e-value	bit-score	Start Position	End Position
OMG	BW999947	1p	scaffold-ma1	6.00E-115	398	309337082	309336564
XAB1	AU312353	1p	scaffold-ma1	2.00E-46	122	297437298	297437486
MGC15407	AU312344	1p	scaffold-ma1	2.00E-65	92.3	288097081	288097206
XPO1	AU312325	1p	scaffold-ma1	2.00E-113	153	289547707	289547901
DEGS	AU312341	1p	scaffold-ma1	5.00E-106	356	269312409	269311948
KIAA0007	AU312332	1p	scaffold-ma1	5.00E-50	120	265943692	265943841
EPRS	AU312324	1p	scaffold-ma1	2.00E-91	174	270708945	270709160
ARID4B	AU312346	1p	scaffold-ma1	1.00E-129	333	252059286	252059699
QKI	AU312356	1p	scaffold-ma1	5.00E-112	124	246094729	246094887
MDN1	AU312339	1p	scaffold-ma1	7.00E-60	109	211517498	211517349
AFTIPHILIN	AU312311	1p	scaffold-ma1	5.00E-75	112	170752748	170752888
SF3B1	AU312337	1q	scaffold-ma1	7.00E-95	215	150078848	150078576
CACNB4	BW999948	1q	scaffold-ma1	1.00E-47	102	127283965	127283819
ZFHX1B	BW999949	1q	scaffold-ma1	6.00E-93	204	123301385	123301101
UMPS	AU312331	1q	scaffold-ma1	8.00E-95	198	113761458	113761724
TCIRG1	BW999950	1q	scaffold-ma1	2.00E-72	164	102088882	102089094
TSG101	AU312316	1q	scaffold-ma1	4.00E-76	113	88358887	88359054
M11S1	AU312350	1q	scaffold-ma1	4.00E-31	94.5	70777673	70777560
GPHN	AU312327	1q	scaffold-ma1	5.00E-68	116	60249829	60249644
DNCH1	AU312310	1q	scaffold-ma1	1.00E-71	145	25060055	25059885
HSPCA	BW999951	1q	scaffold-ma1	2.00E-123	149	25029984	25030184
ISYNA1	AU312338	1q	scaffold-ma1	2.00E-89	178	7770987	7771196
TUBGCP2	AU312343	1q	scaffold-ma1	4.00E-74	136	9697568	9697377
ZFR	AU312309	2p	scaffold-ma2	8.00E-110	208	222653709	222653461
PHAX	AU312322	2p	scaffold-ma2	3.00E-99	224	189308026	189307715
VPS13A	BW999952	2p	scaffold-ma2	9.00E-70	109	179725513	179725656
UBQLN1	BW999953	2p	scaffold-ma2	2.00E-87	132	182156077	182156238
C9orf72	AU312326	2p	scaffold-ma2	5.00E-91	203	164760033	164760347
KIAA0368	BW999954	2p	scaffold-ma2	1.00E-56	116	161287251	161287397
TOPORS	BW999955	2p	scaffold-ma2	8.00E-118	410	162258381	162257809
FAM48A	BW999956	2cen	scaffold-ma2	1.00E-45	102	157286823	157286680
UNQ501	AU312305	2cen	scaffold-ma2	6.00E-118	284	142895238	142895636
DCTN2	AU312317	2q	scaffold-ma2	4.00E-80	122	122527271	122527110
EXOC7	BW999957	2q	scaffold-ma2	3.00E-93	121	92952368	92952526
DDX5	BW999958	2q	scaffold-ma2	7.00E-112	144	108253948	108253775
CCNG1	AU312308	2q	scaffold-ma2	6.00E-70	173	80553964	80553731
CPEB4	AU312333	2q	scaffold-ma2	3.00E-119	250	72297563	72297874
FLJ22318	AU312329	2q	scaffold-ma2	2.00E-105	194	51908839	51908582
DCTN4	AU312349	2q	scaffold-ma2	4.00E-50	99.6	58962806	58962928
C5orf14	AU312304	2q	scaffold-ma2	4.00E-120	329	64853582	64853127
NOSIP[*]	AU312303	2q	scaffold-Z	1.00E-51	93.6	92988551	92988661
RBM5[#]	BW999960	2q	scaffold-mi8	6.00E-78	90.4	9620291	9620181
RBM5[#]	BW999960	2q	scaffold-ma2	7.00E-13	76.1	130725514	130725606
ITPR1	BW999961	2q	scaffold-ma2	9.00E-53	135	23858424	23858585
ENPP2	BW999962	3p	scaffold-ma3	6.00E-90	121	9756367	9756209
YWHAZ	BW999963	3p	scaffold-ma3	2.00E-99	180	16759896	16760114
LRRCC1	BW999964	3p	scaffold-ma3	4.00E-83	150	21993774	21993565
LYPLA1	BW999965	3p	scaffold-ma3	3.00E-107	149	31673258	31673440
SS18	AU312302	3p	scaffold-ma3	1.00E-83	126	36811554	36811724
MBP	AU312318	3p	scaffold-ma3	7.00E-111	179	49049170	49049382
EPB41L3	BW999966	3p	scaffold-ma3	3.00E-84	141	40222999	40222808
TUBB2A	BW999967	3p	scaffold-ma3	8.00E-91	155	59187732	59187532
LRRC16	BW999968	3p	scaffold-ma3	2.00E-100	144	51025171	51025350
SERPINB6[#]	BW999969	3p	scaffold-ma5	5.00E-99	130	36540937	36540755
SERPINB6[#]	BW999969	3p	scaffold-ma3	2.00E-76	113	60484038	60483865
BPHL	BW999970	3p	scaffold-ma3	1.00E-87	118	59199779	59199621

KIF13A	BW999971	3p	scaffold-ma3	3.00E-78	139	53681516	53681349
TPR	BW999972	3q	scaffold-ma3	6.00E-83	122	93408800	93408636
AKR1A1	BW999973	3q	scaffold-ma3	9.00E-75	153	133869419	133869619
ZNF326[*]	BW999974	3q	scaffold-ma2	2.00E-77	120	224940437	224940586
YIPF1	BW999975	3q	scaffold-ma3	6.00E-52	112	127724189	127724353
BCAS2	AU312354	3q	scaffold-ma3	3.00E-51	141	151621402	151621229
KIAA1219	BW999976	3q	scaffold-ma3	4.00E-101	158	155122635	155122844
STAU1	BW999977	3q	scaffold-ma3	2.00E-116	169	165663812	165663594
RBM12	BW999978	3q	scaffold-ma3	2.00E-152	406	154706304	154705780
TPT1	BW999979	4p	scaffold-ma4	2.00E-68	148	1006155	1006349
EIF2S3	AU312306	4p	scaffold-ma4	1.00E-111	126	49115724	49115885
SYAP1	AU312328	4p	scaffold-ma4	3.00E-96	121	46147275	46147135
DSCR3	AU312319	4q	scaffold-ma4	1.00E-74	119	60873037	60872873
DCAMKL1	BW999980	4q	scaffold-ma4	8.00E-49	110	86291138	86291302
ELMOD1	BW999981	4q	scaffold-ma4	1.00E-56	147	93207704	93207522
BCCIP	AU312307	5q	scaffold-ma5	1.00E-46	148	32597249	32597061
SH3MD1	AU312347	5q	scaffold-ma5	2.00E-119	378	45831798	45832379
PPP1R7	BW999982	5q	scaffold-ma5	2.00E-92	228	56956062	56955736
PDCD10	AU312342	5q	scaffold-ma5	4.00E-61	143	74805371	74805547
TLOC1	AU312335	5q	scaffold-ma5	2.00E-45	101	76109988	76110125
UCHL1[*]	BW999983	6p	scaffold-ma7	4.00E-89	210	33298090	33298407
GNAI2[*]	BW999984	6p	scaffold-ma2	2.00E-106	126	49893686	49893841
P4HB[*]	BW999985	6p	scaffold-ma2	2.00E-69	100	97717890	97718012
FLJ12571	AU312352	6q	scaffold-ma6	2.00E-46	117	46698606	46698752
RANGAP1	AU312313	6q	scaffold-ma6	7.00E-71	95	47795604	47795500
LDHB	BW999986	6q	scaffold-ma6	2.00E-60	117	69268248	69268418
SEC3L1	AU312345	7p	scaffold-ma7	3.00E-58	125	55644074	55643916
KIAA1109	AU312348	7q	scaffold-ma7	2.00E-60	124	30398905	30398711
RAP1GDS1	AU312351	7q	scaffold-ma7	2.00E-91	112	12141068	12140931
GAD2	BW999991	Zp	scaffold-Z	1.00E-109	136	17484512	17484336
WAC	AU312355	Zp	scaffold-Z	3.00E-93	209	16303681	16303947
KLF6[*]	BW999992	Zp	scaffold-ma2	1.00E-99	366	47130305	47130796
LOC90693[#]	BW999993	Zp	scaffold-ma7	4.00E-127	301	34444161	34444577
LOC90693[#]	BW999993	Zp	scaffold-Z	1.00E-107	291	34827559	34827182
TAX1BP1	AU312320	Zp	scaffold-Z	1.00E-86	141	36989995	36990174
RAB5A	BW999994	Zp	scaffold-Z	9.00E-94	166	40227424	40227215
CTNNB1	BW999995	Zcen	scaffold-Z	3.00E-129	275	49548885	49549226
AMPH	BW999996	Zcen	scaffold-Z	1.00E-66	101	55612836	55612955
TUBG1	BW999997	Zq	scaffold-Z	5.00E-89	116	17359265	17359113
GH1	BW999998	Zq	scaffold-Z	2.00E-115	179	77397011	77396727
MYST2	BW999999	Zq	scaffold-Z	6.00E-122	293	90785118	90784714
NEF3	BW999987	micro	scaffold-mi1	1.00E-102	352	13833430	13832942
ASB6	AU312340	micro	scaffold-mi7	1.00E-95	161	6270589	6270353
RPL12	BW999988	micro	scaffold-mi7	6.00E-67	95.5	7974658	7974542
FLJ25530	AU312336	micro	scaffold-mi1	4.00E-98	255	8157147	8156806
HSPA8[#]	BW999989	micro	scaffold-ma1	2.00E-124	236	20422342	20422662
HSPA8[#]	BW999989	micro	scaffold-mi1	3.00E-123	259	2089357	2089025
GLCE	AU312330	micro	scaffold-mi10	1.00E-79	234	24861	24577
POLG	AU312315	micro	scaffold-mi3	4.00E-97	116	10042696	10042845
LOC283820	AU312323	micro	scaffold-mi5	8.00E-71	116	3659851	3659708
PARN	AU312312	micro	scaffold-mi7	1.00E-66	73.9	12029447	12029361
ATRX	BW999990	micro	scaffold-mi4	3.00E-63	102	1268001	1268126

Supplemental Table S7. Details of mismatched cDNA markers from *Elaphe quadrivirgata* (Matsubara et al. 2006), their locations in *Crotalus* and *Anolis*, and notes on likelihood of misassembly based on synteny and intrachromosomal Hi-C.

Marker	*Elaphe* Chromosome	*Crotalus* Scaffold	*Anolis* Scaffold	Notes
NOSIP	2q	scaffold-Z	6	Unique or erroneous original cDNA placement; *Anolis* and *Crotalus* synteny suggest correct placement in *Crotalus*; Hi-C data inconsistent with misassembly
ZNF326	3q	scaffold-ma2	4	Possible misassembly error; *Anolis* and *Elaphe* synteny suggest incorrect placement in Crotalus; Hi-C data consistent with regional misassembly
UCHL1	6p	scaffold-ma7	5	Hit to *Anolis* 5 is inconclusive because it is syntenic with snake chromosomes 6 & 7; Hi-C data inconsistent with misassembly
GNAI2	6p	scaffold-ma2	2	Unique or erroneous original cDNA placement; *Anolis* and *Crotalus* synteny suggest correct placement in *Crotalus*; Hi-C data inconsistent with misassembly
P4HB	6p	scaffold-ma2	2	Unique or erroneous original cDNA placement; *Anolis* and *Crotalus* synteny suggest correct placement in *Crotalus*; Hi-C data inconsistent with misassembly
KLF6	Zp	scaffold-ma2	6	Possible misassembly error; *Anolis* and *Elaphe* synteny suggest incorrect placement in *Crotalus*; Hi-C data inconsistent with misassembly

Supplemental Table S8. GC variation in windows of various sizes for 12 squamate species. Values for each species are measured as the standard deviation (SD) of GC content in all sampled windows of a given size. Information for 5, 20, and 80 kb windows are also presented in Fig. 1c. Missing data (i.e., window sizes that were too large and contained greater than the threshold allowed missing data) are denoted with '-'.

Window Size (bp)	*Gekko japonicus*	*Eublepharis macularius*	*Ophisaurus gracilis*	*Shinisaurus crocodilurus*	*Pogona vitticeps*	*Anolis carolinensis*
5,000	0.039295606	0.037140406	0.037038224	0.03488877	0.03681681	0.032312269
20,000	0.028980944	0.027338004	0.029217483	0.027425317	0.030930264	0.021209
40,000	0.025219459	0.024838347	0.027141528	0.025322106	0.029367252	0.017608402
80,000	0.021385708	0.023326607	0.025558162	0.023843432	0.028238318	0.015121097
160,000	0.01811246	0.022646783	0.024536212	0.022632678	0.027330318	0.013089382
240,000	-	0.022203903	0.023356372	0.021943776	0.026943855	0.012088733
320,000	-	0.022121291	0.022899173	0.021312719	0.026617904	0.011287772

Window Size (bp)	*Boa constrictor*	*Python molurus*	*Ophiophagus hannah*	*Thamnophis sirtalis*	*Deinagkistrodon acutus*	*Crotalus viridis*
5,000	0.043942864	0.042024505	0.040098669	0.047076022	0.047062019	0.041210929
20,000	0.034934365	0.035837726	0.031894398	0.037865804	0.03882085	0.032232558
40,000	0.030576918	0.033337717	0.028952912	0.03429097	0.036517713	0.029884634
80,000	0.023292703	0.030197592	0.026685436	0.031202717	0.034964163	0.0281043
160,000	0.014736549	0.02736241	0.024597185	0.02894796	0.033486765	0.026806291
240,000	-	0.024725646	0.023968494	0.026250057	0.032562166	0.02616041
320,000	-	0.023707617	0.023468328	0.024606171	0.031784231	0.025840409

Supplemental Table S9. Details of Illumina Nextera resequencing and RNAseq libraries used for comparative female/male read coverage across the rattlesnake genome and sex-specific gene expression analyses. Raw read data are available on NCBI under accession PRJNA476794.

Library Type	**Read Length**	**Sample ID**	**Tissue**	**Sex**	**Number of Mapped Reads**
Illumina Nextera	150 bp paired end	CV0007	Liver	Male	20,279,801
Illumina Nextera	150 bp paired end	CV0011	Liver	Female	4,975,491
RNAseq	100 bp paired end	Cv3	Liver	Female	3,774,322
RNAseq	100 bp paired end	Cv8	Liver	Female	3,680,195
RNAseq	100 bp paired end	Cv3	Kidney	Female	3,256,208
RNAseq	100 bp paired end	Cv3	Kidney	Female	4,565,008
RNAseq	100 bp paired end	Cv5	Liver	Male	3,330,125
RNAseq	100 bp paired end	Cv6	Liver	Male	3,837,264
RNAseq	100 bp paired end	Cv5	Kidney	Male	3,729,811
RNAseq	100 bp paired end	Cv6	Kidney	Male	4,673,928

Supplementary Table S10. Representative sequences for known snake venom gene families used to annotate venom genes in the rattlesnake genome.

Gene Family	Accession	Sequence Type	Species
5'Nucleotidase	AK291667.1	mRNA	*Homo sapiens*
Acetylcholinesterase	U54591.1	mRNA	*Bungarus fasciatus*
AVItoxin	EU195459.1	mRNA	*Varanus komodoensis*
C-type Lectin	JF895761.1	mRNA	*Crotalus oreganus helleri*
Cobra Venom Factor	U09969.2	mRNA	*Naja kaouthia*
CRISp (cysteine-rich secretory protein)	HQ414088.1	mRNA	*Crotalus adamanteus*
Cystatin	FJ411289.1	mRNA	*Naja kaouthia*
Extendin	EU790960.1	mRNA	*Heloderma suspectum*
Exonuclease	XM_015826835.1	mRNA	*Protobothrops mucrosquamatus*
Hyaluronidase	HQ414098.1	mRNA	*Crotalus adamanteus*
LAAO (L-amino acid oxidase)	HQ414099.1	mRNA	*Crotalus adamanteus*
SVMP I (class I snake venom metalloproteinase)	HM443635.1	mRNA	*Bothrops neuwiedi*
SVMP II (class II snake venom metalloproteinase)	HM443637.1	mRNA	*Bothrops neuwiedi*
SVMP III (class III snake venom metalloproteinase)	HM443632.1	mRNA	*Bothrops neuwiedi*
Nerve growth factor	AF306533.1	mRNA	*Crotalus durissus terrificus*
Phosphodiesterase	HQ414102.1	mRNA	*Crotalus adamanteus*
PLA2_I (vipers)	AF403134.1	mRNA	*Crotalus viridis viridis*
PLA2_II (elapids)	GU190815.1	mRNA	*Bungarus flaviceps*
Sarafotoxin	L07528.1	mRNA	*Atractaspis engaddensis*
Serine Proteinase	HQ414121.1	mRNA	*Crotalus adamanteus*
3FTX (Three-finger Toxin)	DQ273582.1	mRNA	*Ophiophagus hannah*
Veficolin	GU065323.1	mRNA	*Cerberus rynchops*
VEGF (Vascular Endothelial Growth Factor)	AB848141.1	mRNA	*Protobothrops mucrosquamatus*
Vespryn	EU401840.1	mRNA	*Oxyuranus scutellatus*
Waprin	EU401843.1	mRNA	*Oxyuranus scutellatus*
Kunitz (serine peptidase inhibitor, Kunitz type)	JU173666.1	mRNA	*Crotalus adamanteus*
Thrombin-like (thrombin-like venom gland enzyme)	AJ001209.1	mRNA	*Deinagkistrodon acutus*
Ficolin	GBUG01000048.1	mRNA	*Echis coloratus*
Disintegrin	AJ131345.1	mRNA	*Deinagkistrodon acutus*
FactorV (venom coagulation factor V)	XM_015815922.1	mRNA	*Protobothrops mucrosquamatus*
FactorX	XM_015819885.1	mRNA	*Protobothrops mucrosquamatus*
Prokineticin	XM_015822870.1	mRNA	*Protobothrops mucrosquamatus*
Ohanin (ohanin-like)	XM_015818414.1	mRNA	*Protobothrops mucrosquamatus*
Complement C3 (Cadam VF)	JU173742.1	mRNA	*Crotalus adamanteus*
Crotasin	AF250212.1	mRNA	*Crotalus durissus terrificus*
Endothelin	XM_015810852.1	mRNA	*Protobothrops mucrosquamatus*
Kallikrein	GALC01000005.1	mRNA	*Crotalus oreganus helleri*
Lynx1 (Ly6/neurotoxin 1)	XM_014066791.1	mRNA	*Thamnophis sirtalis*
Natriuretic Peptide (bradykinin potentiating peptide and C-type natriuretic peptide precursor isoform 2)	AF308594.2	mRNA	*Crotalus durissus terrificus*
sPla/ryanodine receptor	XM_015823102.1	mRNA	*Protobothrops mucrosquamatus*
WAP four-disulfide core domain protein 5 (Whey Acidic Protein/secretory leuki proteinase inhibitor)	XM_015822353.1	mRNA	*Protobothrops mucrosquamatus*
Myotoxin	HQ414100.1	mRNA	*Crotalus adamanteus*
PLA2	APD70899.1	protein	*Crotalus atrox*
SVMP	Q90282.1	protein	*Crotalus atrox*
Serine Proteinase	F8S114.1	protein	*Crotalus adamanteus*

Supplementary Table S11. Annotated venom gene homologs in the prairie rattlesnake genome. Genes were annotated using materials detailed in Supplementary Table 9.

