The Impact of Transposable Elements on Amniote Evolution

THE UNIVERSITY of ADELAIDE

By

LU ZENG

Department of Genetics and Evolution

School of Biological Sciences

A thesis submitted for the degree of

Doctor of Philosophy

February 2018

Abstract

Transposable elements (TEs) are mobile DNA sequences, often called "jumping genes" because of their ability to replicate to new genomic locations. As a result, TEs make up a significant proportion of the eukaryotic genomes we see today. Growing evidence suggests that TEs are catalysts of genomic change. TE insertions into regulatory regions can lead to new genes, or they can disrupt host sequences and serve as substrates for homologous recombination, generating DNA rearrangements. At the RNA level, TEs can carry transcription factor binding sites, causing alternative splicing and thus impacting gene expression. Some TEs are even capable of jumping between different genomes, using viruses or parasitic insects as transfer vectors. Originally viewed as "junk" DNA, TEs are now recognised as powerful drivers of genome evolution.

However, there are numerous computational challenges to accurately detecting and characterising TEs in genomic data. Many existing tools and pipelines are designed to explicitly remove repetitive, non-unique sequences. Likewise, TE annotation software relies heavily on the use of query sequences and reference databases. This restricts the ability to find new types of TEs (or mutated forms of known TEs), mainly suited to the analysis of model organisms such as human and mouse.

In this thesis, I describe an *ab initio* pipeline for identifying species-specific repeats and segmental duplications with high sensitivity and accuracy. I consider a repeat in the truest sense of the word: any sequence that appears more than once in the genome. Using eight representative species from each branch of amniote evolution, I use this novel method to portray the remarkable diversity of TEs across species and trace different repeats to their families and consensus sequences.

Reptiles are particularly renowned for their unusual TE dynamics. In Chapter 3, I investigate TE evolution in the tuatara genome: a New Zealand reptile. The tuatara is the only surviving member of its order, which flourished around 200 million years ago. Its most recent common ancestor with any other extant group is the lizards and snakes. The tuatara is therefore of great interest to evolutionary geneticists.

In most reptiles, CR1 repeats make up the dominant TE class. My results indicate that the tuatara genome is distinct from other reptiles because the two most dominant TE families are L2 and MIR elements. Furthermore, I describe a likely transfer of L2 elements between

tuatara and monotremes (platypus and echidna), potentially explaining the predominance of L2s in the tuatara genome.

In Chapter 4, I extend my TE analysis to consider gene expression in six species. Due to the prevalence of TEs in the genome, I used a bootstrap approach to minimize the co-occurrence of multiple TE types in one gene. My results show that species-specific associations of TEs with gene expression support a role for TEs in speciation/response to selection by species.

Altogether, this thesis presents novel and *ab initio* approaches for identifying and annotating repetitive elements. By characterizing millions of repeats across different amniote species, and investigating their association with gene expression, I provide evidence for their impact and importance in amniote evolution.

Dedication and Acknowledgements

I can still remember on the 10th of July 2014 when I started my PhD in Adelaide. The first thing I saw when I stepped into the lab was a desk full of balloons and a welcome note. From the start, I knew I would have a wonderful time in the Adelson Lab.

To begin I would like to thank my wonderful supervisors David Adelson and Dan Kortschak. I cannot find the proper words or sentences to express my full appreciation to each of you, though I will try. Dave, you are one of the people I respect most in my life. You are always prepared to help me, and give me your wisdom and your suggestions. Despite your very busy schedule, you still spent so much time helping me throughout my PhD. I especially appreciate our special lab representative - Rory, who is one of the best things that to have in a lab, a lab (Labrador) dog! Dan, I still remember the first thing that you talked to me about when I just started my research life. You said you are glad to help me and that you will be my 'black duck'. To this day I still do not understand what this phrase means. You helped me to correct my English pronunciation, and told me I should not care too much about what other people say. Dave and Dan, I am grateful for your advice on all matters, both professional and personal. Terry, your unwavering enthusiasm and boundless knowledge has been a much needed source of comfort during my times of doubt. I especially want to thank you for taking me on the private lizard exhibition. I can finally show people a picture of the species I studied, it was amazing to hold a live bearded dragon for the first time.