Venom Gene Family	Rattlesnake Scaffold	Start Position (bp)	End Position (bp)
3-Finger toxin	scaffold-ma1	103004868	103021927
3-Finger toxin	scaffold-ma1	102999393	103000958
5' Nucleotidase	scaffold-ma5	46133017	46179118
5' Nucleotidase	scaffold-ma6	55711914	55732365
5' Nucleotidase	scaffold-mi1	18004217	18021456
5' Nucleotidase	scaffold-ma2	45090212	45121335
5' Nucleotidase	scaffold-ma2	134237148	134264183
Acetylcholinesterase	scaffold-ma2	4047955	4053281
Acetylcholinesterase	scaffold-ma2	3948506	3952373
Acetylcholinesterase	scaffold-ma2	4016363	4018146
Acetylcholinesterase	scaffold-ma2	4026170	4045822
Acetylcholinesterase	scaffold-ma5	73971094	73976212
Acetylcholinesterase	scaffold-ma5	74015346	74036663
Acetylcholinesterase	scaffold-un210	16032	17552
Bradykinin potentiating and natriuretic peptide	scaffold-un187	22386	23524
C-type lectin	scaffold-mi5	3276042	3284747
C-type lectin	scaffold-mi5	11650747	11653723
C-type lectin	scaffold-Z	21883578	21895509
C-type lectin	scaffold-Z	21706900	21776775
C-type lectin	scaffold-Z	21786524	21797211
C-type lectin	scaffold-Z	108214710	108236532
Cysteine-rich secretory protein	scaffold-ma1	169434958	169437996
Cysteine-rich secretory protein	scaffold-ma1	169423774	169434684
Cysteine-rich secretory protein	scaffold-ma3	25391938	25416947
Cysteine-rich secretory protein	scaffold-mi6	1021447	1040191
Exonuclease	scaffold-mi7	8097114	8103411
Exonuclease	scaffold-ma1	5804894	5842638
Exonuclease	scaffold-mi3	10271502	10274220
Exonuclease	scaffold-ma6	12590208	12591465
Factor V	scaffold-mi4	8493826	8518402
Factor V	scaffold-mi4	8479637	8493564
Factor V	scaffold-ma4	81074882	81113119
Glutaminyl cyclase	scaffold-ma1	256551622	256564040
Glutaminyl cyclase	scaffold-mi7	5091107	5094268
Hyaluronidase	scaffold-ma6	14952252	14955850
Hyaluronidase	scaffold-ma2	45901201	45920587
Hyaluronidase	scaffold-ma2	49137409	49145188
Hyaluronidase	scaffold-ma2	49106981	49118469
Kunitz peptide	scaffold-mi7	3590975	3597607
Kunitz peptide	scaffold-mi8	4992795	5002390
L-amino acid oxidase	scaffold-ma4	56914906	56948498
L-amino acid oxidase	scaffold-ma4	85461961	85468906
L-amino acid oxidase	scaffold-ma2	4658599	4661642
L-amino acid oxidase	scaffold-ma2	4654769	4658293
Myotoxin/crotamine	scaffold-ma1	289328153	289328605
Nerve growth factor	scaffold-Z	93342025	93347811
Nerve growth factor	scaffold-ma1	76711308	76727703
PLA2	scaffold-mi7	3019970	3021876
PLA2	scaffold-mi7	3027607	3029199
PLA2	scaffold-mi7	3031464	3033348
PLA2	scaffold-mi7	3037103	3038488
PLA2	scaffold-mi7	3042118	3043697

Serine Proteinase	scaffold-mi2	8569773	8575182
Serine Proteinase	scaffold-mi2	8588278	8593660
Serine Proteinase	scaffold-mi2	8628274	8636651
Serine Proteinase	scaffold-mi2	8664603	8670797
Serine Proteinase	scaffold-mi2	8739986	8745649
Serine Proteinase	scaffold-mi2	8752578	8759324
Serine Proteinase	scaffold-mi2	8864675	8879153
Serine Proteinase	scaffold-mi2	8937526	8947481
Serine Proteinase	scaffold-mi2	8960028	8980478
Snake venom metalloproteinase	scaffold-mi1	13901629	14014239
Snake venom metalloproteinase	scaffold-mi1	14022082	14075370
Snake venom metalloproteinase	scaffold-mi1	14091987	14112667
Snake venom metalloproteinase	scaffold-mi1	14147865	14170405
Snake venom metalloproteinase	scaffold-mi1	14174872	14190142
Snake venom metalloproteinase	scaffold-mi1	14211673	14242249
Snake venom metalloproteinase	scaffold-mi1	14248933	14272689
Snake venom metalloproteinase	scaffold-mi1	14281564	14300774
Snake venom metalloproteinase	scaffold-mi1	14368422	14393313
Snake venom metalloproteinase	scaffold-mi1	14401627	14424637
Snake venom metalloproteinase	scaffold-mi1	14310844	14338336
Veficolin/Ficolin	scaffold-mi7	5271880	5282014
Veficolin/Ficolin	scaffold-ma3	179788950	179790745
Veficolin/Ficolin	scaffold-ma1	232337083	232340714
Veficolin/Ficolin	scaffold-ma1	232312034	232335439
Vascular endothelial growth factor	scaffold-ma7	40288572	40327884
Vascular endothelial growth factor	scaffold-ma1	40733075	40747358
Vascular endothelial growth factor	scaffold-ma1	260248287	260272500
Venom Factor	scaffold-Z	79798672	79803249
Venom Factor	scaffold-Z	79749464	79761456
Venom Factor	scaffold-ma2	1573588	1616446
Venom Factor	scaffold-ma2	137559964	137560374
Venom Factor	scaffold-ma2	137553669	137558461
Venom Factor	scaffold-ma2	137623562	137648584
Venom Factor	scaffold-ma2	137651285	137653877
Venom Factor	scaffold-ma2	137710627	137728987
Venom Factor	scaffold-ma2	137753804	137775039
Venom Factor	scaffold-ma2	137735629	137741352
Vespryn/Ohanin	scaffold-ma2	4377779	4385668
Vespryn/Ohanin	scaffold-ma2	109834300	109838076
Waprin	scaffold-ma1	204655764	204666466

Supplemental Table S12. Transcription factors significantly upregulated in the venom gland. Mean distances summarize the distribution of distances between gene venom genes and non-venom genes and the nearest predicted binding site of each transcription factor. No position weight matrix for *NCOA2* was available for a close relative to the rattlesnake, and the *NFI* family transcription factors have a conserved binding motif, and are summarized together under *NFIA*. P-values are from *t*-test comparisons of distance distributions.

Gene ID	Rattlesnake Gene Detail	Mean Distance to	Mean Distance to	p-value
ATF6	augustus_masked-scaffold-ma3-processed-gene-300.3	421,305.1	595,006.2	0.002793
ELF5	maker-scaffold-ma1-augustus-gene-235.5	1,121.3	1,203.9	0.7953
FOXC2	augustus_masked-scaffold-mi6-processed-gene-2.1	202,416.2	251,898.5	0.02967
CREB3L2	maker-scaffold-ma6-augustus-gene-195.2	32,227.9	29,708.3	0.5558
GRHL1	maker-scaffold-ma1-augustus-gene-601.8	78,954.0	86,147.0	0.4343
NCOA2	maker-scaffold-ma3-augustus-gene-89.6	-	-	-
NFIA	maker-scaffold-ma3-augustus-gene-414.2	336,765.8	328,556.3	0.7968
NFIB	maker-scaffold-ma2-augustus-gene-569.3	-	-	-
NFIB	maker-scaffold-ma2-augustus-gene-569.2	-	-	-
NFIX	maker-scaffold-ma2-augustus-gene-473.3	-	-	-
NR4A2	maker-scaffold-ma1-augustus-gene-428.4	100,375.5	92,292.3	0.492
SREBF2	maker-scaffold-ma6-augustus-gene-158.15	306,901.1	328,081.4	0.4302

Supplemental Table S13. Density of predicted GRHL1 and NFI binding sites within given intervals of venom genes and all nonvenom genes. P-values are reported from Fisher's exact tests, which compared the number of predicted binding sites by the total length of sequenced searched between venom and nonvenom gene sets.

Transcription Factor	Interval (kb)	Venom Gene Density	Nonvenom Gene Density	p-value
GRHL1	100 kb	7.44E-06	8.03E-06	0.4022
GRHL1	50 kb	1.49E-05	1.74E-05	0.2127
GRHL1	10 kb	3.00E-05	2.78E-05	0.6875
GRHL1	5 kb	4.44E-05	3.97E-05	0.554
NFI	Promoter (1kb)	1.80E-03	1.43E-03	0.1305

APPENDIX C

CHAPTER 4 SUPPLEMENTARY TABLES

Supplementary File 1. Acession data for samples used in this study. Raw data were quality and adapter trimmed, and cleaned from rRNA transcripts before being analyzed.

Species	Acronym	Testis	Ovary	Brain	Heart	Kidney	Liver	Muscle	Spleen	Small Intestine
Danio rerio (zebrafish)	*Dr*	SRR2177445	SRR2177444	SRR1609735	SRR1609741	SRR1609747	SRR1609750	SRR1609753	SRR1609756	SRR1609744
Xenopus laevis (clawed frog)	*Xl*	SRR2515162	SRR2515157	SRR2515149	SRR2515151	SRR2515153	SRR2515154	SRR2515156	SRR2515160	SRR2515152
Alligator mississippiensis (alligator)	*Am*	SRR3208129	SRR3208147	SRR3208125	SRR3208141	SRR3208142	SRR3208143	SRR3208132	SRR3208127	-
Gallus gallus (chicken)	*Gg*	ERR348580	ERR348571	ERR348563	ERR348583	ERR348561	ERR348586	ERR348578	ERR348585	SRR3194327
Anolis carolinensis (anole)	*Ac*	SRR5412173	SRR543710	SRR540258	SRR540257	SRR579557	SRR391653	SRR391658	-	-
Python molurus (python)	*Pm*	*SRA pending*	*SRA pending*	*SRA pending*	SRR5190732	SRR5190716	SRR5190690	*SRA pending*	*SRA pending*	SRR1746797
Boa constrictor (boa)	*Bc*	SRR7206966	-	SRR7206967	-	SRR7206969; SRR7206968	SRR7206971	SRR7206975	SRR7206963	SRR7206973; SRR7206972
Crotalus viridis (prairie rattlesnake)	*Cv*	*SRA pending*	SRR7401997	SRR7401995	SRR7402006	*SRA pending*	SRR7402007	SRR7401993	SRR7401998	*SRA pending*
Ornithorhynchus anatinus (platypus)	*Oa*	SRR5412241	SRR5412238	SRR5412224	SRR5412229	SRR5412232	SRR5412237	-	-	-
Monodelphis domestica (opossum)	*Md*	SRR500909	SRR500917	SRR500906	SRR500923	SRR500900	SRR500896	SRR500916	SRR868947	-
Homo sapiens (human)	*Hs*	ERR315492	ERR579132	ERR315432	ERR315356; ERR315430	ERR315383	ERR315414	ERR579149	ERR315448	ERR315408; ERR315364
Mus musculus (mouse)	*Mm*	SRR5047953	SRR5047990	SRR5048040	SRR5047921	SRR5047925	SRR5047934	SRR1158599	SRR7207813	SRR5048001

Supplementary File 2A. Within-species normalized expression values of genes involved in the PIWI:piRNA pathway. NA = missing, NiA = gene present, but absent in the genome annotation available. Purple highlights represent duplicated genes. T= testis; O= ovary; B= brain; H= heart; K= kidney; L= liver; M= muscle; S= spleen; SI= small intestine. Decimal values have been rounded to integers.

		piwil1	piwil2	PIWIL3	piwil4	mybl1	tdrd1	tdrd12	tdrd6	tdrd9	tdrkh	exd1	mael	pld6	Dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	Tex19	gpat2	btbd18
Danio rerio	T	9861	1656	*NA*	*NA*	562	1915	584	6304	2327	637	431	*NA*	1216	*NA*	5440	310	424	780	151	269	125	*NA*	1135	66
	O	754	741	*NA*	*NA*	9	1296	169	2091	686	577	660	*NA*	260	*NA*	1400	831	1435	6	341	441	465	*NA*	406	29
	B	2	3	*NA*	*NA*	27	0	4	10	3	231	7	*NA*	1	*NA*	0	7	15	4	73	20	433	*NA*	9	0
	H	5	3	*NA*	*NA*	17	3	3	10	4	109	1	*NA*	1	*NA*	1	6	16	1	52	8	150	*NA*	25	0
	K	22	26	*NA*	*NA*	66	30	9	55	24	64	10	*NA*	8	*NA*	30	20	120	27	163	12	86	*NA*	17	0
	L	0	12	*NA*	*NA*	6	18	0	18	18	23	0	*NA*	6	*NA*	12	6	18	12	198	0	82	*NA*	18	0
	M	485	146	*NA*	*NA*	50	98	35	245	135	96	33	*NA*	148	*NA*	290	33	44	31	81	22	122	*NA*	55	0
	S	125	39	*NA*	*NA*	18	61	5	81	43	71	12	*NA*	27	*NA*	106	16	27	3	78	9	121	*NA*	17	1
	SI	1	21	*NA*	*NA*	16	1	0	19	0	32	1	*NA*	0	*NA*	2	3	1	25	143	3	46	*NA*	1	1

		piwil1.L	piwil1.S	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrd12	tdrkh	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1	tex19	gpat2	btbd18
Xenopus laevis	T	1666	7	8577	*NA*	997	616	4342	1857	822	12	2720	1263	9219	2750	*NA*	3005	105	160	901	1581	*146*	775	*NA*	1705	*NA*
	O	5	297	878	*NA*	2	5	689	317	70	47	11584	510	2917	3378	*NA*	429	9	222	2576	4464	*22*	1844	*NA*	1348	*NA*
	B	0	1	3	*NA*	11	52	4	11	1	2	575	4	3	0	*NA*	74	0	21	622	621	*5*	190	*NA*	32	*NA*
	H	0	1	2	*NA*	2	38	1	2	2	4	288	0	4	0	*NA*	103	0	17	47	248	*0*	243	*NA*	8	*NA*
	K	1	48	70	*NA*	2	50	47	55	26	8	517	42	63	79	*NA*	190	0	86	188	454	*5*	201	*NA*	56	*NA*
	L	0	2	2	*NA*	0	18	5	7	0	2	900	0	0	7	*NA*	97	0	0	95	360	*2*	295	*NA*	5	*NA*
	M	2	0	0	*NA*	0	4	7	0	0	2	387	0	0	0	*NA*	399	0	121	59	288	*2*	169	*NA*	154	*NA*
	S	2	0	5	*NA*	0	28	21	3	2	0	402	2	2	1	*NA*	200	0	18	90	332	*5*	208	*NA*	14	*NA*
	SI	0	0	2	*NA*	1	27	3	20	0	1	306	32	0	0	*NA*	358	0	1	49	408	*1*	248	*NA*	11	*NA*

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Alligator mississippiensis	T	496	99	*NA*	102	511	324	848	208	82	106	94	152	66	0	778	29	165	27	28	71	32	*NA*	51	89
	O	212	75	*NA*	24	23	291	438	84	168	156	51	65	27	0	322	54	256	45	39	46	11	*NA*	74	90
	B	3	0	*NA*	2	12	1	20	16	14	19	8	1	27	0	208	1	1	21	9	16	30	*NA*	5	0
	H	0	2	*NA*	0	9	13	46	5	0	0	27	2	2	0	198	0	0	3	31	6	42	*NA*	0	0
	K	1	1	*NA*	1	3	4	20	7	0	20	107	4	13	0	125	0	0	9	9	11	42	*NA*	1	0
	L	0	0	*NA*	0	2	5	20	2	2	32	152	8	8	0	131	0	0	8	7	8	27	*NA*	2	3
	M	0	0	*NA*	4	11	0	8	4	0	8	16	0	14	0	142	0	0	0	11	12	55	*NA*	1	1
	S	3	0	*NA*	0	1	0	17	1	0	5	35	9	7	0	126	0	0	0	21	20	8	*NA*	5	3
	SI																								

Gallus gallus

	piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
T	1391	649	*NA*	*NA*	2901	86	3701	1164	1907	324	191	396	259	*NA*	703	94	667	112	*NA*	133	141	*NA*	710	137
O	478	341	*NA*	*NA*	450	149	2085	253	433	94	273	463	230	*NA*	58	56	120	13	*NA*	57	44	*NA*	245	94
B	222	41	*NA*	*NA*	15	10	6	80	327	131	28	31	1034	*NA*	0	4	117	16	*NA*	12	313	*NA*	68	20
H	15	46	*NA*	*NA*	8	3	49	33	45	15	4	8	3395	*NA*	1	1	3	2	*NA*	5	29	*NA*	154	54
K	1028	18	*NA*	*NA*	35	6	38	5	9	4	4	5	894	*NA*	4	41	5	3	*NA*	6	46	*NA*	73	55
L	15	73	*NA*	*NA*	0	2	549	6	21	2	7	56	967	*NA*	0	0	11	5	*NA*	1	21	*NA*	81	72
M	21	5	*NA*	*NA*	15	2	5	8	194	5	16	2	6114	*NA*	0	0	0	3	*NA*	18	23	*NA*	77	44
S	111	24	*NA*	*NA*	49	5	7	55	68	62	15	3	784	*NA*	3	1	77	49	*NA*	24	49	*NA*	117	62
SI	7	0	*NA*	*NA*	79	14	22	18	72	43	11	186	1438	*NA*	0	0	22	25	*NA*	4	194	*NA*	276	72

Anolis carolinensis

	piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
T	2658	1081	*NA*	1407	1695	502	3775	591	95	1838	665	1903	239	3	5736	831	2253	1163	1412	140	500	*NA*	65	473
O	1265	1581	*NA*	550	32	322	1823	665	96	104	535	1807	111	38	1238	933	916	55	677	73	170	*NA*	99	885
B	2	66	*NA*	17	41	45	39	4	36	38	108	21	37	0	13	13	13	21	250	0	383	*NA*	8	106
H	70	26	*NA*	4	75	3	7	2	5	15	206	6	65	0	47	2	5	0	608	2	284	*NA*	3	11
K	3	24	*NA*	0	8	23	8	1	3	29	31	85	22	0	224	2	6	10	1420	22	93	*NA*	5	60
L	10	8	*NA*	0	88	6	35	6	0	6	49	12	16	0	163	0	45	6	381	0	218	*NA*	10	53
M	0	0	*NA*	3	37	37	41	0	6	6	621	0	41	0	25	0	3	0	964	150	190	*NA*	3	106
S																								
SI																								

Boa constrictor

	piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
T	537	108	*NA*	173	79	229	413	297	143	160	107	170	18	0	457	70	307	13	316	109	60	*NA*	137	124
O																								
B	3	22	*NA*	0	22	0	68	0	20	27	12	0	0	0	12	2	8	5	163	36	103	*NA*	2	5
H																								
K	151	32	*NA*	41	30	60	103	115	61	50	46	38	7	0	99	13	113	4	283	41	105	*NA*	26	33
L	12	2	*NA*	85	18	12	12	0	31	21	41	0	1	0	150	0	6	2	174	20	115	*NA*	1	4
M	10	0	*NA*	0	13	0	37	0	7	27	0	0	0	0	3	0	13	0	1516	20	143	*NA*	0	3
S	37	2	*NA*	0	121	0	2	0	46	84	17	4	11	0	9	0	29	4	94	36	190	*NA*	3	1
SI	7	2	*NA*	0	49	3	9	3	63	20	15	1	2	0	51	3	0	0	326	10	132	*NA*	3	1

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Python molurus	T	415	31	*NA*	16	4	122	48	132	194	186	41	24	104	0	270	76	89	2	141	84	54	*NA*	189	34
	O	176	56	*NA*	7	5	63	123	82	160	29	192	79	13	0	205	58	97	9	153	37	34	*NA*	24	78
	B	0	5	*NA*	1	1	2	26	5	98	23	23	4	11	0	15	0	5	1	69	2	47	*NA*	3	1
	H	0	0	*NA*	3	3	8	12	2	5	67	38	7	2	0	18	0	3	0	67	0	43	*NA*	0	2
	K	0	0	*NA*	0	0	0	11	0	2	111	16	0	2	0	7	0	11	0	267	1	19	*NA*	2	0
	L	0	0	*NA*	0	0	0	0	0	4	82	8	0	12	0	8	0	8	0	196	0	37	*NA*	33	4
	M	0	0	*NA*	0	0	0	3	0	23	13	8	0	3	0	5	0	0	0	366	0	64	*NA*	0	0
	S	0	3	*NA*	2	1	0	4	0	11	19	5	1	15	0	85	0	1	0	90	0	63	*NA*	1	0
	SI	0	0	*NA*	11	0	0	22	0	0	1023	10	0	3	0	135	0	0	0	116	0	52	*NA*	0	0

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Crotalus viridis	T	1378	76	*NA*	136	24	99	306	293	412	170	8	212	NiA	1	636	107	299	140	NiA	155	14	*NA*	165	294
	O	544	177	*NA*	31	8	370	198	588	271	96	5	59	NiA	5	330	243	111	3	NiA	65	28	*NA*	73	111
	B	22	4	*NA*	1	4	0	6	0	140	30	4	0	NiA	0	6	0	15	12	NiA	45	27	*NA*	1	1
	H	1	0	*NA*	3	2	11	0	17	9	99	1	0	NiA	0	0	1	0	2	NiA	84	61	*NA*	30	2
	K	2	0	*NA*	0	0	57	3	6	11	52	2	2	NiA	0	0	0	0	0	NiA	84	13	*NA*	35	0
	L	0	0	*NA*	0	0	23	2	0	80	16	0	0	NiA	0	4	10	8	10	NiA	70	12	*NA*	53	0
	M	11	0	*NA*	0	0	6	22	0	0	101	28	0	NiA	0	0	0	0	0	NiA	39	6	*NA*	6	0
	S	2	0	*NA*	0	2	3	6	1	13	40	8	5	NiA	0	3	0	2	10	NiA	44	13	*NA*	41	0
	SI	1	2	*NA*	1	1	19	2	6	11	95	4	2	NiA	0	0	1	0	2	NiA	70	45	*NA*	29	1

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Ornithorhynchus anatinus	T	39092	373	*NA*	46	16532	24504	15707	6185	1235	14564	1891	5238	688	*NA*	21681	341	7182	1036	2922	*1619*	487	NA	*3587*	*157*
	O	413	111	*NA*	0	218	309	97	110	173	527	92	277	668	*NA*	409	199	174	532	727	*191*	69	NA	*235*	*3*
	B	76	734	*NA*	6	207	51	24	70	274	581	61	14	1741	*NA*	106	32	68	2264	1406	*94*	0	NA	*118*	*1*
	H	102	46	*NA*	5	154	441	51	124	54	1992	91	5	1031	*NA*	173	11	494	663	1257	*140*	2	NA	*37*	*0*
	K	20	64	*NA*	0	116	34	1935	55	5	961	84	16	716	*NA*	207	29	103	516	1107	*130*	0	NA	*97*	*5*
	L	192	132	*NA*	4	160	61	85	42	29	1044	151	22	338	*NA*	85	19	295	321	1814	*126*	3	NA	*129*	*3*
	M																								
	S																								
	SI																								

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Monodelphis domestica	T	2772	61	*NA*	63	706	1021	1167	1745	1271	384	144	3532	1073	4	3172	111	157	3961	3241	111	88	*NA*	141	210
	O	80	45	*NA*	1	71	48	38	218	150	145	256	572	2	1	254	46	71	25	647	423	105	*NA*	90	54
	B	7	0	*NA*	2	57	5	30	3	523	86	199	27	9	0	4	0	3	38	461	31	84	*NA*	2	47
	H	1	3	*NA*	1	39	1	32	13	84	30	71	0	1	0	0	1	0	19	446	84	35	*NA*	28	10
	K	5	1	*NA*	1	13	2	41	7	23	101	114	1	5	157	8	0	0	5	425	57	34	*NA*	37	17
	L	4	2	*NA*	0	0	2	179	9	9	222	53	6	2	6	11	2	0	11	361	68	40	*NA*	103	9
	M	0	0	*NA*	0	4	4	28	17	135	62	101	0	2	0	2	0	4	9	553	129	45	*NA*	0	19
	S	3	3	*NA*	2	72	3	9	23	105	96	71	4	20	3	4	0	15	1	993	184	28	*NA*	32	57
	SI																								

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Homo sapiens	T	2940	1357	43	397	493	815	1308	1963	972	472	489	2633	567	3	3414	276	945	865	786	1235	411	384	396	69
	O	4	38	1	72	34	17	21	6	42	8	0	1	215	0	0	0	25	16	999	100	1041	0	40	2
	B	67	26	0	35	105	6	33	44	349	14	15	51	133	0	0	0	40	287	509	218	317	0	71	3
	H	0	36	1	42	141	23	35	76	134	6	0	117	105	0	0	0	9	90	635	41	608	0	165	3
	K	3	28	0	25	21	14	17	165	86	3	0	61	78	43	1	0	12	17	722	28	333	0	50	0
	L	0	66	0	21	73	11	30	8	135	0	0	2	102	98	0	0	6	24	551	30	474	0	28	0
	M	0	9	0	21	26	4	26	13	209	9	0	13	64	0	0	0	0	68	1066	111	478	0	145	13
	S	20	9	0	68	261	32	60	41	191	26	0	1	54	10	1	0	3	865	364	147	553	1	409	3
	SI	2	91	0	144	121	5	17	21	86	63	0	1	33	3	0	0	3	19	446	157	478	0	41	0

		piwil1	piwil2	piwil3	piwil4	mybl1	tdrd1	tdrd6	tdrd9	tdrkh	tdrd12	exd1	mael	pld6	dnmt3l	ddx4	asz1	mov10l1	kif17	prmt5	henmt1	dcp1a	tex19	gpat2	btbd18
Mus musculus	T	15651	3296	*NA*	29	341	4432	7908	4085	7987	1435	468	1516	2302	409	3992	179	1939	1369	766	641	1098	987	2552	115
	O	8	52	*NA*	23	7	78	2	25	7	7	26	15	110	56	12	0	19	159	957	12	1040	28	67	2
	B	4	40	*NA*	3	66	8	28	23	458	23	37	0	31	3	72	0	2	478	960	19	590	7	19	2
	H	1	60	*NA*	2	7	3	3	6	44	7	19	7	34	5	1	0	1803	11	1140	0	676	0	10	0
	K	4	5	*NA*	11	12	0	18	2	22	128	33	2	43	1	4	1	1	16	1018	1	537	1	79	1
	L	0	11	*NA*	2	2	0	0	0	21	0	93	0	130	0	0	2	0	11	607	0	498	0	187	0
	M	1	9	*NA*	8	48	35	1	9	358	5	7	0	42	2	3	0	14	8	1171	1	321	3	6	1
	S	1	60	*NA*	0	40	0	0	14	150	2	2	0	1	0	0	0	0	0	2919	0	678	0	10	3
	SI	0	51	*NA*	16	0	5	0	13	3	82	16	0	45	19	0	3	0	8	394	0	599	3	8	3

Supplementary File 2B. Within-species normalized expression values of genes involved in the siRNA pathway. NA = missing, NiA = gene present, but absent in the genome annotation available. T= testis; O= ovary; B= brain; H= heart; K= kidney; L= liver; M= muscle; S= spleen; SI= small intestine. Decimal values have been rounded to integers.

		Dicer1	Drosha	Mrpl44	Prkra	Tarbp2	Tert	Ago1	Ago2	Ago3	Ago4
Danio rerio	T	514	299	111	49	66	275	14	346	191	165
	O	198	856	363	205	478	188	10	55	142	139
	B	516	264	166	187	145	8	114	412	146	605
	H	240	115	265	77	157	21	51	639	59	337
	K	243	155	268	88	205	64	12	356	101	285
	L	449	76	864	128	216	18	23	263	99	181
	M	242	77	356	63	131	133	13	184	133	177
	S	247	105	261	66	138	33	39	419	58	289
	SI	219	99	279	68	281	69	15	218	109	277
Xenopus laevis	T	56	327	432	2041	409	3	40	124	*NA*	101
	O	22	528	4105	3818	1410	45	27	27	*NA*	231
	B	16	501	217	705	296	8	33	21	*NA*	315
	H	47	315	216	472	148	11	17	72	*NA*	221
	K	34	229	305	367	277	27	24	51	*NA*	242
	L	85	127	291	443	395	5	35	102	*NA*	233
	M	7	227	1097	1105	198	7	2	26	*NA*	103
	S	51	226	112	317	397	13	50	123	*NA*	587
	SI	87	158	167	177	555	39	26	46	*NA*	362
Alligator mississippiensis	T	49	25	74	*NA*	31	74	68	19	83	132
	O	39	35	46	*NA*	9	390	14	71	26	40
	B	171	27	32	*NA*	12	7	63	6	10	102
	H	61	39	132	*NA*	24	2	71	0	0	107
	K	95	15	87	*NA*	8	48	31	2	2	108
	L	62	17	183	*NA*	28	2	15	13	7	62
	M	89	20	83	*NA*	13	5	61	8	7	186
	S	38	23	84	*NA*	55	4	39	12	5	141
	SI										
Gallus gallus	T	180	869	1420	*NA*	350	256	1312	*NA*	4974	5647
	O	316	1447	1188	*NA*	282	1272	335	*NA*	2041	1225
	B	336	415	378	*NA*	265	14	1389	*NA*	1276	1128
	H	95	246	2624	*NA*	476	5	938	*NA*	1739	1572
	K	250	1036	1819	*NA*	192	84	697	*NA*	1135	3921
	L	127	301	1770	*NA*	553	5	1004	*NA*	1416	2187
	M	360	481	2992	*NA*	741	0	2118	*NA*	2095	7554
	S	321	801	972	*NA*	191	80	1265	*NA*	1671	4770
	SI	143	165	882	*NA*	689	172	1151	*NA*	592	1535
Anolis carolinensis	T	847	105	110	412	781	615	512	68	156	365
	O	348	502	567	920	336	869	445	100	117	295
	B	1070	199	199	1384	146	254	357	135	121	342
	H	496	230	390	1889	99	12	561	156	120	392
	K	317	230	870	718	210	62	255	35	24	352
	L	370	110	625	1064	165	258	71	22	16	126
	M	583	162	558	2610	147	34	256	165	165	306
	S										
	SI										
Boa constrictor	T	217	150	47	74	47	35	75	20	25	128
	O										
	B	209	110	25	93	42	217	49	19	8	236
	H										
	K	293	141	36	82	154	94	26	18	10	79
	L	371	71	46	123	146	92	29	19	5	65
	M	343	270	50	187	373	240	123	17	0	137
	S	279	151	54	88	49	33	46	181	49	294
	SI	276	96	75	95	85	97	38	16	23	87

		Dicer1	Drosha	Mrpl44	Prkra	Tarbp2	Tert	Ago1	Ago2	Ago3	Ago4
Python molurus	T	67	8	19	0	102	9	32	7	2	37
	O	67	15	190	0	5	26	62	6	9	28
	B	81	9	20	0	47	5	35	5	2	31
	H	165	13	13	0	87	10	133	55	22	127
	K	51	28	22	0	73	39	11	38	0	89
	L	90	29	45	0	29	0	16	8	4	82
	M	95	5	46	0	180	0	46	3	0	62
	S	81	5	26	0	56	3	30	5	1	63
	SI	231	34	0	0	62	31	0	60	10	56
Crotalus viridis	T	131	72	31	27	61	46	10	5	13	119
	O	86	118	278	188	65	354	26	18	7	46
	B	142	58	34	96	193	6	28	3	6	45
	H	165	156	35	62	501	45	11	6	5	139
	K	67	101	84	33	567	10	13	6	13	27
	L	129	96	121	8	568	8	94	0	8	33
	M	78	89	168	34	324	0	11	0	6	45
	S	197	98	70	64	311	19	39	6	1	30
	SI	221	118	40	44	262	57	12	2	14	238
Ornithorhynchus anatinus	T	4851	437	4385	7241	2086	205	378	688	783	*NA*
	O	3182	851	1475	4087	256	124	349	371	419	*NA*
	B	7331	1097	1279	11278	375	43	838	249	803	*NA*
	H	5790	1250	2900	36073	315	4	889	152	133	*NA*
	K	2622	948	1648	4769	286	28	355	137	107	*NA*
	L	3294	788	2931	1766	485	25	341	143	133	*NA*
	M										
	S										
	SI										
Monodelphis domestica	T	331	489	98	648	*NA*	1500	471	68	148	834
	O	438	610	276	981	*NA*	32	414	694	21	370
	B	854	1292	201	1829	*NA*	53	1133	400	14	261
	H	668	360	708	4625	*NA*	29	332	422	19	845
	K	439	488	522	498	*NA*	31	254	335	33	368
	L	969	380	621	47	*NA*	11	275	512	40	529
	M	697	364	1481	32329	*NA*	4	214	637	6	4870
	S	726	470	342	79	*NA*	120	231	5025	268	925
	SI										
Homo sapiens	T	1359	579	346	636	229	21	447	523	329	621
	O	1890	967	277	526	119	0	791	932	189	577
	B	1651	1045	309	548	69	0	1005	819	270	426
	H	1165	579	427	493	135	0	801	713	184	1041
	K	1514	722	565	404	190	0	944	353	192	290
	L	1959	576	681	433	222	0	395	211	143	482
	M	844	1015	964	2222	290	0	917	1271	154	1173
	S	1834	543	280	287	189	3	822	781	182	847
	SI	1188	599	336	388	159	47	749	583	125	597
Mus musculus	T	466	858	102	528	7300	145	336	865	408	489
	O	801	590	69	377	1078	656	1041	1531	295	76
	B	1367	1954	211	296	639	82	1434	1035	1163	339
	H	1148	820	512	289	1412	5	931	2000	628	352
	K	990	840	306	342	853	166	1387	1427	388	97
	L	1309	294	269	397	1207	1585	1467	1356	353	134
	M	846	1141	1100	328	1326	15	90	4389	317	64
	S	519	720	665	71	1177	13	104	1327	212	1
	SI	524	519	120	450	1748	168	804	2363	184	48

Supplementary File 2C. Within-species normalized expression values of genes involved in transcriptional regulation of TE expression. NA = missing, NiA = gene present, but absent in the genome annotation available. T= testis; O= ovary; B= brain; H= heart; K= kidney; L= liver; M= muscle; S= spleen; SI= small intestine. Decimal values have been rounded to integers. Prmts = sum of normalized expression values for *prmt1*, *prmt2*, *prmt3*, *prmt6*, *prmt7*, *prmt8*, *prmt9* (when present).

		dnmt1	dnmt3a	dnmt3b	ehmt1	ehmt2	hdac1	hdac2	hells	kdm1a	marf1	p53	prmts	rybp	set	setdb1	setdb2	suv39h	trim28	zfp90	zfp91
Danio rerio	T	956	171	1129	388	140	1176	210	678	1921	239	348	1412	412	3905	1642	142	256	*NA*	*NA*	133
	O	6661	13	994	639	404	2511	647	1124	700	128	246	2226	891	5415	1991	197	542	*NA*	*NA*	143
	B	338	1088	147	582	148	1347	285	41	1072	217	120	1950	730	3436	709	67	48	*NA*	*NA*	181
	H	149	183	92	334	150	1359	170	15	524	287	402	2434	190	3480	470	95	41	*NA*	*NA*	150
	K	729	165	362	420	141	1447	164	96	636	234	386	1686	159	3533	398	105	141	*NA*	*NA*	91
	L	198	117	152	198	76	1004	82	12	356	158	292	2930	64	3706	321	181	29	*NA*	*NA*	53
	M	184	100	98	214	50	1000	124	55	502	494	308	1339	151	3700	299	125	264	*NA*	*NA*	101
	S	266	168	150	405	139	1300	168	40	491	281	369	1919	151	2681	325	61	86	*NA*	*NA*	119
	SI	248	105	227	338	95	1144	185	81	529	164	549	1247	173	3030	281	62	59	*NA*	*NA*	77
Xenopus laevis	T	1743	*NA*	*NA*	495	1440	611	348	67	287	290	1169	6974	60	2346	1090	16	559	2339	*NA*	3991
	O	3524	*NA*	*NA*	382	2969	2345	541	384	326	261	16333	25599	227	2295	1577	16	9338	2955	*NA*	2434
	B	389	*NA*	*NA*	1207	918	460	894	19	565	1552	615	6415	53	2497	1176	30	61	322	*NA*	1984
	H	131	*NA*	*NA*	688	935	496	339	20	346	1151	1831	4334	41	3560	586	40	46	379	*NA*	1582
	K	583	*NA*	*NA*	681	1142	859	353	81	285	942	1508	3292	111	2562	658	36	263	461	*NA*	1595
	L	129	*NA*	*NA*	609	1027	787	78	21	330	667	2289	3216	55	4449	561	25	99	556	*NA*	3085
	M	207	*NA*	*NA*	423	1315	685	771	13	775	874	506	6481	60	4290	526	40	158	438	*NA*	3246
	S	456	*NA*	*NA*	1339	1371	1216	468	70	489	1557	1431	2994	86	4916	744	57	157	627	*NA*	1814
	SI	211	*NA*	*NA*	774	1432	1200	197	21	406	1163	1397	2557	54	4523	536	54	66	626	*NA*	2369
Alligator mississippiensis	T	105	14	12	58	30	59	102	117	121	125	56	318	94	538	77	58	128	320	*NA*	56
	O	146	8	13	26	219	54	172	160	350	18	28	730	27	369	118	128	41	756	*NA*	45
	B	56	49	10	75	69	73	113	10	221	156	19	396	145	417	57	45	40	253	*NA*	15
	H	38	42	5	97	20	42	159	6	50	135	42	465	57	600	50	50	33	204	*NA*	14
	K	44	48	14	94	26	99	117	10	175	122	33	336	80	413	108	81	28	158	*NA*	17
	L	44	18	35	77	25	89	100	12	50	111	146	300	59	410	111	107	25	102	*NA*	23
	M	83	13	30	95	28	150	79	73	79	172	113	374	55	442	64	133	44	142	*NA*	23
	S	151	74	64	63	92	130	79	47	145	75	481	328	50	595	180	146	29	234	*NA*	27
	SI																				

		dnmt1	dnmt3a	dnmt3b	ehmt1	ehmt2	hdac1	hdac2	hells	kdm1a	marf1	p53	prmts	rybp	set	setdb1	setdb2	suv39h	trim28	zfp90	zfp91
Anolis carolinensis	T	697	70	138	468	113	669	448	466	764	1516	NiA	2743	1600	*NA*	304	585	2660	1586	NiA	1226
	O	915	332	23	762	1165	499	1322	953	501	252	NiA	3072	299	*NA*	706	403	873	1959	NiA	1939
	B	298	464	10	454	118	311	1017	20	678	477	NiA	3108	1178	*NA*	145	455	91	1504	NiA	479
	H	562	282	8	538	140	489	469	24	487	415	NiA	2305	434	*NA*	162	475	62	1377	NiA	511
	K	516	471	5	482	216	666	156	12	649	286	NiA	2167	140	*NA*	319	312	11	1995	NiA	268
	L	197	69	18	244	177	295	299	96	338	224	NiA	2067	804	*NA*	41	753	205	692	NiA	405
	M	187	128	0	284	78	187	156	19	265	284	NiA	3350	427	*NA*	56	477	69	521	NiA	998
	S																				
	SI																				
Boa constrictor	T	346	107	78	231	362	364	169	569	322	115	571	831	33	1500	214	3	68	692	*NA*	89
	O																				
	B	129	176	7	265	87	195	171	58	126	207	170	862	53	855	88	0	12	370	*NA*	129
	H																				
	K	155	110	90	114	169	263	246	122	232	139	629	2538	26	860	233	0	24	361	*NA*	99
	L	127	117	49	94	196	343	112	35	218	153	1177	2308	25	644	293	0	24	249	*NA*	110
	M	237	47	7	110	63	60	493	13	127	177	100	8855	47	1000	17	0	47	143	*NA*	170
	S	513	194	64	179	251	312	118	43	145	449	267	562	67	805	187	5	29	334	*NA*	199
	SI	207	90	17	77	63	258	256	211	135	143	872	2076	39	1549	62	1	50	243	*NA*	105
Python molurus	T	141	52	41	34	19	3	147	114	136	15	38	720	70	1049	128	25	41	431	*NA*	15
	O	105	17	12	55	36	4	113	247	103	81	5	711	28	287	369	1	36	267	*NA*	239
	B	133	107	13	21	18	7	114	11	68	105	17	630	50	486	127	50	4	260	*NA*	21
	H	314	162	5	112	18	13	78	13	58	92	18	1149	53	475	68	75	15	270	*NA*	52
	K	244	123	43	95	24	24	105	0	27	108	9	1386	24	1005	57	34	9	203	*NA*	49
	L	127	86	53	119	45	0	16	20	82	82	29	1333	37	478	98	8	16	164	*NA*	119
	M	147	41	0	31	10	0	183	0	44	23	33	3712	8	1238	64	21	10	98	*NA*	33
	S	181	112	112	31	16	9	104	41	113	30	100	1187	31	830	154	54	29	354	*NA*	26
	SI	540	36	9	73	10	225	117	19	25	24	32	773	43	1276	18	27	56	198	*NA*	57
Crotalus viridis	T	339	134	3	336	42	187	88	115	529	57	8	1096	35	1186	370	25	154	792	*NA*	154
	O	257	42	1	398	84	182	225	125	378	38	4	1632	26	779	588	119	134	319	*NA*	315
	B	172	366	6	391	19	151	172	27	190	121	1	1284	10	1103	79	69	10	563	*NA*	60
	H	195	133	12	126	53	240	120	18	119	219	35	6066	22	829	126	532	37	594	*NA*	50
	K	87	143	2	171	10	189	380	14	116	43	19	4739	13	1312	92	235	16	282	*NA*	24
	L	105	219	0	107	29	246	154	2	150	51	25	3309	31	1268	263	379	76	519	*NA*	21
	M	240	190	11	62	11	151	403	28	129	34	0	5828	6	884	45	246	6	235	*NA*	39
	S	313	602	6	201	19	734	275	33	220	35	117	1596	2	1886	244	412	43	526	*NA*	31
	SI	232	173	14	192	27	235	100	12	144	221	73	3921	11	865	139	298	36	669	*NA*	48

		dnmt1	dnmt3a	dnmt3b	ehmt1	ehmt2	hdac1	hdac2	hells	kdm1a	marf1	p53	prmts	rybp	set	setdb1	setdb2	suv39h	trim28	zfp90	zfp91
Ornithorhynchus anatinus	T	401	1748	224	2092	1877	1193	1224	1129	4130	5640	169	3403	437	6438	4111	1391	1293	2091	*NA*	3808
	O	1201	2318	377	1738	1673	1264	1342	462	2167	3336	755	7619	638	5744	947	758	615	1515	*NA*	2719
	B	551	2087	94	1563	1171	802	1954	331	2717	3796	208	8573	658	7604	438	1063	435	1737	*NA*	3946
	H	1477	3646	37	1740	1159	1861	1309	114	3583	5076	389	8611	296	7708	733	1166	700	1220	*NA*	4589
	K	804	1732	60	1229	1515	1731	1407	78	2089	4525	474	4304	421	5783	1074	932	553	1702	*NA*	2926
	L	573	1508	57	1573	1215	1743	817	324	2275	2713	589	5227	28	7056	831	935	862	2199	*NA*	2857
	M																				
	S																				
	SI																				
Monodelphis domestica	T	1675	87	23	394	318	1412	353	476	1239	649	52	3844	571	2521	453	375	465	*NA*	*NA*	6279
	O	1268	269	148	773	742	626	1113	752	505	853	306	1695	190	3386	245	251	105	*NA*	*NA*	1054
	B	439	242	9	485	924	247	1318	183	667	1581	58	1992	243	4135	262	117	24	*NA*	*NA*	2093
	H	430	149	3	466	380	514	798	96	446	1052	139	1795	94	2808	215	174	7	*NA*	*NA*	1768
	K	650	129	26	737	831	523	925	151	838	1062	163	1289	146	2922	168	226	12	*NA*	*NA*	2098
	L	668	235	0	758	662	553	606	450	474	909	160	1027	188	3097	213	278	24	*NA*	*NA*	2226
	M	358	144	6	454	315	607	1706	101	821	1072	77	2545	84	2804	135	141	4	*NA*	*NA*	2542
	S	1541	225	7	478	381	1293	1582	808	657	1060	342	1059	9	2705	203	895	1055	*NA*	*NA*	1656
	SI																				
Homo sapiens	T	1468	230	183	1233	860	921	1484	459	2498	967	302	2994	1150	1884	984	408	697	3068	420	4970
	O	323	448	33	545	888	1646	1278	116	1458	2243	608	4323	1278	3826	941	506	186	2615	870	1398
	B	613	248	10	426	1101	293	1001	19	832	2147	180	4271	1384	4145	290	139	202	1219	707	1927
	H	395	306	40	520	478	772	693	9	751	1462	244	4757	591	2389	390	299	156	1471	351	1721
	K	349	258	57	573	600	1062	925	61	700	1862	292	2808	730	3389	567	194	149	1509	417	1543
	L	444	235	60	638	282	1210	815	162	598	1065	421	2262	1146	4198	532	198	258	867	380	2164
	M	473	1019	26	853	422	499	985	26	1433	1731	179	6598	209	3838	695	60	162	3211	252	3906
	S	1069	422	30	1137	934	1534	684	87	685	1818	739	4110	1022	2811	798	363	208	1919	448	1413
	SI	683	280	24	632	653	2066	826	117	857	1274	445	3575	779	2659	548	226	181	1650	198	2142
Mus musculus	T	1219	462	204	524	13056	3663	223	203	1685	1017	632	5431	885	985	1424	70	688	12587	67	3474
	O	2236	1746	193	2030	7909	752	268	56	889	960	3261	7161	456	2441	763	14	740	9093	109	650
	B	1422	2407	38	1192	4671	670	1126	65	896	2088	457	4947	1025	2687	1151	34	544	2730	273	1850
	H	687	2095	65	1352	1849	676	477	10	732	1617	1015	3779	380	1147	647	45	219	2805	96	2405
	K	746	1159	81	977	3616	867	240	8	567	903	1036	2843	478	1151	622	40	285	2872	71	1125
	L	521	526	231	1633	4948	742	181	8	429	2104	1116	2499	307	1625	742	95	425	3597	17	1835
	M	807	3002	122	1063	1015	1343	2781	202	573	697	872	6938	189	4975	573	67	378	1845	330	841
	S	6285	791	393	640	732	6397	2264	1896	1274	153	3716	8117	251	10763	650	261	1968	2956	132	994
	SI	1256	851	69	1634	7256	663	237	24	476	732	1447	3262	277	1610	442	16	383	4441	56	559

Supplementary File 2D. Within-species normalized expression values of genes involved in post-transcriptional regulation of TEs. NA = missing, NiA = gene present, but absent in the genome annotation available. T= testis; O= ovary; B= brain; H= heart; K= kidney; L= liver; M= muscle; S= spleen; SI= small intestine. Decimal values have been rounded to integers.