To my wonderful lab mates, you have all been a great group to work with and I have been glad to know you all over the years. Atma, you made one of my biggest dreams come true - having a sibling (especially a sister). Thank you for always cheering me up, sharing your wisdom and experience with me, and lighting my road when it was dark. You will always be a role model to me, I hope all the best things in life can happen to you. Enjoy your upcoming life in America! Reuben, although most of the time I called you 'loobin' or 'boobin', sorry for all the bad jokes made at your expense. Thank you for all the fun you brought for our lab. Zhipeng, your wisdom and advice always came at the most crucial times. James and Brittany, thank you for teaching me some traditional Australian phrases, which I probably should not have learnt. You both taught me to be positive when things are difficult. I especially wanted to say thank you to James for always cheering me up when I am in doubt. I wish you a successful PhD life. Catisha, thank you for all the writing revisions and tea provided. Urwah, thank you for offering me gym classes, they really

helped to dissolve my stress during the last stage of my PhD.

Thank you Janelle Palmer from the Adelaide Graduate Centre, without your help I could not have even started my PhD. thank you to Matt Westlake for providing wonderful computational support with Phoenix. Without your immediate response and quick ability to solve problems, I would have taken far longer in completing my thesis.

Last and by no means least, I would like to thank my parents. Although they worry and do not always agree with the decisions that I made, they stand by my side whatever happens. They are open-minded and respect the life I chose, and make every effort to support me. Thank you for all the love you give to me.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: 27/04/18

Table of Contents

Chapter 1

Introduction

> *"Good science is done by being curious in general, by asking questions all around, by acknowledging the likelihood of being wrong and taking this in good humour for granted, by having a deep fondness for nature, and by being made jumpy and nervous by ignorance."*
>
> — Lewis Thomas

Historically, scientists dismissed transposons as useless or "junk" DNA due to their highly repetitive and non-coding nature. Only recently have scientists begun to entertain the possibility that this so-called "junk" DNA might not be junk after all. In fact, recent studies have found that transposable elements can act as a key driving force in genome evolution. They are a rich source of innovation in genes, regulatory elements and genome structure. Determining how much these transposable elements have altered the regulation of gene expression during evolution can help us predict how likely these elements are to affect future generations.

The impact of transposable elements on gene expression across amniotes

Lu Zeng[1]

[1]School of Biological Sciences, The University of Adelaide, SA 5005, Australia

Abstract

Transposable elements are discrete segments of DNA that can move within genomes. In rodents and primates, approximately 38% to 45% of the genome is made up of transposable elements. Nearly half of the human genome has been identified as comprised of transposable elements. Advances in sequencing technologies have begun to reveal the substantial contribution of these elements to gene expression and genome evolution. Because of their more distant evolutionary relationship to eutheria, monotremes, marsupials, archosaurs and lepidosaurs provide a useful comparison to mammalian gene expression and genome evolution. In this thesis, I describe the implementation of a series of methods and programs for analysing the association between transposable elements and gene expression evolution. I also describe the annotation of transposable elements in amniotes, and analyse the association between transposable elements and gene expression in multiple species, to examine how transposable elements might have contributed to the processes of genome evolution.

Introduction

Definition and classification of repetitive elements

Repetitive elements are segments of nucleic acid sequence that occur in multiple copies throughout a genome. There are three major categories of repetitive elements: tandem repeats

(satellite DNA, minisatellite and microsatellite), segmental duplications, and interspersed repeats (transposon) (Figure 1).