		apobec	aicda	atg5	atm	becn1	calcoco2	dcp1b	dcp2	dcps	hnrpl	rnasel	samhd1	trex1	zc3hav1	zfp36	zfp36l1	zfp36l2
Danio rerio	T	19	1	346	508	41	72	174	177	125	502	*NA*	119	*NA*	*NA*	*NA*	686	65
	O	2	3	607	202	58	4	78	638	365	376	*NA*	313	*NA*	*NA*	*NA*	756	350
	B	9	0	301	137	263	36	55	752	125	509	*NA*	47	*NA*	*NA*	*NA*	902	306
	H	1013	1	179	117	102	190	179	577	67	436	*NA*	59	*NA*	*NA*	*NA*	2103	699
	K	5	3	239	197	59	142	105	353	108	562	*NA*	52	*NA*	*NA*	*NA*	4118	352
	L	0	0	303	163	82	146	228	193	93	537	*NA*	23	*NA*	*NA*	*NA*	5089	292
	M	2383	2	175	142	70	1476	142	338	52	452	*NA*	44	*NA*	*NA*	*NA*	1920	1350
	S	0	1	212	135	68	146	139	502	98	470	*NA*	34	*NA*	*NA*	*NA*	5794	319
	SI	0	9	271	180	86	388	192	209	125	663	*NA*	25	*NA*	*NA*	*NA*	3443	721
Xenopus laevis	T	2	1	411	910	88	350	3623	891	206	*NA*	910	501	866	*NA*	350	395	896
	O	0	0	543	121	31	458	3335	146	717	*NA*	1240	323	649	*NA*	38	23	17750
	B	35	0	72	261	81	1043	280	241	69	*NA*	583	314	104	*NA*	536	501	1138
	H	85	0	194	178	124	2222	109	114	50	*NA*	513	481	125	*NA*	5391	1638	3007
	K	37	1	134	203	89	1651	157	222	74	*NA*	556	274	166	*NA*	5447	2121	3041
	L	25	0	185	238	108	932	113	108	125	*NA*	321	330	194	*NA*	5969	2910	3039
	M	40	4	359	147	176	6159	7	75	35	*NA*	1112	317	269	*NA*	2810	720	3442
	S	38	4	149	485	147	2021	10	285	72	*NA*	590	1123	206	*NA*	4856	4581	4348
	SI	49	1	183	530	177	3873	18	161	62	*NA*	649	359	206	*NA*	3405	1231	3900
Alligator mississippiensis	T	1	1	51	198	118	109	56	21	28	182	28	61	*NA*	43	NiA	87	5
	O	0	0	77	100	58	86	154	32	14	79	2	12	*NA*	7	NiA	110	2
	B	7	4	34	114	29	47	25	22	14	87	18	108	*NA*	14	NiA	30	2
	H	314	0	31	60	69	327	9	20	42	93	6	124	*NA*	30	NiA	121	13
	K	0	0	19	66	29	183	36	23	60	132	13	98	*NA*	35	NiA	167	7
	L	0	2	50	54	44	166	37	27	32	90	12	127	*NA*	54	NiA	392	15
	M	25	1	50	66	47	86	42	29	43	109	8	132	*NA*	23	NiA	313	14
	S	0	8	32	24	36	122	86	5	50	347	12	117	*NA*	7	NiA	434	5
	SI																	
Gallus gallus	T	58	18	285	332	364	159	375	133	425	1103	29	90	523	393	*NA*	944	209
	O	7	7	128	112	853	553	244	92	593	388	115	251	440	236	*NA*	2687	543
	B	0	0	577	499	1316	693	255	214	183	1824	143	252	360	323	*NA*	604	711
	H	0	2	75	311	1272	156	55	26	725	107	78	361	572	53	*NA*	6497	1716
	K	0	0	152	403	842	444	122	45	576	161	1063	100	404	81	*NA*	19725	901
	L	0	1	61	498	943	624	88	29	488	148	396	112	681	627	*NA*	15546	2348
	M	0	0	102	357	2013	748	352	46	103	767	21	182	223	31	*NA*	4732	4465
	S	1	32	85	499	717	415	221	112	920	283	315	553	1267	325	*NA*	21842	2493
	SI	0	4	430	369	384	1495	115	7	298	997	158	506	674	441	*NA*	3658	7036
Anolis carolinensis	T	185	57	502	168	1450	298	317	401	638	1535	21	358	10	*NA*	0	NiA	303
	O	1	0	676	44	753	230	163	573	337	1316	0	1152	22	*NA*	0	NiA	254
	B	45	3	436	108	442	441	80	309	152	1881	13	236	16	*NA*	1	NiA	199
	H	42	1	358	83	413	2132	126	207	85	1235	48	1470	170	*NA*	6	NiA	601
	K	53	0	184	66	394	474	146	40	80	2466	16	193	27	*NA*	2	NiA	1090
	L	20	10	279	171	409	1953	63	185	118	2143	53	718	362	*NA*	10	NiA	332
	M	41	3	605	56	1126	10173	94	112	50	1173	12	196	153	*NA*	0	NiA	337
	S																	
	SI																	
Boa constrictor	T	71	0	32	150	79	236	58	163	37	1074	35	298	*NA*	*NA*	47	174	174
	O																	
	B	2	0	17	68	166	424	64	239	15	438	41	124	*NA*	*NA*	20	93	93
	H																	
	K	5	0	14	81	108	530	67	175	44	1595	25	145	*NA*	*NA*	107	158	158
	L	20	0	8	64	199	741	49	216	38	2295	44	191	*NA*	*NA*	1051	1343	1343
	M	13	0	0	67	93	910	213	180	60	633	10	137	*NA*	*NA*	93	47	47
	S	155	0	27	236	618	390	62	267	7	826	118	562	*NA*	*NA*	65	303	303
	SI	267	0	30	77	106	532	68	285	41	741	121	280	*NA*	*NA*	102	438	438

		apobec	aicda	atg5	atm	becn1	calcoco2	dcp1b	dcp2	dcps	hnrpl	rnasel	samhd1	trex1	zc3hav1	zfp36	zfp36l1	zfp36l2
Python molurus	T	12	0	75	91	55	54	24	19	102	474	79	279	*NA*	*NA*	16	159	98
	O	0	0	243	19	89	104	11	32	196	99	4	285	*NA*	*NA*	19	297	36
	B	1	0	42	61	103	182	20	23	30	317	135	63	*NA*	*NA*	297	139	38
	H	3	0	17	23	67	185	13	48	47	317	325	23	*NA*	*NA*	135	204	92
	K	0	0	58	72	35	50	27	42	4	344	183	11	*NA*	*NA*	12	312	63
	L	0	0	12	49	78	49	8	37	25	446	147	8	*NA*	*NA*	944	576	49
	M	0	0	90	33	121	332	10	10	59	185	62	8	*NA*	*NA*	136	64	111
	S	19	0	49	90	77	178	28	27	43	333	166	161	*NA*	*NA*	1316	909	155
	SI	0	0	8	54	191	174	5	67	16	114	607	37	*NA*	*NA*	30	365	82
Crotalus viridis	T	69	3	82	69	290	63	1296	60	48	468	20	99	*NA*	*NA*	NiA	148	188
	O	70	0	59	37	112	26	74	195	42	154	58	350	*NA*	*NA*	NiA	714	50
	B	7	1	142	37	213	57	49	58	72	345	49	57	*NA*	*NA*	NiA	211	90
	H	16	3	74	21	170	296	53	86	20	290	64	15	*NA*	*NA*	NiA	673	243
	K	5	0	119	11	159	136	49	38	27	287	55	10	*NA*	*NA*	NiA	382	92
	L	16	0	70	6	96	160	10	12	55	736	33	25	*NA*	*NA*	NiA	2433	82
	M	201	6	117	22	358	727	22	17	50	252	682	11	*NA*	*NA*	NiA	341	50
	S	12	32	174	69	109	55	38	25	44	594	272	13	*NA*	*NA*	NiA	1654	142
	SI	19	2	84	31	136	272	71	86	31	360	47	13	*NA*	*NA*	NiA	548	180
Ornithorhynchus anatinus	T	2	0	877	233	2060	1873	479	911	*NA*	3017	71	1314	32	*NA*	20	857	28
	O	69	0	623	213	2148	3132	250	206	*NA*	2017	220	1077	146	*NA*	168	6318	124
	B	36	0	743	170	3009	2554	225	644	*NA*	2292	79	1965	60	*NA*	40	2051	24
	H	14	0	532	170	6639	5498	949	270	*NA*	2382	58	737	30	*NA*	275	2225	82
	K	16	0	530	166	3507	3407	257	146	*NA*	2085	187	482	120	*NA*	239	4464	107
	L	18	0	753	164	3957	3766	175	191	*NA*	1622	193	543	24	*NA*	282	8361	33
	M																	
	S																	
	SI																	
Monodelphis domestica	T	30	2	346	115	751	1510	424	145	120	1444	7	282	11	370	43	459	94
	O	28	10	275	295	384	1430	24	448	228	1176	90	556	28	1267	102	1950	450
	B	1	1	251	265	770	1365	42	404	181	1616	43	387	20	415	20	465	82
	H	1	1	231	216	681	2435	395	180	176	1051	58	623	48	467	315	1022	538
	K	0	1	391	206	524	2611	47	157	314	1098	74	334	38	577	183	1421	289
	L	0	0	553	307	440	2388	38	203	220	914	342	591	226	1020	273	3752	531
	M	0	0	375	103	954	3672	66	105	272	1053	49	827	51	264	570	924	311
	S	0	73	645	676	1006	1254	332	381	115	1095	502	766	2	1523	49	1089	1
	SI																	
Homo sapiens	T	423	5	339	801	528	2048	587	526	190	2407	94	1017	109	516	564	1712	937
	O	468	0	586	219	650	2332	357	752	167	2295	586	1993	38	834	5523	7805	5976
	B	44	0	463	929	871	908	119	948	146	1902	229	621	87	537	444	2153	901
	H	84	0	370	100	952	2763	90	288	258	2128	235	1224	132	383	2912	2741	2370
	K	62	0	425	760	747	1568	201	587	240	2341	312	562	104	669	1194	3639	1725
	L	94	0	747	965	833	2818	124	674	706	2534	196	954	49	636	3951	10393	2881
	M	55	0	414	269	1975	3254	175	405	635	3075	90	746	209	149	708	1241	1006
	S	493	34	350	277	778	1921	214	107	279	2820	350	3667	254	1147	1951	4648	4474
	SI	125	20	740	841	994	1757	196	427	143	2620	408	1789	125	630	3423	2568	4065
Mus musculus	T	1429	1	101	244	1773	86	3721	248	766	3847	92	246	104	170	180	1260	350
	O	1049	8	150	98	859	13	720	211	583	3701	105	846	1406	1217	1293	7342	7151
	B	602	0	310	455	1316	1	482	940	244	3698	390	396	390	120	70	458	511
	H	953	1	299	367	737	0	210	544	389	3792	145	780	642	612	826	1594	689
	K	1092	0	366	260	863	0	344	389	511	3082	123	821	553	588	789	3138	1391
	L	290	0	118	153	896	0	151	387	111	3077	44	553	782	2136	3409	8650	1738
	M	275	0	702	264	1206	0	127	262	396	2839	388	415	235	512	372	2019	385
	S	9358	0	160	254	3580	0	329	442	133	5915	298	5661	1543	2725	2287	384	5551
	SI	881	8	287	51	1892	0	367	138	391	2754	588	1828	1623	1671	6508	3395	12122

Supplementary File 3. Total normalized expression values of genes involved in the PIWI pathway, the siRNA pathway and other genes that negatively regulate TE expression.

Total normalized expression values are reported for each tissue by species.

		Recent-TE %	PIWI pth %	siRNA pth %	Transcription %	Post-transcription %	Total regulators %
Danio rerio	T	1.94	0.74	0.04	0.35	0.06	1.19
	O	0.08	0.18	0.04	0.40	0.05	0.67
	B	0.94	0.01	0.04	0.20	0.05	0.30
	H	0.47	0.00	0.02	0.11	0.06	0.19
	K	0.14	0.01	0.01	0.08	0.05	0.15
	L	0.04	0.00	0.00	0.02	0.01	0.03
	M	0.11	0.01	0.01	0.04	0.04	0.10
	S	0.34	0.01	0.02	0.09	0.08	0.19
	SI	0.28	0.00	0.01	0.08	0.06	0.16
Xenopus laevis	T	0.02	0.20	0.02	0.11	0.05	0.37
	O	0.01	0.14	0.04	0.31	0.12	0.61
	B	0.05	0.01	0.01	0.13	0.03	0.19
	H	0.03	0.00	0.01	0.06	0.05	0.11
	K	0.03	0.01	0.01	0.09	0.08	0.20
	L	0.02	0.00	0.00	0.04	0.03	0.08
	M	0.01	0.00	0.01	0.04	0.03	0.08
	S	0.05	0.01	0.01	0.12	0.11	0.25
	SI	0.03	0.01	0.01	0.07	0.06	0.14
Alligator mississippiensis	T	0.06	0.40	0.05	0.22	0.09	0.76
	O	0.01	0.21	0.06	0.28	0.06	0.61
	B	0.03	0.03	0.03	0.15	0.04	0.25
	H	0.02	0.01	0.01	0.07	0.04	0.13
	K	0.03	0.02	0.02	0.12	0.05	0.22
	L	0.02	0.01	0.01	0.05	0.03	0.11
	M	0.02	0.02	0.03	0.15	0.07	0.28
	S	0.03	0.02	0.03	0.21	0.09	0.35
	SI						
Gallus gallus	T	0.05	0.05	0.05	0.12	0.02	0.23
	O	0.08	0.02	0.02	0.12	0.02	0.18
	B	0.01	0.01	0.02	0.07	0.03	0.12
	H	0.02	0.01	0.01	0.07	0.02	0.10
	K	0.02	0.00	0.01	0.04	0.03	0.08
	L	0.01	0.00	0.01	0.03	0.03	0.07
	M	0.00	0.00	0.01	0.03	0.01	0.05
	S	0.05	0.00	0.03	0.11	0.09	0.24
	SI	0.75	0.00	0.01	0.03	0.02	0.06
Anolis carolinensis	T	0.25	0.13	0.02	0.07	0.03	0.25
	O	0.02	0.11	0.04	0.13	0.04	0.32
	B	0.23	0.01	0.02	0.06	0.03	0.12
	H	0.19	0.01	0.02	0.04	0.03	0.10
	K	0.23	0.01	0.01	0.04	0.02	0.09
	L	0.10	0.00	0.01	0.02	0.02	0.06
	M	0.07	0.00	0.01	0.01	0.02	0.05
	S						
	SI						
Boa constrictor	T	0.03	0.11	0.02	0.18	0.07	0.39
	O						
	B	0.02	0.01	0.02	0.07	0.03	0.12
	H						
	K	0.03	0.03	0.02	0.14	0.07	0.26
	L	0.03	0.01	0.01	0.07	0.09	0.18
	M	0.00	0.01	0.01	0.04	0.01	0.06
	S	0.05	0.02	0.03	0.11	0.09	0.24
	SI	0.01	0.01	0.02	0.12	0.06	0.21

		Recent-TE %	PIWI pth %	siRNA pth %	Transcription %	Post-transcription %	Total regulators %
Python molurus	T	0.08	0.17	0.02	0.24	0.11	0.54
	O	0.01	0.13	0.03	0.21	0.11	0.47
	B	0.09	0.02	0.01	0.13	0.09	0.25
	H	0.10	0.01	0.02	0.09	0.05	0.16
	K	0.17	0.02	0.02	0.17	0.06	0.26
	L	0.08	0.01	0.01	0.08	0.06	0.16
	M	0.01	0.01	0.01	0.07	0.01	0.09
	S	0.07	0.02	0.02	0.23	0.23	0.49
	SI	0.21	0.06	0.02	0.15	0.07	0.31
Crotalus viridis	T	0.21	0.22	0.02	0.26	0.13	0.64
	O	0.06	0.22	0.07	0.34	0.12	0.74
	B	0.26	0.02	0.03	0.23	0.07	0.35
	H	0.27	0.01	0.04	0.35	0.07	0.47
	K	0.14	0.01	0.03	0.23	0.04	0.30
	L	0.17	0.01	0.02	0.12	0.07	0.21
	M	0.15	0.00	0.01	0.16	0.05	0.23
	S	0.25	0.01	0.04	0.35	0.15	0.55
	SI	0.34	0.01	0.04	0.31	0.08	0.45
Ornithorhynchus anatinus	T	0.19	0.68	0.08	0.16	0.05	0.97
	O	0.06	0.03	0.05	0.18	0.10	0.37
	B	0.07	0.03	0.08	0.14	0.05	0.30
	H	0.11	0.01	0.10	0.10	0.05	0.26
	K	0.27	0.03	0.04	0.14	0.08	0.28
	L	0.09	0.01	0.02	0.06	0.04	0.14
	M						
	S						
	SI						
Monodelphis domestica	T	0.10	0.23	0.04	0.20	0.07	0.54
	O	0.16	0.05	0.06	0.23	0.14	0.49
	B	0.21	0.02	0.06	0.16	0.07	0.31
	H	0.05	0.00	0.04	0.06	0.04	0.15
	K	0.10	0.01	0.02	0.11	0.07	0.21
	L	0.09	0.01	0.02	0.06	0.06	0.14
	M	0.04	0.00	0.17	0.06	0.04	0.27
	S	0.25	0.02	0.12	0.23	0.15	0.52
	SI						
Homo sapiens	T	0.03	0.23	0.05	0.27	0.13	0.68
	O	0.02	0.02	0.05	0.21	0.27	0.55
	B	0.01	0.01	0.04	0.14	0.07	0.26
	H	0.01	0.01	0.02	0.08	0.08	0.19
	K	0.01	0.01	0.03	0.11	0.09	0.25
	L	0.01	0.01	0.02	0.06	0.09	0.17
	M	0.00	0.01	0.02	0.06	0.03	0.12
	S	0.03	0.03	0.05	0.21	0.25	0.54
	SI	0.02	0.01	0.04	0.15	0.16	0.36
Mus musculus	T	0.24	0.38	0.04	0.21	0.06	0.69
	O	0.15	0.02	0.04	0.27	0.18	0.51
	B	0.39	0.02	0.04	0.17	0.06	0.29
	H	0.17	0.01	0.02	0.07	0.04	0.14
	K	0.14	0.01	0.03	0.09	0.07	0.20
	L	0.10	0.00	0.02	0.06	0.06	0.14
	M	0.49	0.01	0.04	0.12	0.04	0.22
	S	0.09	0.02	0.03	0.26	0.21	0.52
	SI	0.15	0.00	0.02	0.07	0.10	0.20

Summary statics of TE regulators expression values by species (top) and for the whole dataset (bottom).
NM = Non-mammal species; M = Mammal species.

	PIWI		**siRNA**		**Transcription**		**Post-Transcription**		**Total Regulators**	
	Avg Germline	*Avg Somatic*	*Avg Germline*	*Avg Somatic*	*Avg Germline*	*Avg Somatic*	*Avg Germline*	*Avg Somatic*	*Avg Germline*	*Avg Somatic*
Danio rerio	0.459	0.006	0.041	0.016	0.373	0.089	0.057	0.048	0.930	0.159
Xenopus laevis	0.167	0.007	0.030	0.008	0.210	0.078	0.082	0.057	0.489	0.150
Alligator mississippiensis	0.305	0.019	0.053	0.023	0.250	0.127	0.076	0.054	0.684	0.223
Gallus gallus	0.034	0.004	0.036	0.014	0.119	0.054	0.020	0.032	0.207	0.103
Anolis carolinensis	0.123	0.006	0.027	0.015	0.101	0.036	0.037	0.026	0.289	0.084
Boa constrictor	0.110	0.014	0.022	0.016	0.182	0.090	0.072	0.058	0.386	0.179
Python molurus	0.147	0.021	0.026	0.014	0.222	0.131	0.111	0.082	0.506	0.247
Crotalus viridis	0.224	0.011	0.048	0.030	0.300	0.250	0.125	0.076	0.686	0.366
Ornithorhynchus anatinus	0.353	0.019	0.068	0.061	0.173	0.108	0.077	0.056	0.671	0.245
Monodelphis domestica	0.144	0.011	0.053	0.072	0.215	0.112	0.105	0.071	0.518	0.266
Homo sapiens	0.127	0.013	0.051	0.032	0.240	0.114	0.197	0.112	0.615	0.270
Mus musculus	0.198	0.010	0.042	0.029	0.240	0.120	0.119	0.083	0.603	0.244

		Testis	Ovary	Germline	Somatic	Germline/Somatic
PIWI pth	*Avg*	0.295	0.103	0.195	0.012	16.85
	Avg NM	0.253	0.144	0.202	0.011	18.26
	Avg M	0.381	0.030	0.183	0.013	14.47
siRNA pth	*Avg*	0.038	0.047	0.040	0.026	1.54
	Avg NM	0.030	0.043	0.036	0.017	2.11
	Avg M	0.054	0.053	0.047	0.046	1.03
Transcription	*Avg*	0.182	0.230	0.197	0.105	1.87
	Avg NM	0.186	0.241	0.212	0.106	2.01
	Avg M	0.174	0.211	0.171	0.103	1.66
Post-Transcription	*Avg*	0.089	0.124	0.102	0.071	1.44
	Avg NM	0.077	0.089	0.082	0.059	1.40
	Avg M	0.114	0.187	0.134	0.096	1.39
Total Regulators	*Avg*	0.605	0.501	0.532	0.213	2.50
	Avg NM	0.546	0.514	0.531	0.193	2.76
	Avg M	0.723	0.479	0.535	0.257	2.08

Supplementary File 4. Proportion of TE-derived transcripts in vertebrate transcriptomes. Percentages of TE derived transcripts were calculated after within-species RNAseq data normalization in DESeq2.