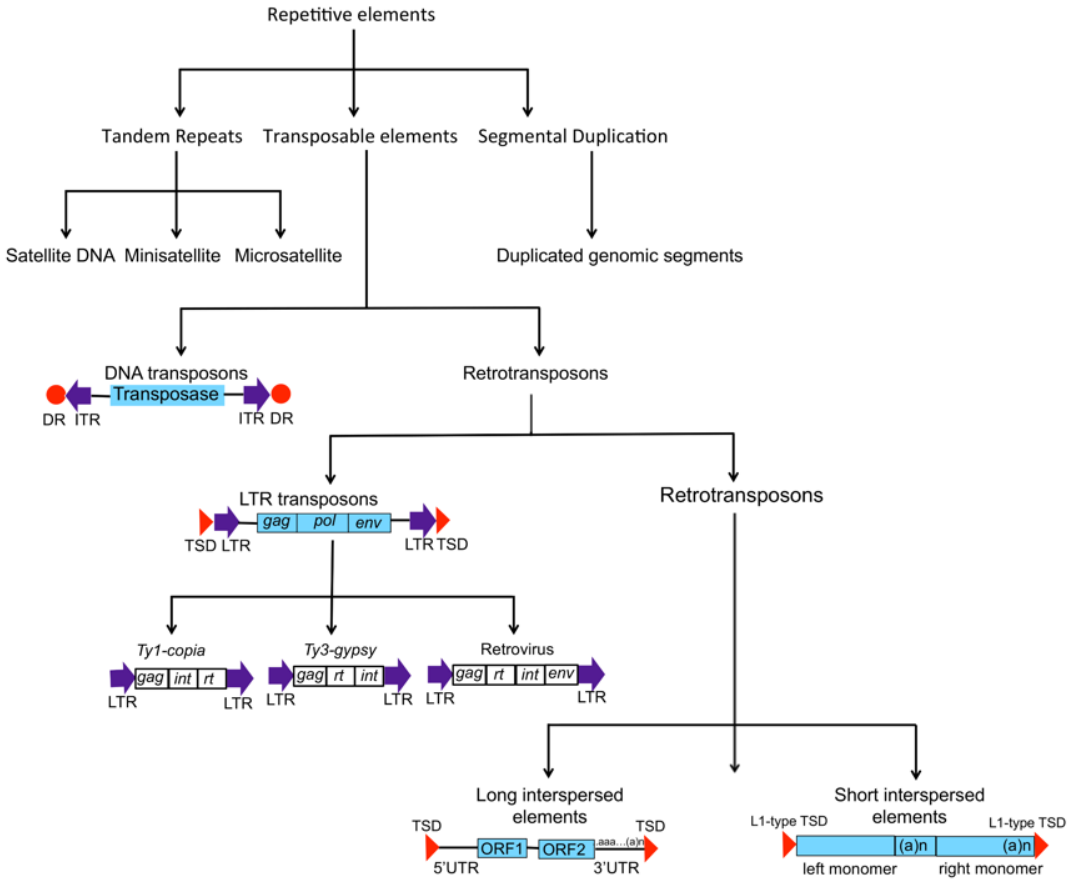


Figure 1: **Repetitive elements classification and structure.** Shows the classification of different repetitive elements, and the structure of main transposable elements.

The majority of repetitive elements in the human genome are derived from transposable elements (TEs) [1] [2] that can move within the genome, potentially giving rise to mutations or altering genome size and structure [3].

Typical eukaryotic genomes can contain millions of copies of transposable elements (TEs) and other repetitive elements. TEs fall into two major classes: Class I, those moving/replicating via

a copy-and-paste mechanism using an RNA intermediate (retrotransposons), and Class II, those moving via direct cut-and-paste of their DNA sequences (DNA transposons). Figure 2 shows the mechanisms of TE replication.

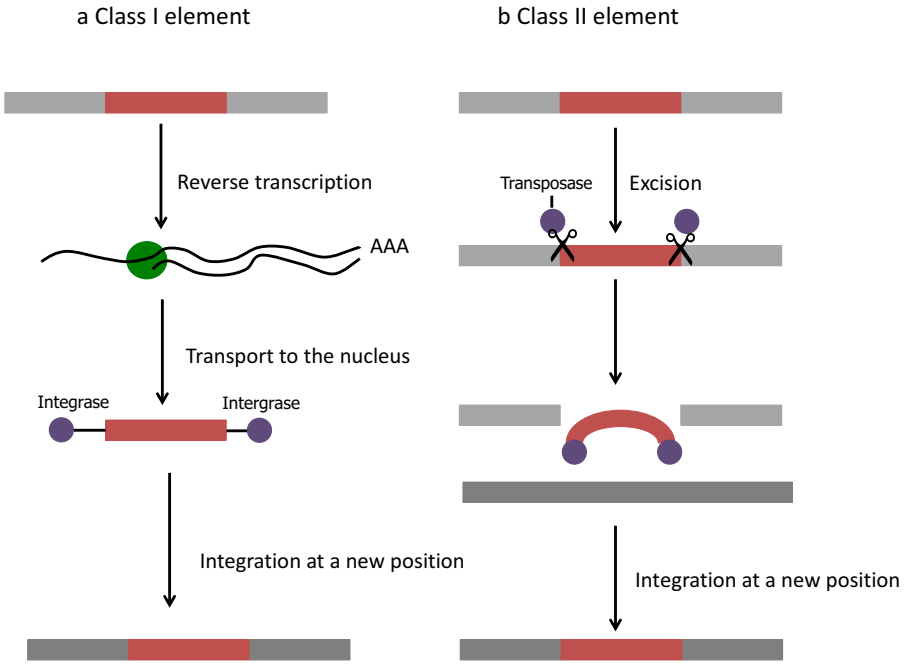


Figure 2: **Mechanisms of replication/transposition**. This figure shows a diagrammatic representation of Class I (retrotransposons) and Class II (DNA transposons) transposable element replication mechanisms.