Total-TE	**Testis**	**Ovary**	**Brain**	**Heart**	**Kidney**	**Liver**	**Muscle**	**Spleen**	**SI**	***Tot Avg***	***Avg Somatic***	***Avg Germline***
Danio rerio	9.10	1.15	12.70	4.76	2.56	0.72	2.01	4.07	4.54	4.62	4.48	5.13
Xenopus laevis	2.85	1.37	9.16	5.59	5.92	3.00	3.07	7.84	5.27	4.90	5.69	2.11
Alligator mississippiensis	15.01	5.70	11.65	4.86	9.68	5.62	8.25	6.84		8.45	7.82	10.36
Gallus gallus	11.33	2.00	1.20	1.14	0.91	0.96	0.26	4.24	1.92	2.66	1.52	6.66
Anolis carolinensis	7.68	1.10	7.89	5.33	6.02	4.35	2.25			4.94	5.17	4.39
Boa constrictor	14.53		15.32		14.72	13.53	2.06	15.91	9.08	12.16	11.77	-
Python molurus	8.01	3.57	7.12	6.80	8.44	3.89	1.20	6.43	6.79	5.81	5.81	5.79
Crotalus viridis	13.41	5.49	16.60	18.30	10.17	11.79	9.59	17.06	21.32	13.75	14.97	9.45
Ornithorhynchus anatinus	11.70	5.66	7.84	9.61	13.81	8.18				9.47	9.86	8.68
Monodephis domestica	8.40	10.98	16.12	4.39	7.43	5.97	3.52	23.44		10.03	10.14	9.69
Homo sapiens	5.71	4.14	2.17	1.54	1.75	1.61	1.12	4.88	3.42	2.93	2.36	4.93
Mus musculus	7.85	4.31	12.80	3.96	4.04	2.51	2.18	2.91	2.49	4.79	4.42	6.08
Avg	9.63	4.13	10.05	6.03	7.12	5.18	3.23	9.36	6.85	6.86	6.82	7.00
StDev	3.66	2.90	5.08	4.68	4.48	4.13	2.97	6.94	6.29	5.04	5.27	4.29

Recent-TE	**Testis**	**Ovary**	**Brain**	**Heart**	**Kidney**	**Liver**	**Muscle**	**Spleen**	**SI**	***Tot Avg***	***Avg Somatic***	***Avg Germline***
Danio rerio	1.94	0.08	0.94	0.47	0.14	0.04	0.11	0.34	0.28	0.48	0.33	1.01
Xenopus laevis	0.02	0.01	0.05	0.03	0.03	0.02	0.01	0.05	0.03	0.03	0.03	0.01
Alligator mississippiensis	0.06	0.01	0.03	0.02	0.03	0.02	0.02	0.03		0.03	0.02	0.04
Gallus gallus	0.05	0.08	0.01	0.02	0.02	0.01	0.00	0.05	0.75	0.11	0.12	0.06
Anolis carolinensis	0.25	0.02	0.23	0.19	0.23	0.10	0.07			0.16	0.16	0.14
Boa constrictor	0.03		0.02		0.03	0.03	0.00	0.05	0.01	0.02	0.02	-
Python molurus	0.08	0.01	0.09	0.10	0.17	0.08	0.01	0.07	0.21	0.09	0.10	0.04
Crotalus viridis	0.21	0.06	0.26	0.27	0.14	0.17	0.15	0.25	0.34	0.21	0.23	0.13
Ornithorhynchus anatinus	0.19	0.06	0.07	0.11	0.27	0.09				0.13	0.13	0.13
Monodephis domestica	0.10	0.16	0.21	0.05	0.10	0.09	0.04	0.25		0.12	0.12	0.13
Homo sapiens	0.03	0.02	0.01	0.01	0.01	0.01	0.00	0.03	0.02	0.02	0.01	0.03
Mus musculus	0.24	0.15	0.39	0.17	0.14	0.10	0.49	0.09	0.15	0.21	0.22	0.19
Avg	0.27	0.06	0.19	0.13	0.11	0.06	0.08	0.12	0.22	0.14	0.13	0.17
StDev	0.53	0.05	0.26	0.14	0.09	0.05	0.14	0.11	0.25	0.24	0.16	0.39

Δ % TE mapping reads	Testis	Ovary	Brain	Heart	Kidney	Liver	Muscle	Spleen	SI	*Tot Avg*	*Avg Somatic*	*Avg Germline*
Danio rerio	80.25	92.92	93.46	90.60	94.49	94.66	94.57	91.99	94.12	91.90	93.41	86.58
Xenopus laevis	99.40	99.48	99.47	99.56	99.45	99.51	99.56	99.43	99.51	99.49	99.50	99.44
Alligator mississippiensis	99.24	99.58	99.48	99.37	99.42	99.28	99.43	99.28		99.38	99.38	99.41
Gallus gallus	99.23	92.38	98.97	96.94	96.33	97.92	97.74	97.81	22.26	88.84	86.85	95.81
Anolis carolinensis	93.95	96.45	94.94	93.75	93.13	95.64	94.60			94.64	94.41	95.20
Boa constrictor	99.70		99.75		99.65	99.65	99.73	99.50	99.75	99.67	99.67	-
Python molurus	99.12	99.75	98.86	98.64	98.17	98.08	99.00	98.96	97.05	98.62	98.39	99.43
Crotalus viridis	98.64	99.02	98.71	98.78	98.73	98.72	98.54	98.78	98.72	98.74	98.71	98.83
Ornithorhynchus anatinus	98.55	98.99	99.14	98.97	98.34	99.01				98.83	98.86	98.77
Monodephis domestica	98.90	98.70	98.91	98.89	98.76	98.55	98.93	99.19		98.85	98.87	98.80
Homo sapiens	99.46	99.47	99.60	99.66	99.44	99.48	99.55	99.44	99.54	99.52	99.53	99.47
Mus musculus	97.22	96.76	97.34	95.93	96.67	96.01	78.56	97.03	94.12	94.41	93.67	96.99
Avg	96.97	97.59	98.22	97.37	97.71	98.04	96.38	98.14	88.14	96.79	96.65	97.27
StDev	5.49	2.68	2.00	2.89	2.13	1.69	6.20	2.30	26.72	8.28	9.17	4.30

Genomic %	Total	K2D<2
Danio rerio	49.76	3.08
Xenopus laevis	26.69	0.13
Alligator mississippiensis	37.04	0.31
Gallus gallus	8.31	0.30
Anolis carolinensis	35.28	4.70
Boa constrictor	28.80	0.43
Python molurus	22.09	0.26
Crotalus viridis	35.83	1.32
Ornithorhynchus anatinus	53.55	3.20
Monodephis domestica	54.28	2.95
Homo sapiens	49.91	1.68
Mus musculus	41.89	5.53

Genomic TE loci	Total	Recent-TE	% Recent-TE
Danio rerio	2930931	60145	2.05
Xenopus laevis	3050695	15727	0.52
Alligator mississippiensis	2438072	18564	0.76
Gallus gallus	346075	2312	0.67
Anolis carolinensis	2076314	48279	2.33
Boa constrictor	1716891	9026	0.53
Python molurus	1783137	17821	1.00
Crotalus viridis	2004731	22485	1.12
Ornithorhynchus anatinus	4634624	47507	1.03
Monodephis domestica	4398463	17915	0.41
Homo sapiens	4630459	10350	0.22
Mus musculus	3688958	56683	1.54

Supplementary File 5A. Recently inserted TE copies abundance (recent-TEs) across somatic and germline tissues in vertebrates. Abundance of TE major families was calculated as percentage of the normalized transcriptome to the inclusion of Unknown elements.

		DNA transposons				LTRs			Other RT	LINEs					SINEs		
		Others	TcMar	hAT	RC/Helitron	Other	ERV	Gypsy		Other	CR1-L3	L2	RTE	L1	Other	Alu-SVA	MIR
Danio rerio	T	0.6525	0.1252	0.0739	0.0999	0.0490	0.0740	0.5601	0.0307	0.0140	0	0.2271	0.0012	0.0059	0.0009	0	0
	O	0.0120	0.0028	0.0059	0.0079	0.0024	0.0075	0.0169	0.0077	0.0012	0	0.0169	0.0001	0.0005	0.0001	0	0
	B	0.1607	0.0389	0.0923	0.0020	0.0415	0.0405	0.1177	0.0369	0.0107	0	0.3790	0.0009	0.0101	0.0013	0	0
	H	0.0942	0.0266	0.0430	0.0012	0.0208	0.0229	0.0604	0.0119	0.0056	0	0.1678	0.0005	0.0038	0.0010	0	0
	K	0.0369	0.0100	0.0163	0.0035	0.0070	0.0088	0.0225	0.0067	0.0019	0	0.0287	0.0002	0.0011	0.0003	0	0
	L	0.0096	0.0025	0.0054	0.0004	0.0030	0.0026	0.0054	0.0025	0.0005	0	0.0057	0.0002	0.0004	0.0000	0	0
	M	0.0289	0.0128	0.0117	0.0015	0.0044	0.0046	0.0152	0.0062	0.0009	0	0.0234	0.0001	0.0005	0.0001	0	0
	S	0.0670	0.0143	0.0351	0.0011	0.0156	0.0267	0.0421	0.0104	0.0042	0	0.1133	0.0004	0.0017	0.0009	0	0
	SI	0.0660	0.0137	0.0307	0.0008	0.0132	0.0170	0.0338	0.0099	0.0035	0	0.0857	0.0004	0.0018	0.0003	0	0
Xenopus laevis	T	0.0039	0.0107	0.0019	0	0	0	0	0	0	0.0003	0	0	5.19E-05	0	0	0
	O	0.0009	0.0057	0.0002	0	0	0	0	7.86E-06	0	0.0002	0	0	7.86E-06	0	0	0
	B	0.0093	0.0274	0.0116	0.0001	2.23E-05	4.02E-05	8.93E-06	0.0001	0	0.0009	0	0	0.0001	0	0	0
	H	0.0042	0.0123	0.0072	0.0001	1.62E-05	8.08E-06	0.0001	2.42E-05	0	0.0005	0	0	8.08E-05	0	0	0
	K	0.0057	0.0193	0.0067	3.42E-05	6.21E-06	3.42E-05	3.11E-06	9.94E-05	0	0.0005	0	0	4.97E-05	0	0	0
	L	0.0022	0.0076	0.0045	0	4.92E-06	4.92E-06	1.48E-05	3.45E-05	0	0.0002	0	0	5.42E-05	0	0	0
	M	0.0025	0.0070	0.0021	0.0002	3.53E-05	0	0.0001	1.06E-05	0	0.0012	0	0	7.06E-06	0	0	0
	S	0.0090	0.0242	0.0122	0.0001	8.18E-06	5.72E-05	4.09E-06	0.0002	0	0.0007	0	0	0.0001	0	0	0
	SI	0.0061	0.0125	0.0065	1.53E-05	6.11E-06	0.0002	6.11E-06	4.28E-05	0	0.0004	0	0	2.75E-05	0	0	0
Alligator mississippiensis	T	0.0018	0.0017	0	0.0004	0.0005	0.0401	0.0047	0.0022	0	0.0113	0	0	0	0	0	0
	O	0.0003	0.0004	0	7.5825E-05	3.79E-05	0.0085	0.0006	0.0003	0	0.0016	0	0	0	0	0	0
	B	0.0004	0.0022	0	0.0004	0.0001	0.0112	0.0043	0.0024	0	0.0101	0	0	0	0	0	0
	H	0.0016	0.0009	0	0.0001	0.0001	0.0055	0.0011	0.0013	0	0.0047	0	0	0	0	0	0
	K	6.20E-05	0.0012	0	0.0001	0.0002	0.0126	0.0043	0.0019	0	0.0084	0	0	0	0	0	0
	L	4.79E-05	0.0007	0	4.79257E-05	4.79E-05	0.0129	0.0019	0.0009	0	0.0042	0	0	0	0	0	0
	M	0.0005	0.0005	0	0.0003	0	0.0083	0.0040	0.0019	0	0.0082	0	0	0	0	0	0
	S	0.0007	0.0009	0	0.0002	0	0.0154	0.0023	0.0008	0	0.0050	0	0	0	0	0	0
	SI																
Gallus gallus	T	0	0	0	0	0	0.0382	0	0	0	0.0107	0	0	0	0	0	0
	O	0	0	0	0	0	0.0767	0	0	0	0.0010	0	0	0	0	0	0
	B	0	0	0	0	0	0.0052	0	0	0	0.0011	0	0	0	0	0	0
	H	0	0	0	0	0	0.0162	0	0	0	0.0015	0	0	0	0	0	0
	K	0	0	0	0	0	0.0149	0	0	0	0.0020	0	0	0	0	0	0
	L	0	0	0	0	0	0.0076	0	0	0	0.0025	0	0	0	0	0	0
	M	0	0	0	0	0	0.0028	0	0	0	0.0002	0	0	0	0	0	0
	S	0	0	0	0	0	0.0441	0	0	0	0.0044	0	0	0	0	0	0
	SI	0	0	0	0	0	0.7491	0	0	0	0.0004	0	0	0	0	0	0

		DNA transposons				LTRs			Other RT	LINEs					SINEs		
		Others	TcMar	hAT	RC/Helitron	Other	ERV	Gypsy		Other	CR1-L3	L2	RTE	L1	Other	Alu-SVA	MIR
Anolis carolinensis	T	0.0003	0.0080	0.0100	0.0285	0.0259	0.0039	0.0344	0.0203	0.0113	0.0264	0.0218	0.0095	0.0524	0.0002	0	0
	O	0.0003	0.0005	0.0011	0.0007	0.0022	0.0001	0.0022	0.0035	0.0006	0.0011	0.0021	0.0010	0.0043	1.57E-05	0	0
	B	0.0005	0.0097	0.0109	0.0236	0.0126	0.0013	0.0382	0.0223	0.0103	0.0337	0.0306	0.0092	0.0245	0.0006	0	0
	H	0.0006	0.0084	0.0095	0.0192	0.0141	0.0015	0.0308	0.0194	0.0089	0.0303	0.0203	0.0067	0.0232	0.0002	0	0
	K	0.0008	0.0067	0.0141	0.0169	0.0128	0.0035	0.0599	0.0250	0.0127	0.0402	0.0197	0.0075	0.0139	0.0004	0	0
	L	0.0003	0.0031	0.0050	0.0118	0.0051	0.0018	0.0091	0.0080	0.0047	0.0106	0.0220	0.0030	0.0182	4.51E-05	0	0
	M	0.0002	0.0022	0.0016	0.0067	0.0033	0.0003	0.0064	0.0052	0.0032	0.0085	0.0120	0.0018	0.0144	2.03E-05	0	0
	S																
	SI																
Boa constrictor	T	0.0002	0.0028	0.0017	0	0.0030	0	3.20E-05	0.0055	0.0009	1.60E-05	0	0.0091	0.0003	0.0003	0	0.0003
	O																
	B	0.0003	0.0021	0.0027	0	0.0008	0	5.71E-05	0.0039	8.56E-05	0.0002	0	0.0099	0.0002	0.0001	0	8.56E-05
	H																
	K	0.0002	0.0024	0.0031	0	0.0018	0	2.02E-05	0.0070	0.0050	4.05E-05	0	0.0081	0.0003	0.0006	0	8.09E-05
	L	8.23E-05	0.0019	0.0036	0	0.0011	0	0	0.0119	0.0014	2.74E-05	0	0.0058	0.0003	0.0001	0	5.49E-05
	M	1.03E-05	0.0003	0.0003	0	0.0002	0	0	0.0007	7.24E-05	0	0	0.0011	4.14E-05	4.14E-05	0	0
	S	0.0001	0.0048	0.0026	0	0.0017	0	2.69E-05	0.0055	0.0178	1.34E-05	0	0.0119	0.0002	9.41E-05	0	0.0005
	SI	8.95E-05	0.0017	0.0012	0	0.0009	0	2.98E-05	0.0032	0.0001	0	0	0.0036	7.95E-05	0.0003	0	8.95E-05
Python molurus	T	0.0029	0.0142	0.0021	0	0.0066	0.0054	0.0019	0.0009	0.0002	0	0	0.0300	0	0	0	0
	O	0.0003	0.0010	0.0011	0	0.0001	0.0004	3.48E-05	3.48E-05	3.48E-05	0.0004	0	0.0036	0	0	0	0
	B	0.0032	0.0176	0.0047	0	0.0036	0.0047	0.0013	0.0007	0.0004	2.93E-05	0	0.0430	0	0	0	0
	H	0.0015	0.0193	0.0027	0	0.0038	0.0085	0.0005	0.0009	0.0006	0	0	0.0394	0	0	0	0
	K	0.0084	0.0394	0.0039	0	0.0006	0.0035	0.0011	0.0010	0.0007	0	0	0.1006	0	0	0	0
	L	0.0045	0.0146	0.0011	0	0.0025	0.0065	0.0009	0.0002	0.0002	0	0	0.0428	0	0	0	0
	M	0.0004	0.0024	0.0003	0	0.0003	0.0002	0	0.0001	3.02E-05	0	0	0.0056	0	0	0	0
	S	0.0027	0.0152	0.0046	0	0.0042	0.0033	0.0029	0.0002	0.0005	0.0015	0	0.0262	0	0	0	0
	SI	0.0049	0.0393	0.0015	0	0.0002	0.0018	9.24E-05	0.0010	4.62E-05	0	0	0.1566	0	0	0	0
Crotalus viridis	T	0.0067	0.0056	0.0622	0.0002	0.0018	0.0011	0.0172	0.0230	0.0051	0.0249	0.0008	0.0593	0.0007	0	0	5.72E-05
	O	0.0047	0.0022	0.0203	0	0.0002	0.0003	0.0030	0.0062	0.0005	0.0111	0.0001	0.0076	9.02E-05	0	0	0
	B	0.0128	0.0080	0.1160	0.0001	0.0012	0.0012	0.0230	0.0184	0.0043	0.0257	0.0004	0.0443	0.0001	0	0	0.0003
	H	0.0137	0.0100	0.0875	0.0001	0.0021	0.0008	0.0264	0.0166	0.0091	0.0389	0.0011	0.0658	7.21E-05	0	0	7.21E-05
	K	0.0060	0.0060	0.0482	0	0.0006	0.0003	0.0150	0.0113	0.0016	0.0174	0.0002	0.0364	0.0002	0	0	4.62E-05
	L	0.0058	0.0049	0.0710	3.58E-05	0.0013	0.0006	0.0204	0.0177	0.0047	0.0140	0.0007	0.0286	0.0001	0	0	7.16E-05
	M	0.0059	0.0042	0.0494	0.0001	0.0007	0.0003	0.0143	0.0326	0.0034	0.0145	0.0001	0.0283	0	0	0	0
	S	0.0113	0.0089	0.1028	0	0.0011	0.0009	0.0351	0.0170	0.0014	0.0173	0.0002	0.0533	0.0001	0	0	0.0002
	SI	0.0142	0.0105	0.1123	0.0002	0.0035	0.0015	0.0333	0.0224	0.0115	0.0485	0.0015	0.0853	0.0002	0	0	0.0001
Ornithorhynchus anatinus	T	0	0	4.50E-05	0	0	0.0112	0	0	0	8.44E-06	0.1182	0	0	3.94E-05	0	0.0623
	O	0	0	1.28E-05	0	0	0.0036	0	0	0	0	0.0333	0	0	9.57E-06	0	0.0236
	B	0	0	2.03E-05	0	0	0.0005	0	0	0	3.38E-06	0.0424	0	0	2.37E-05	0	0.0303
	H	0	0	1.12E-05	0	0	0.0006	0	0	0	0	0.0592	0	0	8.93E-05	0	0.0493
	K	0	0	3.76E-05	0	0	0.0051	0	0	0	2.35E-05	0.2037	0	0	4.70E-06	0	0.0562
	L	0	0	1.33E-05	0	0	0.0003	0	0	0	7.96E-06	0.0505	0	0	1.06E-05	0	0.0372
	M																
	S																
	SI																

		DNA transposons				LTRs			Other RT	LINEs					SINEs		
		Others	TcMar	hAT	RC/Helitron	Other	ERV	Gypsy		Other	CR1-L3	L2	RTE	L1	Other	Alu-SVA	MIR
Monodelphis domestica	T	0	0	0.0008	0	0	0.0704	0	0	0	0	0	0.0001	0.0259	8.93E-05	0	0.0019
	O	3.77E-05	0	0.0011	0	0	0.1184	0	0	0	0	9.43E-06	0.0002	0.0350	2.83E-05	0	0.0033
	B	0	0	0.0016	0	0	0.1305	0	0	0	0	3.38E-05	0.0002	0.0698	0.0001	0	0.0042
	H	7.50E-06	0	0.0007	0	0	0.0333	0	0	0	0	0	5.25E-05	0.0136	1.50E-05	0	0.0007
	K	1.68E-05	0	0.0014	0	0	0.0671	0	0	0	0	2.53E-05	9.27E-05	0.0286	5.90E-05	0	0.0015
	L	0	0	0.0005	0	0	0.0701	0	0	0	0	0	3.58E-05	0.0192	2.68E-05	0	0.0014
	M	0	0	0.0006	0	0	0.0267	0	0	0	0	9.03E-06	4.52E-05	0.0103	1.81E-05	0	0.0010
	S	1.46E-05	0	0.0053	0	0	0.1430	0	0	0	0	5.85E-05	0.0004	0.0928	0.0002	0	0.0042
	SI																
Homo sapiens	T	0	0	2.41E-05	0	0	0.0091	0	0	0	0	0	0	0.0098	0	0.0135	0
	O	0	0	3.59E-06	0	0	0.0107	0	0	0	0	0	0	0.0064	0	0.0054	0
	B	0	0	1.77E-05	0	0	0.0010	0	0	0	0	0	0	0.0046	0	0.0030	0
	H	0	0	3.84E-06	0	0	0.0010	0	0	0	0	0	0	0.0022	0	0.0020	0
	K	0	0	0	0	0	0.0033	0	0	0	0	0	0	0.0037	0	0.0028	0
	L	0	0	0	0	0	0.0022	0	0	0	0	0	0	0.0029	0	0.0032	0
	M	0	0	9.86E-06	0	0	0.0004	0	0	0	0	0	0	0.0023	0	0.0022	0
	S	0	0	1.79E-05	0	0	0.0074	0	0	0	0	0	0	0.0069	0	0.0142	0
	SI	0	0	0	0	0	0.0048	0	0	0	0	0	0	0.0050	0	0.0062	0
Mus musculus	T	0	0	4.91E-05	0	0	0.1043	0	0	0	0	0	0	0.1253	0.0030	0.0030	0
	O	0	0	0.0002	0	0	0.1076	0	0	0	0	0	0	0.0338	0.0017	0.0019	0
	B	0	0	7.36E-05	0	0	0.1286	0	1.17E-05	0	0	0	0	0.2464	0.0068	0.0055	0
	H	0	0	4.90E-05	0	0	0.1109	0	0	0	0	0	0	0.0531	0.0021	0.0014	0
	K	0	0	1.33E-05	0	0	0.0892	0	0	0	0	0	0	0.0463	0.0020	0.0023	0
	L	0	0	1.85E-05	0	0	0.0622	0	0	0	0	0	0	0.0380	0.0012	0.0009	0
	M	0	0	4.93E-06	0	0	0.0357	0	0	0	0	0	0	0.4391	0.0135	0.0058	0
	S	0	0	0	0	0	0.0638	0	0	0	0	0	0	0.0149	0.0073	0.0022	0
	SI	0	0	6.11E-05	0	0	0.0439	0	0	0	0	0	0	0.1040	0.0009	0.0009	0

Supplementary File 5B. Total-TE abundance across somatic and germline tissues in vertebrates. Abundance of TE major families was calculated as percentage of the normalized transcriptome to the inclusion of Unknown elements.

		DNA transposons				LTRs			Other RT	LINEs					SINEs		
		Others	TcMar	hAT	RC/Helitron	Other	ERV	Gypsy		Other	CR1-L3	L2	RTE	L1	Other	Alu-SVA	MIR
Danio rerio	T	3.3208	0.9473	1.7861	0.3023	0.1947	0.7868	0.7846	0.1764	0.0609	0	0.4960	0.0538	0.0516	0.0388	0	0
	O	0.4279	0.0796	0.3290	0.0467	0.0341	0.0705	0.0588	0.0234	0.0078	0	0.0418	0.0071	0.0034	0.0060	0	0
	B	5.7572	1.7646	3.0192	0.2279	0.1910	0.2546	0.2427	0.2814	0.0589	0.0001	0.5577	0.0867	0.0633	0.0755	0	0
	H	1.9567	0.6080	1.0573	0.0855	0.0829	0.3592	0.1134	0.0856	0.0317	0	0.2606	0.0218	0.0277	0.0235	0	0
	K	1.1082	0.3663	0.6398	0.0694	0.0466	0.0886	0.0560	0.0500	0.0126	0	0.0706	0.0130	0.0092	0.0131	0	0
	L	0.3086	0.1026	0.1738	0.0223	0.0193	0.0119	0.0200	0.0143	0.0044	0	0.0212	0.0033	0.0028	0.0062	0	0
	M	0.9098	0.3038	0.4593	0.0408	0.0338	0.0695	0.0423	0.0383	0.0092	0	0.0553	0.0124	0.0062	0.0103	0	0
	S	1.6683	0.5194	0.9422	0.1096	0.0813	0.2193	0.1067	0.0854	0.0259	0	0.1998	0.0192	0.0185	0.0326	0	0
	SI	2.0347	0.5715	1.0253	0.1117	0.0843	0.2795	0.0860	0.0685	0.0216	0	0.1613	0.0182	0.0209	0.0235	0	0
Xenopus laevis	T	1.6027	0.1924	0.4605	0.0009	0.0103	0.1750	0.0581	0.0942	0.0040	0.0788	0.0318	0.0003	0.0605	0.0001	0	0.0004
	O	0.8130	0.1082	0.2884	0.0001	0.0022	0.0080	0.0154	0.0486	0.0004	0.0143	0.0088	0.0002	0.0203	0.0018	0	0.0003
	B	5.0982	0.8755	1.5595	0.0046	0.0258	0.0786	0.2138	0.4464	0.0098	0.2915	0.1201	0.0015	0.1743	0.0012	0	0.0027
	H	3.0022	0.4580	0.6858	0.0018	0.0310	0.1333	0.6859	0.2123	0.0073	0.1312	0.0543	0.0005	0.0697	0.0001	0	0.0016
	K	3.4537	0.5608	0.8767	0.0032	0.0165	0.0643	0.1096	0.2493	0.0087	0.2013	0.0873	0.0006	0.1179	0.0002	0	0.0016
	L	1.7314	0.2495	0.4826	0.0015	0.0080	0.0283	0.0771	0.1397	0.0036	0.0842	0.0432	0.0008	0.0712	0.0001	0	0.0010
	M	1.8859	0.2768	0.5061	0.0023	0.0058	0.0277	0.0371	0.1012	0.0029	0.0742	0.0297	0.0009	0.0385	0.0001	0	0.0004
	S	4.2939	0.7332	1.1680	0.0054	0.0270	0.1748	0.1807	0.5014	0.0148	0.2551	0.1221	0.0016	0.1482	0.0002	0	0.0024
	SI	2.9713	0.5041	0.7999	0.0021	0.0676	0.0535	0.0996	0.2386	0.0114	0.1683	0.0900	0.0009	0.1022	0.0002	0	0.0015
Alligator mississippiensis	T	1.4130	0.2764	1.5043	0.1098	0.0984	1.2429	0.5345	0.2943	0.1003	1.9433	0.0955	0.2148	0.2454	0.0784	0	0.0979
	O	0.6642	0.1064	0.5695	0.0106	0.0531	0.2861	0.3735	0.0868	0.0272	0.4064	0.0440	0.1024	0.1282	0.0421	0	0.0669
	B	1.5940	0.2648	1.3496	0.0287	0.0685	0.3495	0.2417	0.2400	0.0331	1.6218	0.0838	0.1033	0.0607	0.0689	0	0.1133
	H	0.6296	0.1105	0.5192	0.0113	0.0247	0.1521	0.0861	0.1040	0.0102	0.6750	0.0268	0.0502	0.0428	0.0265	0	0.0301
	K	1.2276	0.2139	1.1012	0.0208	0.0703	0.3560	0.2389	0.2099	0.0282	1.3408	0.0779	0.0856	0.0617	0.0488	0	0.0943
	L	0.6476	0.1022	0.8395	0.0106	0.0314	0.2090	0.1276	0.1071	0.0155	0.6732	0.0363	0.0306	0.0392	0.0304	0	0.0376
	M	1.0285	0.2097	0.9934	0.0212	0.0549	0.2667	0.1524	0.2077	0.0221	1.0815	0.0649	0.0679	0.0506	0.0510	0	0.0713
	S	0.6320	0.1229	0.7838	0.0145	0.0655	0.4345	0.1626	0.1808	0.0191	0.8808	0.0400	0.0482	0.0964	0.0759	0	0.0560
	SI																
Gallus gallus	T	0.4972	0.2117	0.7695	0	0.0062	2.7076	0.0001	0	0	6.7978	0.0553	0.0003	0.0015	0.0819	0	0.0408
	O	0.0453	0.0226	0.0755	0	0.0005	0.9161	8.77E-05	0	0	0.8995	0.0085	9.96E-05	0.0021	0.0113	0	0.0067
	B	0.0341	0.0312	0.0571	0	3.40E-05	0.1641	1.53E-05	0	0	0.8839	0.0036	1.19E-05	3.23E-05	0.0140	0	0.0124
	H	0.0126	0.0102	0.0299	0	0.0001	0.2488	4.24E-06	0	0	0.8268	0.0018	0	2.97E-05	0.0041	0	0.0060
	K	0.0074	0.0078	0.0229	0	1.15E-05	0.2045	2.29E-06	0	0	0.6564	0.0011	2.29E-06	1.61E-05	0.0020	0	0.0040
	L	0.0052	0.0064	0.0164	0	3.07E-05	0.1323	0	0	0	0.7950	0.0008	0	1.28E-05	0.0020	0	0.0018
	M	0.0041	0.0023	0.0077	0	3.94E-06	0.0537	1.97E-06	0	0	0.1905	0.0006	1.97E-06	7.87E-06	0.0006	0	0.0008
	S	0.1236	0.0643	0.2305	0	0.0012	0.9743	9.83E-05	0	0	2.7567	0.0178	1.97E-05	0.0006	0.0229	0	0.0150
	SI	0.0131	0.0085	0.0213	0	0.0001	1.5808	2.73E-05	0	0	0.2854	0.0037	0	4.55E-05	0.0031	0	0.0027

		DNA transposons				LTRs			Other RT	LINEs					SINEs		
		Others	TcMar	hAT	RC/Helitron	Other	ERV	Gypsy		Other	CR1-L3	L2	RTE	L1	Other	Alu-SVA	MIR
Anolis carolinensis	T	0.6759	0.7878	1.0283	0.6457	0.2041	0.0649	0.5659	0.5162	0.3726	0.7952	1.0178	0.4667	0.4739	0.0649	0	0.0009
	O	0.2347	0.0642	0.2370	0.0310	0.0220	0.0041	0.0551	0.0652	0.0217	0.0970	0.1365	0.0402	0.0826	0.0059	0	0.0002
	B	1.0877	1.0432	1.4696	0.4610	0.1409	0.0334	0.4132	0.4081	0.3165	0.7776	0.8573	0.3063	0.5039	0.0668	0	0.0029
	H	0.6839	0.7573	0.8822	0.3347	0.1191	0.0540	0.3195	0.3038	0.2336	0.5292	0.5305	0.2189	0.3179	0.0405	0	0.0009
	K	0.7873	0.5194	1.0820	0.4372	0.1125	0.0590	0.6104	0.3489	0.3178	0.5889	0.6072	0.3261	0.1823	0.0355	0	0.0007
	L	0.5098	0.7703	0.7386	0.2325	0.0580	0.0475	0.2304	0.1289	0.4627	0.2704	0.5162	0.1319	0.2256	0.0210	0	0.0006
	M	0.3466	0.3001	0.4048	0.1267	0.0297	0.0178	0.0975	0.0928	0.1005	0.2562	0.2200	0.0907	0.1519	0.0168	0	0.0003
	S																
	SI																
Boa constrictor	T	0.9269	1.4428	1.9123	0.0314	0.9501	1.0124	0.4421	0.6558	0.2167	1.3600	1.3462	1.3531	0.9827	0.5076	0.0005	0.7057
	O																
	B	1.0850	1.7711	2.5058	0.0226	0.9026	0.1467	0.6313	0.5874	0.2385	1.3476	1.1092	1.7600	1.0125	0.5358	0.0003	0.8134
	H																
	K	0.9505	1.6641	1.9775	0.0221	0.8711	0.4268	0.4893	0.7115	0.2499	1.4459	1.3926	1.3685	1.0839	0.5286	0.0003	0.8146
	L	1.0037	1.6748	1.7512	0.0198	0.6932	0.1399	0.4344	0.7220	0.1896	1.3871	1.2568	1.1240	1.0817	0.4755	0.0002	0.8998
	M	0.1539	0.2498	0.3046	0.0054	0.0981	0.0273	0.0449	0.0865	0.0282	0.2416	0.1605	0.1729	0.1565	0.1093	0	0.1293
	S	0.9361	2.0644	2.0018	0.0346	0.9961	0.2485	0.4850	0.7761	0.2812	1.4869	1.6470	1.5443	1.0882	0.5609	0.0004	0.8743
	SI	0.6470	1.0842	1.2415	0.0217	0.7890	0.1334	0.2147	0.4357	0.1492	0.8045	0.8707	0.8091	0.6240	0.3144	0.0006	0.4841
Python molurus	T	0.5096	0.7136	1.0243	0.0185	0.7842	0.9141	0.1349	0.2957	0.1312	0.5766	0.6663	0.5152	0.8246	0.2708	0	0.2399
	O	0.3870	0.1699	0.7240	0.0065	0.3111	0.0475	0.1714	0.1509	0.0319	0.3361	0.3553	0.2069	0.2678	0.1987	0	0.0760
	B	0.5602	0.8234	1.0267	0.0167	0.3668	0.1149	0.1083	0.3235	0.1546	0.5897	0.6757	0.6418	0.7985	0.2661	0	0.2599
	H	0.2952	0.5782	0.4781	0.0192	0.7242	0.1109	0.0777	0.2412	0.1967	0.4774	0.6411	0.7367	1.0811	0.6717	0	0.1297
	K	0.4731	1.0914	0.6789	0.0304	0.5211	0.1559	0.1291	0.4999	0.2622	0.7508	0.8002	0.9809	1.0551	0.2749	0	0.2825
	L	0.2449	0.5525	0.4204	0.0097	0.3990	0.0994	0.0657	0.1669	0.0878	0.2566	0.3372	0.3385	0.3029	0.2342	0	0.1513
	M	0.0904	0.1923	0.1947	0.0027	0.0606	0.0129	0.0171	0.0720	0.0241	0.0959	0.0950	0.0818	0.1172	0.0597	0	0.0355
	S	0.3903	0.7595	0.9215	0.0124	0.4697	0.3056	0.1023	0.2726	0.1618	0.4409	0.5577	0.4722	0.7303	0.2172	0	0.2691
	SI	0.4909	0.7233	0.5519	0.0284	0.4313	0.1212	0.0999	0.3918	0.2835	0.6155	0.6691	0.8131	0.7544	0.2286	0	0.2002
Crotalus viridis	T	0.6822	1.3564	2.2243	0.0184	0.2014	0.4197	0.7532	1.5577	0.4070	2.5032	0.6712	0.9157	1.1830	0.0797	0	0.4208
	O	0.6478	0.5227	0.7547	0.0096	0.0396	0.1013	0.1211	0.4436	0.1289	0.8941	0.1917	0.2343	1.2304	0.0244	0	0.1126
	B	1.5064	1.9242	3.2303	0.0414	0.1483	0.5195	0.5390	1.6398	0.3286	3.1865	0.7005	0.9932	1.1917	0.1187	0	0.5217
	H	0.9970	1.8001	3.1242	0.0253	0.7043	0.8089	0.9078	1.9534	0.3978	3.5274	0.7869	0.9715	1.6800	0.0893	0	0.4972
	K	0.7005	1.1342	2.1075	0.0144	0.1144	0.4276	0.4490	1.1550	0.1979	1.8374	0.3845	0.5046	0.7851	0.0499	0	0.2975
	L	0.6022	1.2020	2.2028	0.0205	0.1892	0.3980	0.5311	1.0138	0.2925	2.0201	0.4107	0.5650	1.9757	0.0619	0	0.2841
	M	0.6745	1.0401	1.5790	0.0130	0.0997	0.4286	0.3749	0.9119	0.3567	1.7706	0.4048	0.5540	1.0735	0.0546	0	0.2445
	S	1.2494	1.6580	2.9724	0.0148	0.2185	2.7385	0.4835	1.2724	0.3952	2.5144	0.5683	0.8460	1.6574	0.0773	0	0.3666
	SI	1.0957	2.1549	3.6146	0.0337	0.6329	1.1414	1.2216	2.1832	0.4669	4.3322	0.8626	1.1515	1.7271	0.1005	0	0.5668
Ornithorhynchus anatinus	T	0.0288	0.1814	0.0870	0.0003	0.0415	0.2304	0.0121	0.0035	0.0047	0.2072	4.9455	0.2802	0.0065	0.1337	6.23E-05	5.5237
	O	0.0208	0.1279	0.0568	0.0003	0.0141	0.1606	0.0035	0.0019	0.0021	0.1149	2.1943	0.1256	0.0076	0.0678	0.0001	2.7566
	B	0.0315	0.1438	0.0602	0.0003	0.0008	0.0979	0.0034	0.0028	0.0017	0.1871	3.2082	0.1703	0.0054	0.1111	0.0005	3.8057
	H	0.0312	0.1380	0.0708	1.41E-04	0.0005	0.0721	0.0031	0.0019	0.0037	0.1603	3.7571	0.1303	0.0071	0.0847	2.36E-05	5.1355
	K	0.0255	0.2159	0.0726	1.14E-04	0.0047	0.2221	0.0055	0.0019	0.0036	0.2123	6.0289	0.2877	0.0051	0.1310	2.84E-05	6.5926
	L	0.0151	0.1242	0.0513	4.39E-05	0.0061	0.0568	0.0052	0.0011	0.0013	0.1412	3.2890	0.1184	0.0021	0.0877	2.68E-05	4.2797
	M																
	S																
	SI																

		DNA transposons				LTRs			Other RT	LINEs					SINEs		
		Others	TcMar	hAT	RC/Helitron	Other	ERV	Gypsy		Other	CR1-L3	L2	RTE	L1	Other	Alu-SVA	MIR
Monodelphis domestica	T	0.0061	0.1068	0.1927	0.0003	0.0128	1.8863	0.0089	0.0003	0.0004	0.4062	0.9771	0.2730	2.5529	0.2398	6.39E-05	1.7327
	O	0.0073	0.1492	0.2440	0.0004	0.0205	2.2605	0.0062	0.0005	0.0002	0.5159	1.2885	0.3067	3.1754	0.1353	9.29E-05	2.8659
	B	0.0192	0.2065	0.3482	0.0005	0.0190	2.2953	0.0117	0.0037	0.0006	0.7899	2.2415	0.3990	5.6226	0.1664	0.0003	3.9917
	H	0.0020	0.0548	0.1031	0.0010	0.0036	0.6128	0.0012	0.0002	7.21E-05	0.2093	0.5673	0.1339	1.1564	0.0531	1.44E-05	1.2870
	K	0.0044	0.0947	0.1736	0.0011	0.0057	1.2520	0.0031	0.0004	0.0002	0.3365	0.8960	0.1782	2.3155	0.0965	5.48E-05	2.0670
	L	0.0031	0.0636	0.1235	0.0006	0.0040	1.1371	0.0019	0.0001	9.31E-05	0.2358	0.6593	0.1389	1.7932	0.0719	1.69E-05	1.7349
	M	0.0022	0.0524	0.0851	0.0007	0.0023	0.4946	0.0008	0.0000	4.37E-05	0.1747	0.4724	0.1095	0.9429	0.0447	2.62E-05	1.1341
	S	0.0166	0.4186	0.6204	0.0007	0.0280	3.6841	0.0095	0.0009	0.0010	1.2497	2.9043	0.8874	7.2454	0.3302	0.0021	6.0366
	SI																
Homo sapiens	T	0.0111	0.1893	0.3147	0.0014	0.0068	1.1687	0.0196	0.0006	0.0005	0.0727	0.5458	0.0249	1.3424	0.0131	1.4345	0.5605
	O	0.0080	0.1703	0.2750	0.0012	0.0049	0.6182	0.0097	0.0005	0.0007	0.0659	0.4520	0.0206	0.9571	0.0089	1.0552	0.4879
	B	0.0036	0.0886	0.1252	0.0007	0.0025	0.3920	0.0069	0.0003	0.0002	0.0290	0.2357	0.0086	0.6418	0.0046	0.4330	0.1986
	H	0.0028	0.0579	0.0966	0.0004	0.0018	0.2484	0.0054	1.06E-04	8.74E-05	0.0232	0.1660	0.0071	0.3173	0.0027	0.4267	0.1854
	K	0.0027	0.0634	0.1015	0.0006	0.0022	0.3274	0.0071	1.46E-04	0.0004	0.0226	0.1977	0.0068	0.4537	0.0025	0.3885	0.1747
	L	0.0020	0.0723	0.0948	0.0007	0.0018	0.2605	0.0058	0.0003	0.0002	0.0217	0.1616	0.0086	0.4669	0.0023	0.3411	0.1657
	M	0.0011	0.0383	0.0655	0.0003	0.0008	0.1712	0.0020	6.87E-05	5.89E-05	0.0153	0.1336	0.0033	0.1941	0.0014	0.3757	0.1182
	S	0.0073	0.1719	0.2835	0.0012	0.0042	0.6596	0.0103	0.0004	0.0003	0.0620	0.4979	0.0207	1.0387	0.0109	1.5845	0.5259
	SI	0.0063	0.1345	0.2145	0.0013	0.0036	0.5017	0.0095	0.0008	0.0004	0.0480	0.3472	0.0156	0.7932	0.0084	0.9378	0.3952
Mus musculus	T	0.0042	0.0566	0.0020	0.0006	0.0026	2.7115	0.0032	3.50E-06	9.79E-05	0.0249	0.2121	0.0023	1.4456	1.8103	1.0724	0.2688
	O	0.0009	0.0270	0.0005	0.0002	0.0005	1.3816	0.0009	0	2.95E-05	0.0114	0.1086	0.0008	0.5858	1.2336	0.7000	0.1421
	B	0.0081	0.1336	0.0054	0.0011	0.0039	3.1699	0.0050	1.76E-04	0.0006	0.0485	0.3546	0.0055	3.0069	3.2594	1.9833	0.4763
	H	0.0013	0.0341	0.0008	0.0002	0.0009	1.1181	0.0014	1.70E-05	3.73E-05	0.0146	0.0911	0.0010	0.6260	1.1101	0.7138	0.1539
	K	0.0012	0.0278	0.0006	0.0003	0.0006	1.2360	0.0009	1.29E-05	1.61E-05	0.0126	0.0831	0.0009	0.6385	1.1342	0.6548	0.1399
	L	0.0004	0.0170	0.0002	0.0001	0.0002	0.7832	0.0005	9.06E-06	4.53E-06	0.0073	0.0430	0.0007	0.4181	0.7617	0.3441	0.0679
	M	0.0010	0.0185	0.0007	0.0003	0.0003	0.5461	0.0006	0	1.39E-05	0.0129	0.0401	0.0007	0.3661	0.6006	0.4196	0.0866
	S	0.0013	0.0236	0.0007	0.0006	0.0002	0.8512	0.0004	0	1.68E-05	0.0064	0.0422	0.0010	0.4369	0.7803	0.5855	0.0861
	SI	0.0005	0.0131	0.0003	0.0002	0.0002	0.7982	0.0005	0	7.46E-06	0.0119	0.0520	0.0004	0.4039	0.7186	0.3488	0.0756

Supplementary file 6. Linear model testing of a correlation between the fraction of TEs in germline tissue transcriptomes (after normalization) and genomic abundance of major TE subfamilies.

	Testis		Ovary	
Total-TE transcripts vs total genomic abundance (%)	Adjusted R-squared	p-value	Adjusted R-squared	p-value
Danio rerio	0.91	2.12E-08	0.88	1.57E-07
Xenopus laevis	0.94	2.54E-10	0.95	1.46E-10
Alligator mississippiensis	0.86	1.40E-07	0.83	6.50E-07
Gallus gallus	0.97	2.93E-10	0.65	5.68E-04
Anolis carolinensis	0.38	0.009	0.21	0.048
Boa constrictor	0.31	0.014		
Python molurus	0.23	0.035	0.16	0.066
Crotalus viridis	0.92	2.51E-09	0.44	0.003
Ornithorhynchus anatinus	0.98	2.69E-14	0.96	1.08E-11
Monodelphis domestica	0.95	9.01E-11	0.91	7.44E-09
Homo sapiens	0.87	6.97E-08	0.84	3.01E-07
Mus musculus	0.54	6.81E-04	0.37	0.007

	Testis		Ovary	
Recent-TE transcripts vs recent-TE genomic abudance (%)	Adjusted R-squared	p-value	Adjusted R-squared	p-value
Danio rerio	0.57	0.001	0.22	0.050
Xenopus laevis	0.66	1.32E-04	0.49	0.002
Alligator mississippiensis	0.69	3.99E-05	0.60	2.74E-04
Gallus gallus	0.64	0.010	0.33	0.078
Anolis carolinensis	0.24	0.036	0.04	0.228
Boa constrictor	0.93	5.51E-09		
Python molurus	0.82	2.51E-06	0.71	5.36E-05
Crotalus viridis	0.84	1.07E-06	0.84	1.17E-06
Ornithorhynchus anatinus	0.99	1.18E-11	0.94	3.77E-08
Monodelphis domestica	0.22	0.068	0.16	0.110
Homo sapiens	0.58	0.001	0.25	0.048
Mus musculus	0.77	2.36E-04	0.12	0.160

	Testis		Ovary	
Recent-TE transcripts vs total genomic abudance (%)	Adjusted R-squared	p-value	Adjusted R-squared	p-value
Danio rerio	0.44	0.004	0.05	0.212
Xenopus laevis	0.18	0.056	0.02	0.284
Alligator mississippiensis	-0.05	0.590	-0.06	0.738
Gallus gallus	0.09	0.162	-0.06	0.605
Anolis carolinensis	-0.08	0.893	-0.08	0.886
Boa constrictor	0.05	0.194		
Python molurus	0.13	0.095	0.15	0.078
Crotalus viridis	0.34	0.010	0.58	3.81E-04
Ornithorhynchus anatinus	0.92	3.03E-09	0.97	9.88E-13
Monodelphis domestica	0.29	0.018	0.24	0.032
Homo sapiens	0.77	5.69E-06	0.64	1.32E-04
Mus musculus	0.92	1.85E-09	0.44	0.003

Supplementary File 7. Relationship between recent-TE and silencing mechanism expression levels in the germline across vertebrate species. Percentages (left) were calculated after within-species RNAseq data normalization in DESeq2. Multiple linear regression analyses (right) were performed after testing that model assumptions (e.g., normality) where not violated; when necessary, data were log2-transformed to meet the model assumptions.