Retrotransposons can be subdivided into two groups: those with long terminal repeats (LTRs), and those without LTRs (non-LTRs). Non-LTR retrotransposons include two sub-types: autonomous Long INterspersed Elements (LINEs), and non-autonomous Short INterspersed Elements (SINEs) which are dependent on autonomous elements for their replication, both of which are widespread in eukaryotic genomes [1].

A common LTR retrotransposon typically encodes two polyproteins, termed Gag and Pol [4]. The capside protein (Gag) usually contains matrix, capsid, and nucleocapsid domains; Pol consists of aspartic proteinase (AP), reverse transcriptase (RT), ribonuclease (RN), and integrase (INT) domains, the latter three (RT, RN, INT) are responsible for retrotranscribing cDNA from

RNA intermediates and inserting it into the host genome. Endogenous retroviruses constitute a specific class of LTR REs that sometimes additionally contain an open reading frame (ORF) for an envelope protein (Env), which enables ERVs to move from one cell to another. In contrast, all other LTR retrotransposons either completely lack or contain a vestigial remnant of an Env gene and can only reinsert into their own host genome [5] [6] [7]. There are, however, ERVs that have secondarily lost their Env gene and thus their infectious ability. Such ERVs are limited to retrotransposing rather than infecting other cells as typical retroviruses do [8].

LINEs are genetic elements that contribute significantly to eukaryotic genomes [9]. Full-length LINEs are around 6kb long and usually consist of a 5′UTR containing an internal RNA polymerase II promoter, which enables them to be transcribed [10]. Due to their replication method, copying themselves rather than cutting-and-pasting as with transposons, LINEs comprise 17% of the human genome [11]

SINEs are short repeated DNA sequences, and their full length is usually less than 500 base pairs [9]. SINEs are transcribed by RNA polymerase III, which also transcribes other non-coding RNAs such as tRNA, 5S ribosomal RNA, and other small RNAs. SINEs make up about 11% of the human genome [11].

The content of repetitive elements in different species

Birds and mammals are at the extremes among amniotes in genome size and repetitive element content. Avian genomes are relatively small with generally $<$10% TEs [12] [13] [14], whereas mammalian genomes are about three times as large and often contain a TE density of 50% [15] [16] [17]. However, a recent study found that the axolotl genome has an even greater TE density at 65.6% for a total of 18.6 Gb of repetitive sequence [18]. With the exception of the anole lizard that exhibits a highly diverse landscape of TE activity [19] [20], previously analysed amniote genomes are largely dominated by activity of a single LINE superfamily, namely L1 in placental and marsupial mammals [15] [16], and L2 in monotreme mammals [17], which are the most

ancient families of TEs in the human genome [15]. In contrast, CR1 LINEs are the "major genome component" [21] in birds [12] [13] [14], crocodilians [22], snakes [23], and turtles [24]. Given the observation that ancient CR1 lineages are also present in mammals, it has been suggested that the dominance of CR1 in sauropsids resembles the genome organization of the amniote ancestor [25] [26]. These findings support TEs as important elements for the study of amniote evolution.

Table 1 shows an example of the specific composition of transposable elements in human, bovine, and mouse genomes. The high proportion of transposable elements in different organisms genomes indicates that their existence may play critical roles in gene regulation and genome evolution.

Table 1: The abundance of different classes of transposable elements in the bovine, human and mouse genomes (D L. Adelson, personal communication [27])

Group	Number	Total bp	Percentage coverage of genome		
			Bovine	Human	Mouse
Non-LTR retrotransposons (LINE)					
L1	616,259	328,664,804	11.26352	17.07	19.14
RTE(BovB)	376,067	313,409,818	10.74072	NA	0.02
L2	132,485	34,553,185	1.18416	3.07	0.37
CR1	14,524	3,083,954	0.10569	0.27	0.06
Total	**1,139,335**	**679,711,761**	**23.29409**	**20.40**	**19.59**
SINEs					
BOV-A2	377,697	68,880,046	2.360556	NA	NA
Bov-tA	1,461,800	225,579,571	7.730733	NA	NA
ART2A	348,768	121,997,595	4.18092	NA	NA
tRNA	388,920	57,981,206	1.98705	NA	0.00
MIR	301,335	40,569,445	1.39034	2.42	0.55
Other	4,322	432,334	0.01482	10.68	6.78
Total	**2,882,842**	**515,440,197**	**17.66441**	**13.11**	**7.34**
ERVs	277,632	93,363,384	3.19961	8.56	9.84
DNA transposons	244,174	57,157,641	1.95882	3.00	0.89
LTR Other	34,352	12,395,410	0.42480	0.00	0.01
Interspersed repeat total	**4,578,335**	**1,358,068,393**	**46.54174**	**45.089**	**37.65**
SSR Total	**5,653,575**	**66,275,552**	**2.27130**	**0.78**	**4.16**

Functions of transposable elements

Not only do transposable elements contribute to genome content, but they have also been found to significantly affect genome structure, genome evolution, and gene expression. These ideas will be discussed below.

Retrotransposons can impact on genome structure

Past studies have shown that retrotransposons can impact on human genome structure through various mechanisms, and can dramatically affect genome evolution at the DNA level. For example, retrotransposition of L1 and *Alu* might cause insertional mutagenesis if their target is a genic region. In addition, retrotransposons (red box) can influence genome structure in the following ways (Figure 3): creating and repairing DNA double-strand breaks, promoting gene conversion, leading to insertion-mediated deletions, and causing ectopic recombination and transduction.

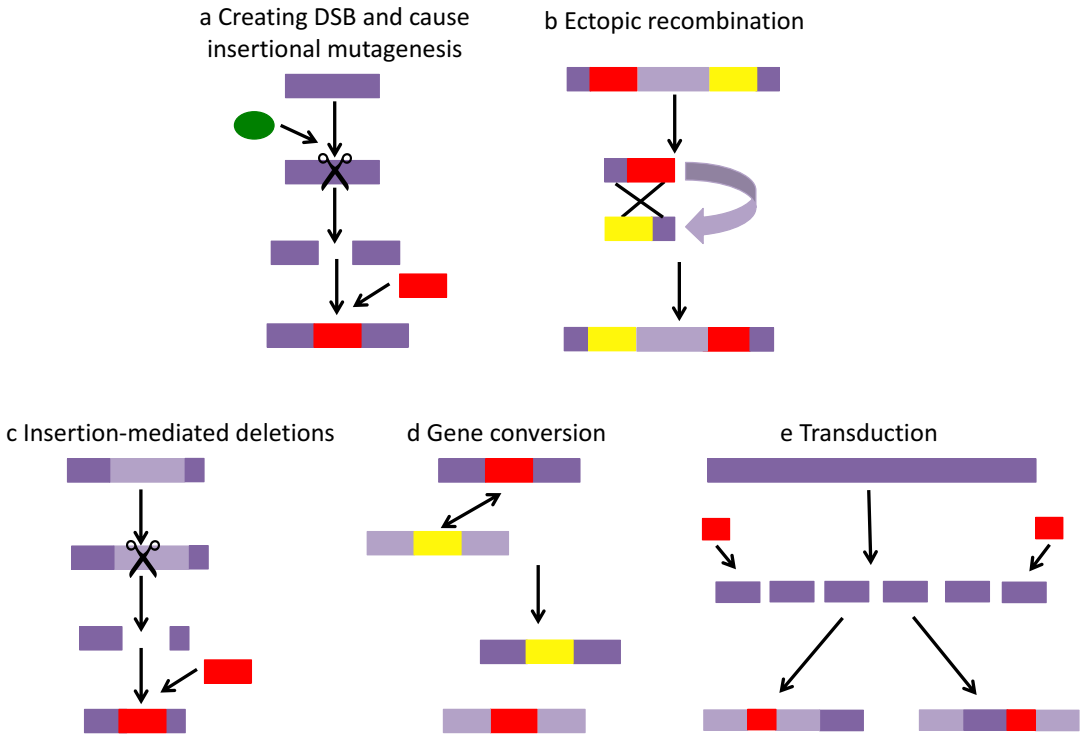


Figure 3: **Impact of retrotransposons on human genome structure.**

The exaptation of transposable elements

In 1992, Brosius and Gould [28] suggested the term "exaptation" for the phenomenon of "junk" DNA sequences (such as TEs) acquiring a novel function in the genome [28]. Although repetitive elements have been regarded as parasitic DNA elements with no intrinsic functions for the host [29], early studies have found that TE insertions into specific genomes have contributed new genes, coding sequences, and regulatory regions [30]. Moreover, recent research supports the view that changes in gene regulation play a significant role in morphological evolution [31]. Therefore, considering the huge TE content of many vertebrate genomes, the exaptation of TEs into new promoters, enhancers, and other regulatory elements is likely to drive the evolution of transcriptional networks.