Testis	TE	Total regulators	PIWI pathway	siRNA pathway	Transcriptional regulation	Post-transcriptional regulation
Danio rerio	1.9380	1.1932	0.7399	0.0439	0.3480	0.0613
Xenopus laevis	0.0175	0.3669	0.1958	0.0160	0.1080	0.0471
Alligator mississippiensis	0.0626	0.7607	0.3999	0.0509	0.2193	0.0906
Gallus gallus	0.0490	0.2318	0.0497	0.0476	0.1185	0.0173
Anolis carolinensis	0.2529	0.2550	0.1339	0.0183	0.0740	0.0288
Boa constrictor	0.0254	0.3864	0.1101	0.0223	0.1821	0.0718
Python molurus	0.0763	0.5384	0.1666	0.0209	0.2375	0.1133
Crotalus viridis	0.2087	0.6376	0.2245	0.0234	0.2575	0.1323
Ornithorhynchus anatinus	0.1918	0.9750	0.6780	0.0804	0.1635	0.0530
Monodelphis domestica	0.0993	0.5411	0.2341	0.0427	0.1973	0.0670
Homo sapiens	0.0324	0.6829	0.2323	0.0509	0.2717	0.1280
Mus musculus	0.2356	0.6950	0.3789	0.0424	0.2054	0.0619

	Linear Model		PIC	
Testis	Adj. R^2	P-val	Adj. R^2	P-val
Total regulators	0.393	0.0173	0.571	0.0043
PIWI	0.409	0.0150	0.413	0.0196
siRNA	-0.088	0.7448	0.144	0.1357
Transcription	0.357	0.0238	0.389	0.0239
Post-Transcription	-0.056	0.5314	-0.111	0.9759

Ovary	TE	Total regulators	PIWI pathway	siRNA pathway	Transcriptional regulation	Post-transcriptional regulation
Danio rerio	0.0820	0.6670	0.1785	0.0373	0.3981	0.0532
Xenopus laevis	0.0072	0.6105	0.1378	0.0445	0.3118	0.1164
Alligator mississippiensis	0.0118	0.6079	0.2106	0.0554	0.2813	0.0605
Gallus gallus	0.0777	0.1819	0.0179	0.0245	0.1202	0.0219
Anolis carolinensis	0.0198	0.3225	0.1128	0.0363	0.1289	0.0446
Boa constrictor	0.0093	0.4728	0.1272	0.0309	0.2059	0.1088
Python molurus	0.0563	0.7351	0.2235	0.0721	0.3433	0.1180
Crotalus viridis	0.0605	0.3662	0.0273	0.0548	0.1832	0.1009
Ornithorhynchus anatinus	0.1581	0.4940	0.0547	0.0628	0.2336	0.1430
Monodelphis domestica	0.0225	0.5466	0.0219	0.0511	0.2082	0.2654
Homo sapiens	0.1453	0.5103	0.0178	0.0422	0.2740	0.1757
Mus musculus	0.0820	0.6670	0.1785	0.0373	0.3981	0.0532

	Linear Model		PIC	
Ovary	Adj. R^2	P-val	Adj. R^2	P-val
Total regulators	-0.1041	0.8158	-0.117	0.8167
PIWI	0.08926	0.193	-0.1226	0.8998
siRNA	-0.08073	0.6271	-0.1206	0.8638
Transcription	-0.09962	0.7661	-0.0011	0.3489
Post-Transcription	-0.09589	0.7318	-0.05284	0.4802

Supplementary File 8. Phylogenetically independent contrast Spearman rank-order correlation testing for correlations between recent-TE and TE regulatory mechanism proportional expression across vertebrate species.

Correlation coefficient (ρ)

All Tissues	TE	PIWI	siRNA	Transcription	Post-Transcription	Total regulators
TE	X	0.27	0.02	0.14	-0.18	0.12
PIWI	0.640	X	0.44	0.42	-0.20	0.51
siRNA	0.960	0.340	X	0.19	0.19	0.38
Transcription	0.737	0.344	0.682	X	0.62	0.97
Post-Transcription	0.682	0.682	0.682	0.094	X	0.64
Total regulators	0.746	0.230	0.399	1.26E-06	0.088	X

Adjusted p-values

Correlation coefficient (ρ)

Germline Tissues	TE	PIWI	siRNA	Transcription	Post-Transcription	Total regulators
TE	X	0.51	0.04	0.30	-0.34	0.24
PIWI	0.223	X	0.55	0.54	-0.01	0.77
siRNA	0.975	0.202	X	0.27	0.03	0.60
Transcription	0.504	0.202	0.551	X	0.49	0.81
Post-Transcription	0.463	0.975	0.975	0.231	X	0.41
Total regulators	0.575	0.020	0.151	0.011	0.337	X

Adjusted p-values

Correlation coefficient (ρ)

Somatic Tissues	TE	PIWI	siRNA	Transcription	Post-Transcription	Total regulators
TE	X	-0.47	0.00	0.09	-0.15	-0.01
PIWI	0.240	X	0.17	0.72	0.44	0.73
siRNA	0.993	0.762	X	0.17	0.23	0.33
Transcription	0.888	0.021	0.762	X	0.62	0.92
Post-Transcription	0.762	0.274	0.699	0.074	X	0.82
Total regulators	0.993	0.020	0.490	8.92E-05	0.004	X

Adjusted p-values

Correlation coefficient (ρ)

Testis	TE	PIWI	siRNA	Transcription	Post-Transcription	Total regulators
TE	X	0.56	-0.02	0.22	-0.24	0.37
PIWI	0.112	X	0.61	0.34	0.03	0.90
siRNA	0.954	0.082	X	0.12	-0.05	0.62
Transcription	0.628	0.459	0.839	X	0.67	0.61
Post-Transcription	0.617	0.954	0.954	0.065	X	0.28
Total regulators	0.418	2.75E-04	0.082	0.082	0.556	X

Adjusted p-values

Correlation coefficient (ρ)

Ovary	TE	PIWI	siRNA	Transcription	Post-Transcription	Total regulators
TE	X	-0.37	-0.20	0.20	-0.16	0.07
PIWI	0.420	X	0.34	0.39	-0.16	0.43
siRNA	0.667	0.428	X	0.44	0.56	0.62
Transcription	0.667	0.418	0.363	X	0.34	0.91
Post-Transcription	0.677	0.677	0.204	0.428	X	0.55
Total regulators	0.830	0.363	0.158	4.60E-04	0.204	X

Adjusted p-values

REFERENCES

CLC Genomics Workbench 9.0.1. https://www.qiagenbioinformatics.com/

Abrieu, A., M. Doree, and D. Fisher. 2001. The interplay between cyclin-B-Cdc2 kinase (MPF) and MAP kinase during maturation of oocytes. *Journal of Cell Science* 114:257-267.

Adams, R. H., H. Blackmon, J. Reyes-Velasco, D. R. Schield, D. C. Card, A. L. Andrew, N. Waynewood, and T. A. Castoe. 2016. Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. *Genome* 59:295-310.

Adams, R. H., D. R. Schield, D. C. Card, and T. A. Castoe. 2018. Assessing the impacts of positive selection on coalescent-based species tree estimation and species delimitation. *Systematic Biology* 67(6): 1076-1090.

Agrawal, A., Q. M. Eastman, and D. G. Schatz. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394:744-751.

Agren, J. A. and S. I. Wright. 2011. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? *Chromosome Research* 19:777-786.

Alfoldi, J., F. Di Palma, M. Grabherr, C. Williams, L. S. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. E. Glor, J. D. Jaffe, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477:587-591.

Anders, S. and W. Huber. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11.

Anders, S., D. J. McCarthy, Y. S. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* 8:1765-1786.

Andrew, A. L., B. W. Perry, D. C. Card, D. R. Schield, R. P. Ruggiero, S. E. McGaugh, A. Choudhary, S. M. Secor, and T. A. Castoe. 2017. Growth and stress response mechanisms underlying post-feeding regenerative organ growth in the Burmese python. *BMC Genomics* 18:338.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data.

Anwar, S. L., W. Wulaningsih, and U. Lehmann. 2017. Transposable elements in human cancer: Causes and consequences of deregulation. *International Journal of Molecular Sciences* 18.

Aravin, A. and T. Tuschl. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett* 579:5830-5840.

Aravin, A. A., R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K. F. Toth, T. Bestor, and G. J. Hannon. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular Cell* 31:785-799.

Backstrom, N., W. Forstmeier, H. Schielzeth, H. Mellenius, K. Nam, E. Bolund, M. T. Webster, T. Ost, M. Schneider, B. Kempenaers, et al. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research* 20:485-495.

Baillie, J. K., M. W. Barnett, K. R. Upton, D. J. Gerhardt, T. A. Richmond, F. De Sapio, P. M. Brennan, P. Rizzu, S. Smith, M. Fell, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479:534-537.

Baker, R. J., J. J. Bull, and G. A. Mengden. 1972. Karyotypic studies of 38 species of North-American snakes. *Copeia*:257-265.

Bao, W., K. K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6:11.

Barron, M. G., A. S. Fiston-Lavier, D. A. Petrov, and J. Gonzalez. 2014. Population genomics of transposable elements in drosophila. *Annual Review of Genetics, Vol 48* 48:561-581.

Beck, C. R., J. L. Garcia-Perez, R. M. Badge, and J. V. Moran. 2011. LINE-1 elements in structural variation and disease. *Annual Review of Genomics and Human Genetics, Vol 12* 12:187-215.

Beckstette, M., R. Homann, R. Giegerich, and S. Kurtz. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* 7.

Bedrosian, T. A., S. Linker, and F. H. Gage. 2016. Environment-driven somatic mosaicism in brain disorders. *Genome Medicine* 8.

Bellott, D. W., H. Skaletsky, T. J. Cho, L. Brown, D. Locke, N. Chen, S. Galkina, T. Pyntikova, N. Koutseva, T. Graves, et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nature Genetics* 49:387-394.

Benton, M. J. and P. C. J. Donoghue. 2007. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution* 24:26-53.

Bernt, M., A. Donath, F. Juhling, F. Externbrink, C. Florentz, G. Fritzsch, J. Putz, M. Middendorf, and P. F. Stadler. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* 69:313-319.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14:708-715.

Blass, E., M. Bell, and S. Boissinot. 2012. Accumulation and rapid decay of Non-LTR retrotransposons in the genome of the three-spine stickleback. *Genome Biology and Evolution* 4:687-702.

Blumenstiel, J. P., X. Chen, M. M. He, and C. M. Bergman. 2014. An age-of-allele test of neutrality for transposable element insertions. *Genetics* 196:523-538.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578-579.

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

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

Boitard, S., W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz. 2016. Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian computation approach. *Plos Genetics* 12:e1005877.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.

Bouckaert, R., J. Heled, D. Kuhnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10:e1003537.

Bourque, G., B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J. L. Chew, Y. Ruan, C. L. Wei, H. H. Ng, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research* 18:1752-1762.

Braasch, I., A. R. Gehrke, J. J. Smith, K. Kawasaki, T. Manousaki, J. Pasquier, A. Amores, T. Desvignes, P. Batzel, J. Catchen, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* 48(4):427.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2:10.

Brouha, B., J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, and H. H. Kazazian, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* 100:5280-5285.

Burns, K. H. 2017. Transposable elements in cancer. *Nat Rev Cancer* 17:415-424.

Bushnell, B. 2014. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Callinan, P. A. and M. A. Batzer. 2006. Retrotransposable elements and human disease. *Genome Dyn* 1:104-115.

Canapa, A., M. Barucca, M. A. Biscotti, M. Forconi, and E. Olmo. 2016. Transposons, genome size, and evolutionary insights in animals. *Cytogenetic and Genome Research* 147:217-239.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado, and M. Yandell. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18:188-196.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335-336.

Carbon, S., A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, G. O. H. Ami, and G. Web Presence Working. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* 25:288-289.

Card, D. C., R. H. Adams, D. R. Schield, B. W. Perry, A. B. Corbin, G. I. Pasquesi, K. Row, M. J. Van Kleeck, J. M. Daza, and W. Booth. 2019. Genomic basis of convergent island phenotypes in boa constrictors. *Genome Biology and Evolution*.

Casewell, N. R., R. A. Harrison, W. Wuster, and S. C. Wagstaff. 2009. Comparative venom gland transcriptome surveys of the saw-scaled vipers (Viperidae: *Echis*) reveal substantial intra-family gene diversity and novel venom transcripts. *BMC Genomics* 10.

Casewell, N. R., G. A. Huttley, and W. Wuster. 2012. Dynamic evolution of venom proteins in squamate reptiles. *Nature Communications* 3.

Casewell, N. R., S. C. Wagstaff, R. A. Harrison, C. Renjifo, and W. Wuster. 2011. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Molecular Biology and Evolution* 28:2637-2649.

Casper, J., A. S. Zweig, C. Villarreal, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, D. Karolchik, et al. 2018. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* 46:D762-D769.

Castoe, T. A., A. P. J. de Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K. Fujita, R. P. Ruggiero, J. F. Degner, J. M. Daza, W. J. Gu, et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *PNAS* 110:20645-20650.

Castoe, T. A., K. T. Hall, M. L. G. Mboulas, W. J. Gu, A. P. J. de Koning, S. E. Fox, A. W. Poole, V. Vemulapalli, J. M. Daza, T. Mockler, et al. 2011. Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biology and Evolution* 3:641-653.

Castoe, T. A., A. W. Poole, W. Gu, A. P. Jason de Koning, J. M. Daza, E. N. Smith, and D. D. Pollock. 2010. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Mol Ecol Resour* 10:341-347.

Chalopin, D., M. Naville, F. Plard, D. Galiana, and J. N. Volff. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution* 7:567-580.

Chapman, J. A., I. Ho, S. Sunkara, S. J. Luo, G. P. Schroth, and D. S. Rokhsar. 2011. Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* 6.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10:195-205.

Charlesworth, B., P. Sniegowski, and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in Eukaryotes. *Nature* 371:215-220.

Chuong, E. B., N. C. Elde, and C. Feschotte. 2016. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*.

Clark, K., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2016. GenBank. *Nucleic Acids Res* 44:D67-72.

Cohn, M. J. and C. Tickle. 1999. Developmental basis of limblessness and axial patterning in snakes. *Nature* 399:474-479.

Consortium, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-74.

Cordaux, R. and M. A. Batzer. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10:691-703.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.

Darrow, E. M., M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S. C. Huang, A. L. Sanborn, I. Machol, M. Shamim, et al. 2016. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *PNAS* 113:E4504-E4512.

De Cecco, M., S. W. Criscione, A. L. Peterson, N. Neretti, J. M. Sedivy, and J. A. Kreiling. 2013. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging (Albany NY)* 5:867-883.

de Koning, A. P. J., W. J. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. 2011. Repetitive elements may comprise over two-thirds of the human genome. *Plos Genetics* 7.

Deakin, J. E., M. J. Edwards, H. Patel, D. O'Meally, J. M. Lian, R. Stenhouse, S. Ryan, A. M. Livernois, B. Azad, C. E. Holleley, et al. 2016. Anchoring genome sequence to chromosomes of the central bearded dragon (*Pogona vitticeps*) enables reconstruction of

ancestral squamate macrochromosomes and identifies sequence content of the Z chromosome. *BMC Genomics* 17.
Deininger, P., M. E. Morales, T. B. White, M. Baddoo, D. J. Hedges, G. Servant, S. Srivastav, M. E. Smither, M. Concha, D. L. DeHaro, et al. 2017. A comprehensive approach to expression of L1 loci. *Nucleic Acids Res* 45:e31.
Dixon, J. R., D. U. Gorkin, and B. Ren. 2016. Chromatin domains: The unit of chromosome organization. *Molecular Cell* 62:668-680.
Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, et al. 2012. Landscape of transcription in human cells. *Nature* 489:101-108.
Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.
Dowell, N. L., M. W. Giorgianni, V. A. Kassner, J. E. Selegue, E. E. Sanchez, and S. B. Carroll. 2016. The deep origin and recent loss of venom toxin genes in rattlesnakes. *Current Biology* 26:2434-2445.
Drummond, A. J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.
Dunn, C. W., F. Zapata, C. Munro, S. Siebert, and A. Hejnol. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *PNAS* 115:E409-E417.
Durand, N. C., M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3:95-98.
Duret, L. and N. Galtier. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* 10:285-311.
Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.
Elliott, T. A. and T. R. Gregory. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc Lond B Biol Sci* 370:20140331.
Ernst, C., D. T. Odom, and C. Kutter. 2017. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nat Commun* 8:1411.
Fane, M., L. Harris, A. G. Smith, and M. Piper. 2017. Nuclear factor one transcription factors as epigenetic regulators in cancer. *International Journal of Cancer* 140:2634-2641.
Faulkner, G. J. and V. Billon. 2018. L1 retrotransposition in the soma: a field jumping ahead. *Mob DNA* 9:22.
Faulkner, G. J. and J. L. Garcia-Perez. 2017. L1 mosaicism in mammals: Extent, effects, and evolution. *TRENDS in Genetics* 33:802-816.
Faulkner, G. J., Y. Kimura, C. O. Daub, S. Wani, C. Plessy, K. M. Irvine, K. Schroder, N. Cloonan, A. L. Steptoe, T. Lassmann, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics* 41:563-571.
Feldman, A., N. Sabath, R. A. Pyron, I. Mayrose, and S. Meiri. 2016. Body sizes and diversification rates of lizards, snakes, amphisbaenians and the tuatara. *Global Ecology and Biogeography* 25:187-197.

Figuet, E., B. Nabholz, M. Bonneau, E. M. Carrio, K. Nadachowska-Brzyska, H. Ellegren, and N. Galtier. 2016. Life history traits, protein evolution, and the nearly neutral theory in Amniotes. *Molecular Biology and Evolution* 33:1517-1527.

Fryxell, K. J. and W. J. Moon. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution* 22:650-658.

Fujita, M. K., S. V. Edwards, and C. P. Ponting. 2011. The Anolis lizard genome: an amniote genome without isochores. *Genome Biology and Evolution* 3:974-984.

Gamble, T., T. A. Castoe, S. V. Nielsen, J. L. Banks, D. C. Card, D. R. Schield, G. W. Schuett, and W. Booth. 2017. The discovery of XY sex chromosomes in a boa and python. *Current Biology* 27:2148-2153.

Gao, J., Q. Li, Z. Wang, Y. Zhou, P. Martelli, F. Li, Z. Xiong, J. Wang, H. Yang, and G. Zhang. 2017. Sequencing, de novo assembling, and annotating the genome of the endangered Chinese crocodile lizard *Shinisaurus crocodilurus*. *Gigascience* 6:gix041.

Garcia-Perez, J. L., T. J. Widmann, and I. R. Adams. 2016. The impact of transposable elements on mammalian development. *Development* 143:4101-4114.

Gasior, S. L., T. P. Wakeman, B. Xu, and P. L. Deininger. 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of Molecular Biology* 357:1383-1393.

Geffeney, S., E. D. Brodie, P. C. Ruben, and E. D. Brodie. 2002. Mechanisms of adaptation in a predator-prey arms race: TTX-resistant sodium channels. *Science* 297:1336-1339.

Georges, A., Q. Li, J. Lian, D. O'Meally, J. Deakin, Z. Wang, P. Zhang, M. Fujita, H. R. Patel, C. E. Holleley, et al. 2015. High-coverage sequencing and annotated assembly of the genome of the Australian dragon lizard *Pogona vitticeps*. *Gigascience* 4:45.

Gilbert, C., S. S. Hernandez, J. Flores-Benabib, E. N. Smith, and C. Feschotte. 2012. Rampant horizontal transfer of SPIN transposons in squamate reptiles. *Molecular Biology and Evolution* 29:503-515.

Gilbert, C., J. M. Meik, D. Dashevsky, D. C. Card, T. A. Castoe, and S. Schaack. 2014. Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proceedings of the Royal Society B-Biological Sciences* 281.

Gilbert, C., S. Schaack, J. K. Pace, 2nd, P. J. Brindley, and C. Feschotte. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* 464:1347-1350.

Goodier, J. L. 2016. Restricting retrotransposons: a review. *Mob DNA* 7:16.

Goubert, C., J. Thomas, L. M. Payer, J. M. Kidd, J. Feusier, W. S. Watkins, K. H. Burns, L. B. Jorde, and C. Feschotte. 2019. TypeTE: a tool to genotype mobile element insertions from whole genome resequencing data. *bioRxiv*:791665.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. D. Zeng, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644-U130.

Graves, J. A. M. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. *Nature Reviews Genetics* 17:33-46.

Green, R. E., E. L. Braun, J. Armstrong, D. Earl, N. Nguyen, G. Hickey, M. W. Vandewege, J. A. St John, S. Capella-Gutierrez, T. A. Castoe, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* 346:1254449.

Gregory, T. R. 2017. Animal Genome Size Database. http://www.genomesize.com

Gronostajski, R. M. 2000. Roles of the NFI/CTF gene family in transcription and development. *Gene* 249:31-45.
Hackett, J. A., T. Kobayashi, S. Dietmann, and M. A. Surani. 2017. Activation of lineage regulators and transposable elements across a pluripotent spectrum. *Stem Cell Reports* 8:1645-1658.
Hajkova, P., S. Erhardt, N. Lane, T. Haaf, O. El-Maarri, W. Reik, J. Walter, and M. A. Surani. 2002. Epigenetic reprogramming in mouse primordial germ cells. *Mechanisms of Development* 117:15-23.
Handel, M. A. and J. C. Schimenti. 2010. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat Rev Genet* 11:124-136.
Hangauer, M. J., I. W. Vaughn, and M. T. McManus. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *Plos Genetics* 9:e1003569.
Hanna, C. B. and J. D. Hennebold. 2014. Ovarian germline stem cells: an unlimited source of oocytes? *Fertil Steril* 101:20-30.
Hargreaves, A. D., M. T. Swain, M. J. Hegarty, D. W. Logan, and J. F. Mulley. 2014. Genomic and transcriptomic insights into the regulation of snake venom production. *bioRxiv*:008474.
Harrell, F. H. J. 2019. Hmisc V4.2-0.
Harris, R. S. 2007. Improved pairwise alignmnet of genomic DNA.
He, J., X. Fu, M. Zhang, F. He, W. Li, M. M. Abdul, J. Zhou, L. Sun, C. Chang, Y. Li, et al. 2019. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat Commun* 10:34.
Hellen, E. H. and J. F. Brookfield. 2013. Alu elements in primates are preferentially lost from areas of high GC content. *PeerJ* 1:e78.
Hillier, L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. M. Groenen, M. E. Delany, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695-716.
Houwing, S., L. M. Kamminga, E. Berezikov, D. Cronembold, A. Girard, H. van den Elst, D. V. Filippov, H. Blaser, E. Raz, C. B. Moens, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* 129:69-82.
Huang, C. R. L., K. H. Burns, and J. D. Boeke. 2012. Active transposition in genomes. *Annual Review of Genetics, Vol 46* 46:651-675.
Hutchins, A. P. and D. Pei. 2015. Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. *Sci Bull (Beijing)* 60:1722-1733.
Ikeda, N., T. Chijiwa, K. Matsubara, N. Oda-Ueda, S. Hattori, Y. Matsuda, and M. Ohno. 2010. Unique structural characteristics and evolution of a cluster of venom phospholipase A2 isozyme genes of *Protobothrops flavoviridis* snake. *Gene* 461:15-25.
Jacobs, F. M. J., D. Greenberg, N. Nguyen, M. Haeussler, A. D. Ewing, S. Katzman, B. Paten, S. R. Salama, and D. Haussler. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* 516:242.
Jiao, W. B., G. G. Accinelli, B. Hartwig, C. Kiefer, D. Baker, E. Severing, E. M. Willing, M. Piednoel, S. Woetzel, E. Madrid-Herrero, et al. 2017. Improving and correcting the

contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research* 27:778-786.
Jin, Y., O. H. Tam, E. Paniagua, and M. Hammell. 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* 31:3593-3599.
Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res* 36:W5-9.
Julien, P., D. Brawand, M. Soumillon, A. Necsulea, A. Liechti, F. Schutz, T. Daish, F. Grutzner, and H. Kaessmann. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *Plos Biology* 10:e1001328.
Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110:462-467.
Kapusta, A., A. Suh, and C. Feschotte. 2017. Dynamics of genome size evolution in birds and mammals. *PNAS* 114:E1460-E1469.
Kato, Y., M. Kaneda, K. Hata, K. Kumaki, M. Hisano, Y. Kohara, M. Okano, E. Li, M. Nozaki, and H. Sasaki. 2007. Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Human Molecular Genetics* 16:2272-2280.
Kazazian, H. H. 1998. Mobile elements and disease. *Current Opinion in Genetics & Development* 8:343-350.
Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *PNAS* 100:11484-11489.
Kerchove, C. M., M. S. Luna, M. B. Zablith, M. F. Lazari, S. S. Smaili, and N. Yamanouye. 2008. α1-adrenoceptors trigger the snake venom production cycle in secretory cells by activating phosphatidylinositol 4, 5-bisphosphate hydrolysis and ERK signaling pathway. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 150:431-437.
Kim, M. and W. McGinnis. 2011. Phosphorylation of Grainy head by ERK is essential for wound-dependent regeneration but not for development of an epidermal barrier. *PNAS* 108:650-655.
Kirino, Y., N. Kim, M. de Planell-Saguer, E. Khandros, S. Chiorean, P. S. Klein, I. Rigoutsos, T. A. Jongens, and Z. Mourelatos. 2009. Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. *Nature Cell Biology* 11:652-U478.
Kohany, O., A. J. Gentles, L. Hankus, and J. Jurka. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7:474.
Kolde, R. 2012. Pheatmap: pretty heatmaps. *R package version* 61:915.
Kopylova, E., L. Noe, and H. Touzet. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28:3211-3217.
Kordis, D. and F. Gubensek. 1997. Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A2 genes and in genomes of Viperidae snakes. *Eur J Biochem* 246:772-779.

Kordis, D. and F. Gubensek. 1998. Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. *PNAS* 95:10704-10709.
Kordis, D. and F. Gubenšek. 1998. The Bov-B lines found in *Vipera ammodytes* toxic PLA2 genes are widespread in snake genomes. *Toxicon* 36:1585-1590.
Kreiling, J. A., B. C. Jones, J. G. Wood, M. De Cecco, S. W. Criscione, N. Neretti, S. L. Helfand, and J. M. Sedivy. 2017. Contribution of retrotransposable elements to aging. Pp. 297-321. Human retrotransposons in health and disease. Springer.
Kumar, S., G. Stecher, M. Suleski, and S. B. Hedges. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* 34:1812-1819.
Lachumanan, R., A. Armugam, C. H. Tan, and K. Jeyaseelan. 1998. Structure and organization of the cardiotoxin genes in *Naja naja sputatrix*. *Febs Letters* 433:119-124.
Lanfear, R., B. Calcott, S. Y. Ho, and S. Guindon. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29:1695-1701.
Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.
Le Rouzic, A., T. S. Boutin, and P. Capy. 2007. Long-term evolution of transposable elements. *PNAS* 104:19375-19380.
Le Rouzic, A., T. Payen, and A. Hua-Van. 2013. Reconstructing the evolutionary history of transposable elements. *Genome Biology and Evolution* 5:77-86.
Leinonen, R., H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. *Nucleic Acids Res* 39:D19-21.
Levine, A. J., D. T. Ting, and B. D. Greenbaum. 2016. P53 and the defenses against genome instability caused by transposons and repetitive elements. *Bioessays* 38:508-513.
Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
Li, H. and R. Durbin. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493-U484.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.
Li, W. Z. and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.
Liao, Y., G. K. Smyth, and W. Shi. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923-930.
Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289-293.
Lim, R. S. and T. Kai. 2015. A piece of the pi(e): The diverse roles of animal piRNAs and their PIWI partners. *Semin Cell Dev Biol* 47-48:17-31.
Lim, S. L., E. Tsend-Ayush, R. D. Kortschak, R. Jacob, C. Ricciardelli, M. K. Oehler, and F. Grutzner. 2013. Conservation and expression of PIWI-interacting RNA pathway genes in male and female adult gonad of amniotes. *Biol Reprod* 89:136.

Lin, C. Y., V. B. Vega, J. S. Thomsen, T. Zhang, S. L. Kong, M. Xie, K. P. Chiu, L. Lipovich, D. H. Barnett, F. Stossi, et al. 2007. Whole-genome cartography of estrogen receptor alpha binding sites. *Plos Genetics* 3:e87.

Lippman, Z., A. V. Gendrel, M. Black, M. W. Vaughn, N. Dedhia, W. R. McCombie, K. Lavine, V. Mittal, B. May, K. D. Kasschau, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430:471-476.

Liu, B., Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, and W. Fan. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.2012*.

Liu, Y., Q. Zhou, Y. Wang, L. Luo, J. Yang, L. Yang, M. Liu, Y. Li, T. Qian, Y. Zheng, et al. 2015. *Gekko japonicus* genome reveals evolution of adhesive toe pads and tail regeneration. *Nat Commun* 6:10033.

Loreto, E. L. S. and C. M. Pereira. 2017. Somatizing the transposons action. *Mob Genet Elements* 7:1-9.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.

Lynch, M. and J. S. Conery. 2003. The origins of genome complexity. *Science* 302:1401-1404.

Lynch, M. and B. Walsh. 2007. *The origins of genome architecture*. Sinauer Associates Sunderland.

Lynch, V. J., M. C. Nnamani, A. Kapusta, K. Brayer, S. L. Plaza, E. C. Mazur, D. Emera, S. Z. Sheikh, F. Grutzner, S. Bauersachs, et al. 2015. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep* 10:551-561.

Mackessy, S. 2008. Venom composition in rattlesnakes: trends and biological significance. *The biology of rattlesnakes* 495:510.

Mackessy, S. P. 2010. The field of reptile toxinology: snakes, lizards and their venoms. *Handbook of Venoms and Toxins of Reptiles*. CRC Press, Boca Raton.

Makałowski, W., T. Kischka, and I. Makałowska. 2017. Contribution of transposable elements to human proteins. *eLS*.

Malki, S., G. W. van der Heijden, K. A. O'Donnell, S. L. Martin, and A. Bortvin. 2014. A role for retrotransposon LINE-1 in fetal oocyte attrition in mice. *Developmental Cell* 29:521-533.

Marcais, G. and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-770.

Marin, R., D. Cortez, F. Lamanna, M. M. Pradeepa, E. Leushkin, P. Julien, A. Liechti, J. Halbert, T. Bruning, K. Mossinger, et al. 2017. Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome Research* 27:1974-1987.

Matsubara, K., H. Tarui, M. Toriba, K. Yamada, C. Nishida-Umehara, K. Agata, and Y. Matsuda. 2006. Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *PNAS* 103:18190-18195.

Mazet, O., W. Rodriguez, and L. Chikhi. 2015. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology* 104:46-58.

Mazet, O., W. Rodriguez, S. Grusea, S. Boitard, and L. Chikhi. 2016. On the importance of being structured: instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity* 116:362-371.
Molaro, A., I. Falciatori, E. Hodges, A. A. Aravin, K. Marran, S. Rafii, W. R. McCombie, A. D. Smith, and G. J. Hannon. 2014. Two waves of de novo methylation during mouse germ cell development. *Genes & Development* 28:1544-1549.
Murphy, W. J., T. H. Pringle, T. A. Crider, M. S. Springer, and W. Miller. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research* 17:413-421.
Nadachowska-Brzyska, K., R. Burri, L. Smeds, and H. Ellegren. 2016. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology* 25:1058-1072.
Nadalin, F., F. Vezzi, and A. Policriti. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13.
Navarro, F., J. Hoops, L. Bellfy, E. Cerveira, Q. Zhu, C. Zhang, C. Lee, and M. Gerstein. 2019. TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements. *bioRxiv*:648667.
Neafsey, D. E., J. P. Blumenstiel, and D. L. Hartl. 2004. Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies. *Molecular Biology and Evolution* 21:2310-2318.
Neff, B. D. and M. R. Gross. 2001. Microsatellite evolution in vertebrates: Inference from AC dinucleotide repeats. *Evolution* 55:1717-1733.
Nielsen, R. and M. A. Beaumont. 2009. Statistical inferences in phylogeography. *Molecular Ecology* 18:1034-1047.
Novick, P., J. Smith, D. Ray, and S. Boissinot. 2010. Independent and parallel lateral transfer of DNA transposons in tetrapod genomes. *Gene* 449:85-94.
O'Connor, R. E., M. N. Romanov, L. G. Kiazim, P. M. Barrett, M. Farre, J. Damas, M. Ferguson-Smith, N. Valenzuela, D. M. Larkin, and D. K. Griffin. 2018. Reconstruction of the diapsid ancestral genome permits chromosome evolution tracing in avian and non-avian dinosaurs. *Nature Communications* 9.
Olmo, E. 2005. Rate of chromosome changes and speciation in reptiles. *Genetica* 125:185-203.
Organ, C. L., R. G. Moreno, and S. V. Edwards. 2008. Three tiers of genome evolution in reptiles. *Integrative and Comparative Biology* 48:494-504.
Orozco-TerWengel, P. 2016. The devil is in the details: the effect of population structure on demographic inference. *Heredity* 116:349-350.
Oshlack, A., M. D. Robinson, and M. D. Young. 2010. From RNA-seq reads to differential expression results. *Genome Biology* 11.
Ozata, D. M., I. Gainetdinov, A. Zoch, D. O'Carroll, and P. D. Zamore. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* 20:89-108.
Paradis, E. 2010. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 26:419-420.
Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.
Pasquesi, G. I. M., R. H. Adams, D. C. Card, D. R. Schield, A. B. Corbin, B. W. Perry, J. Reyes-Velasco, R. P. Ruggiero, M. W. Vandewege, J. A. Shortt, et al. 2018. Squamate reptiles

challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nature Communications* 9.
Perry, B. W., D. C. Card, J. W. McGlothlin, G. I. M. Pasquesi, R. H. Adams, D. R. Schield, N. R. Hales, A. B. Corbin, J. P. Demuth, F. G. Hoffmann, et al. 2018. Molecular adaptations for sensing and securing prey and insight into amniote genome diversity from the garter snake genome. *Genome Biology and Evolution* 10:2110-2129.
Petrov, D. A. 2002. Mutational equilibrium model of genome size evolution. *Theoretical Population Biology* 61:531-544.
Petrov, D. A., Y. T. Aminetzach, J. C. Davis, D. Bensasson, and A. E. Hirsh. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. *Molecular Biology and Evolution* 20:880-892.
Petrov, D. A., A. S. Fiston-Lavier, M. Lipatov, K. Lenkov, and J. Gonzalez. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. *Molecular Biology and Evolution* 28:1633-1644.
Piskurek, O. and N. Okada. 2007. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. *PNAS* 104:12046-12051.
Platt, R. N., S. F. Mangum, and D. A. Ray. 2016. Pinpointing the vesper bat transposon revolution using the Miniopterus natalensis genome. *Mobile DNA* 7.
Platt, R. N., M. W. Vandewege, and D. A. Ray. 2018. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research* 26:25-43.
Ponnusamy, M., K. W. Yan, C. Y. Liu, P. F. Li, and K. Wang. 2017. PIWI family emerging as a decisive factor of cell fate: An overview. *European Journal of Cell Biology* 96:746-757.
Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* 26:342-350.
Pyron, R. A., F. T. Burbrink, and J. J. Wiens. 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol Biol* 13:93.
Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
Radis-Baptista, G., T. Kubo, N. Oguiura, M. Svartman, T. M. B. Almeida, R. F. Batistic, E. B. Oliveira, A. M. Vianna-Morgante, and T. Yamane. 2003. Structure and chromosomal localization of the gene for crotamine, a toxin from the South American rattlesnake, *Crotalus durissus terrificus*. *Toxicon* 42:747-752.
Rambaut, A. and A. J. Drummond. 2007. Tracer v1. 4.
Ramirez, F., V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Gruning, J. Villaveces, B. Habermann, A. Akhtar, and T. Manke. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications* 9.
Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. 2015. A 3D Map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 162:687-688.
Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447:425-432.
The UniProt Consortium. 2018. UniProt: the universal protein knowledgebase *Nucleic Acids Research* 46:2699-2699.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3:217-223.

Reyes-Velasco, J., D. C. Card, A. L. Andrew, K. J. Shaney, R. H. Adams, D. R. Schield, N. R. Casewell, S. P. Mackessy, and T. A. Castoe. 2015. Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. *Molecular Biology and Evolution* 32:173-183.

Reznik, B., S. A. Cincotta, R. G. Jaszczak, L. J. Mateo, J. Shen, M. Cao, L. Baskin, P. Ye, W. An, and D. J. Laird. 2019. Heterogeneity of transposon expression and activation of the repressive network in human fetal germ cells. *Development* 146.

Rice, E. S., S. Kohno, J. St John, S. Pham, J. Howard, L. F. Lareau, B. L. O'Connell, G. Hickey, J. Armstrong, A. Deran, et al. 2017. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Research* 27:686-696.

Richardson, S. R., P. Gerdes, D. J. Gerhardt, F. J. Sanchez-Luque, G. O. Bodea, M. Munoz-Lopez, J. S. Jesuadian, M. H. C. Kempen, P. E. Carreira, J. A. Jeddeloh, et al. 2017. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Res* 27:1395-1405.

Rizzon, C., G. Marais, M. Gouy, and C. Biemont. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res* 12:400-407.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140.

Robinson, M. D. and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11.

Rokyta, D. R., A. R. Lemmon, M. J. Margres, and K. Aronow. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). *BMC Genomics* 13.

Ruggiero, R. P., Y. Bourgeois, and S. Boissinot. 2017. LINE insertion polymorphisms are abundant but at low frequencies across populations of *Anolis carolinensis*. *Frontiers in Genetics* 8.

Sagata, N. 1996. Meiotic metaphase arrest in animal oocytes: Its mechanisms and biological significance. *Trends in Cell Biology* 6:22-28.

Sayers, E. W., R. Agarwala, E. E. Bolton, J. R. Brister, K. Canese, K. Clark, R. Connor, N. Fiorini, K. Funk, T. Hefferon, et al. 2019. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 47:D23-D28.

Schield, D. R., D. C. Card, N. R. Hales, B. W. Perry, G. I. M. Pasquesi, H. Blackmon, R. H. Adams, A. B. Corbin, C. F. Smith, B. Ramesh, et al. 2019. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* 29(4):590-601.

Schrider, D. R., A. G. Shanku, and A. D. Kern. 2016. Effects of linked selective sweeps on demographic inference and model selection. *Genetics* 204:1207-1223.

Secor, S. M. and J. Diamond. 1998. A vertebrate model of extreme physiological regulation. *Nature* 395:659-662.

Sen, S. K., K. D. Han, J. X. Wang, J. Lee, H. Wang, P. A. Callinan, M. Dyer, R. Cordaux, P. Liang, and M. A. Batzer. 2006. Human genomic deletions mediated by recombination between Alu elements. *American Journal of Human Genetics* 79:41-53.
Shi, X., A. Seluanov, and V. Gorbunova. 2007. Cell divisions are required for L1 retrotransposition. *Mol Cell Biol* 27:1264-1270.
Silva, J. C., E. L. Loreto, and J. B. Clark. 2004. Factors that affect the horizontal transfer of transposable elements. *Current Issues in Molecular Biology* 6:57-71.
Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-3212.
Siomi, M. C., K. Sato, D. Pezic, and A. A. Aravin. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12:246-258.
Slotkin, R. K. and R. Martienssen. 2007. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics* 8:272-285.
Smeds, L., T. Kawakami, R. Burri, P. Bolivar, A. Husby, A. Qvarnstrom, S. Uebbing, and H. Ellegren. 2014. Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. *Nature Communications* 5.
Smit, A.F.A., R. Hubley, 2008-2017. RepeatModeler Open-1.0.9.
Smit, A.F.A., R. Hubley, P. Green. RepeatMasker Genomic Datasets (last accessed 2019).
Smit, A.F.A., R. Hubley, P. Green. 2015-2019. RepeatMasker Open-4.0.
Solovyev, V., P. Kosarev, I. Seledsov, and D. Vorobyev. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* 7 Suppl 1:S10 11-12.
Song, B., S. F. Cheng, Y. B. Sun, X. Zhong, J. Q. Jin, R. Guan, R. W. Murphy, J. Che, Y. P. Zhang, and X. Liu. 2015. A genome draft of the legless anguid lizard, *Ophisaurus gracilis*. *Gigascience* 4.
Song, M. Z. and S. Boissinot. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* 390:206-213.
Soumillon, M., A. Necsulea, M. Weier, D. Brawand, X. Zhang, H. Gu, P. Barthes, M. Kokkinaki, S. Nef, A. Gnirke, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* 3:2179-2190.
Srikulnath, K., C. Nishida, K. Matsubara, Y. Uno, A. Thongpan, S. Suputtitada, S. Apisitwanich, and Y. Matsuda. 2009. Karyotypic evolution in squamate reptiles: comparative gene mapping revealed highly conserved linkage homology between the butterfly lizard (*Leiolepis reevesii rubritaeniata*, Agamidae, Lacertilia) and the Japanese four-striped rat snake (*Elaphe quadrivirgata*, Colubridae, Serpentes). *Chromosome Research* 17:975-986.
Stanke, M. and B. Morgenstern. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* 33:W465-W467.
Stein, P., N. V. Rozhkov, F. Li, F. L. Cardenas, O. Davydenko, L. E. Vandivier, B. D. Gregory, G. J. Hannon, and R. M. Schultz. 2015. Essential role for endogenous siRNAs during meiosis in mouse oocytes. *Plos Genetics* 11:e1005013.
Sudmant, P. H., M. S. Alexis, and C. B. Burge. 2015. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol* 16:287.
Suh, A., G. Churakov, M. P. Ramakodi, R. N. Platt, J. Jurka, K. K. Kojima, J. Caballero, A. F. Smit, K. A. Vliet, F. G. Hoffmann, et al. 2015. Multiple lineages of ancient CR1

retroposons shaped the early genome evolution of Amniotes. *Genome Biology and Evolution* 7:205-217.
Sun, Y. H., L. H. Xie, X. Y. Zhuo, Q. A. Chen, D. L. Ghooneim, B. Zhang, J. Jagne, C. B. Yang, and X. Z. Li. 2017. Domestic chickens activate a piRNA defense against avian leukosis virus. *Elife* 6.
Surani, M. A., K. Hayashi, and P. Hajkova. 2007. Genetic and epigenetic regulators of pluripotency. *Cell* 128:747-762.
R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.
Ting, S. B., J. Caddy, N. Hislop, T. Wilanowski, A. Auden, L. L. Zhao, S. Ellis, P. Kaur, Y. Uchida, W. M. Holleran, et al. 2005. A homolog of Drosophila grainy head is essential for epidermal integrity in mice. *Science* 308:411-413.
Tollis, M. and S. Boissinot. 2013. Lizards and LINEs: Selection and demography affect the fate of L1 Retrotransposons in the genome of the green anole (*Anolis carolinensis*). *Genome Biology and Evolution* 5:2019-2019.
van de Lagemaat, L. N., J. R. Landry, D. L. Mager, and P. Medstrand. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *TRENDS in Genetics* 19:530-536.
Vandewege, M. W., R. N. Platt, D. A. Ray, and F. G. Hoffmann. 2016. Transposable element targeting by piRNAs in Laurasiatherians with distinct transposable element histories. *Genome Biology and Evolution* 8:1327-1337.
Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. 2001. The sequence of the human genome. *Science* 291:1304-1351.
Vicoso, B., J. J. Emerson, Y. Zektser, S. Mahajan, and D. Bachtrog. 2013. Comparative sex chromosome genomics in snakes: Differentiation, evolutionary strata, and lack of global dosage compensation. *Plos Biology* 11.
Vogt, J., K. Bengesser, K. B. Claes, K. Wimmer, V. F. Mautner, R. van Minkelen, E. Legius, H. Brems, M. Upadhyaya, J. Hogel, et al. 2014. SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. *Genome Biol* 15:R80.
Vonk, F. J., N. R. Casewell, C. V. Henkel, A. M. Heimberg, H. J. Jansen, R. J. R. McCleary, H. M. E. Kerkkamp, R. A. Vos, I. Guerreiro, J. J. Calvete, et al. 2013. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *PNAS* 110:20651-20656.
Voss, S. R., D. K. Kump, S. Putta, N. Pauly, A. Reynolds, R. J. Henry, S. Basa, J. A. Walker, and J. J. Smith. 2011. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Research* 21:1306-1312.
Walsh, A. M., R. D. Kortschak, M. G. Gardner, T. Bertozzi, and D. L. Adelson. 2013. Widespread horizontal transfer of retrotransposons. *PNAS* 110:1012-1016.
Warren, W. C., D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Kunstner, S. Searle, S. White, A. J. Vilella, S. Fairley, et al. 2010. The genome of a songbird. *Nature* 464:757-762.

Watanabe, T., Y. Totoki, A. Toyoda, M. Kaneda, S. Kuramochi-Miyagawa, Y. Obata, H. Chiba, Y. Kohara, T. Kono, T. Nakano, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453:539-U539.

Weber, C. C., B. Boussau, J. Romiguier, E. D. Jarvis, and H. Ellegren. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biology* 15.

Weick, E. M. and E. A. Miska. 2014. piRNAs: from biogenesis to function. *Development* 141:3458-3471.

Weirauch, M. T., A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158:1431-1443.

Weissman, I. L. and F. H. Gage. 2016. A mechanism for somatic brain mosaicism. *Cell* 164:593-595.

Wylie, A., A. E. Jones, and J. M. Abrams. 2016. p53 in the game of transposons. *Bioessays* 38:1111-1116.

Xiong, Z. J., F. Li, Q. Y. Li, L. Zhou, T. Gamble, J. Zheng, L. Kui, C. Li, S. B. Li, H. M. Yang, et al. 2016. Draft genome of the leopard gecko, *Eublepharis macularius*. *Gigascience* 5.

Xue, A. T., R. P. Ruggiero, M. J. Hickerson, and S. Boissinot. 2018. Differential effect of selection against LINE retrotransposons among Vertebrates inferred from whole-genome data and demographic modeling. *Genome Biology and Evolution*.

Xue, W., J. T. Li, Y. P. Zhu, G. Y. Hou, X. F. Kong, Y. Y. Kuang, and X. W. Sun. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* 14.

Yin, W., Z. J. Wang, Q. Y. Li, J. M. Lian, Y. Zhou, B. Z. Lu, L. J. Jin, P. X. Qiu, P. Zhang, W. B. Zhu, et al. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nat Commun* 7:13107.

Zachos, J., M. Pagani, L. Sloan, E. Thomas, and K. Billups. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292:686-693.

Zeng, L., S. M. Pederson, R. D. Kortschak, and D. L. Adelson. 2018. Transposable elements and gene expression during the evolution of amniotes. *Mob DNA* 9:17.

Zhang, B., Y. S. Mao, S. D. Diermeier, I. V. Novikova, E. P. Nawrocki, T. A. Jones, Z. Lazar, C. S. Tung, W. Luo, S. R. Eddy, et al. 2017. Identification and characterization of a class of MALAT1-like genomic loci. *Cell Reports* 19:1723-1738.

Zhang, G. J., C. Li, Q. Y. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346:1311-1320.

Zheng, Y. and J. J. Wiens. 2016. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. *Molecular phylogenetics and evolution* 94:537-547.

Zhou, X., Z. Zuo, F. Zhou, W. Zhao, Y. Sakaguchi, T. Suzuki, T. Suzuki, H. Cheng, and R. Zhou. 2010. Profiling sex-specific piRNAs in zebrafish. *Genetics* 186:1175-1185.