Despite TEs commonly being regarded as "junk" DNA sequences, a recent study comparing 29

mammalian genomes found that all transposable elements, including LINEs, SINEs, LTRs, and DNA transposons in the human genome have contributed to regulatory innovation on the human lineage [32]. Among these mammalian genomes, at least 11% of gene regulatory sequences in the human genome showing cross-species conservation were co-opted from repetitive elements. They also found that certain repetitive element sequence insertions in these genome regions are more likely to be maintained by purifying selection.

For example, experiments using transgenic mice show that MIR elements were exapted into non-classical enhancers with the ability to influence the expression of nearby genes, and the dispersion of the MER20 element in placental mammals has helped create completely new cell types during development [33]. L1 elements were also found to have provided more putative regulatory sequences than other transposable elements [32]. As they contribute to the largest number of conserved non-exonic elements (CNEE) to the human genome, where CNEE are thought to act as gene regulatory regions that control the special and temporal expression of genes during development [34] [35] [36]. In addition, LINEs and LTRs may have an intrinsic ability to recruit RNA polymerase II, and thus have great potential to be exapted as promoters [30]. However, most TE-derived sequences in vertebrate genomes are still considered to have no function with respect to gene regulation.

Retrotransposons can impact on genome evolution

Continued TE activity and accumulation in the genome over millions of years, non-LTR retrotransposons are expected to have contributed significantly to the evolution of genomes.

Alu elements are ancestrally derived from the 7SL RNA gene [37]. The evolution of *Alu* elements is dominated by the accumulation of new *Alu* inserts, which are rarely removed by non-specific deletion processes. For example, comparisons between the chimpanzee and human genomes showed that about 2,400 and 5,000 lineage-specific insertions have been fixed over 6 million years since divergence [38] [39]. Moreover, *Alu* elements are enriched in gene-rich

regions, while L1 elements are enriched in the gene-poor regions [1] [40]. However, the younger *Alu* and L1 elements do not show significant disparity in their genomic distributions, making it most likely that the differences in location are the result of losses of L1 and *Alu* elements in different genomic regions. This may explain why larger L1 elements might be subject to more negative selection when located in or near genes, while *Alu* are more stably maintained within genes.

Although highly active L1s are rare, as most 5 ends of the L1 copies are truncated, these L1s account for the majority of retrotransposition insertions in humans. Several thousand full-length L1 elements exist in the human reference genome [41], and 80-100 copies are potentially still active. But only a small number of these L1 elements are highly active in the genome and are known as "hot" L1s [42].

Not only do repetitive elements impact genome evolution within species, but they can also influence genome evolution through horizontal transfer (HT) between different organisms. Horizontal transfer is defined as the non-parental transmission of genetic material between individuals, and was thought to rarely occur in multicellular eukaryotes [43]. In most cases, TEs don't have virus-like envelope proteins, thus they need a vector to facilitate HT. LTR retrotransposons and DNA transposons, unlike retrotransposons, have more stable double-stranded DNA intermediates, which are thought to have a higher likelihood of horizontal transfer between organisms than non-LTR retrotransposons [44] [45]. However, it has been shown that the HT of BovB elements (LINE retrotransposon) is widespread, providing evidence for HT of genetic material that has transformed vertebrate genomes [46].

Retrotransposons can influence gene expression

As described above, retrotransposons have dramatically affected genome evolution at the DNA level. Accumulating evidence also shows that retrotransposons have substantially shaped human genome evolution at the RNA level. According to recent studies, the insertion of

retrotransposons can disrupt gene expression [47], causing numerous diseases and syndromes [48]. In addition, expression and export of retrotransposon transcripts in exosomes has been shown to affect gene expression in target cells [49].

Retrotransposons tend to regulate gene expression through several mechanisms, for example, retrotransposons can carry transcription-factor binding sites, which can bind to transcription factors to upregulate or downregulate the expression of neighbouring genes. Retrotransposons can also be recruited as a coding sequence and be integrated into a gene by alternative splicing (Figure 4).

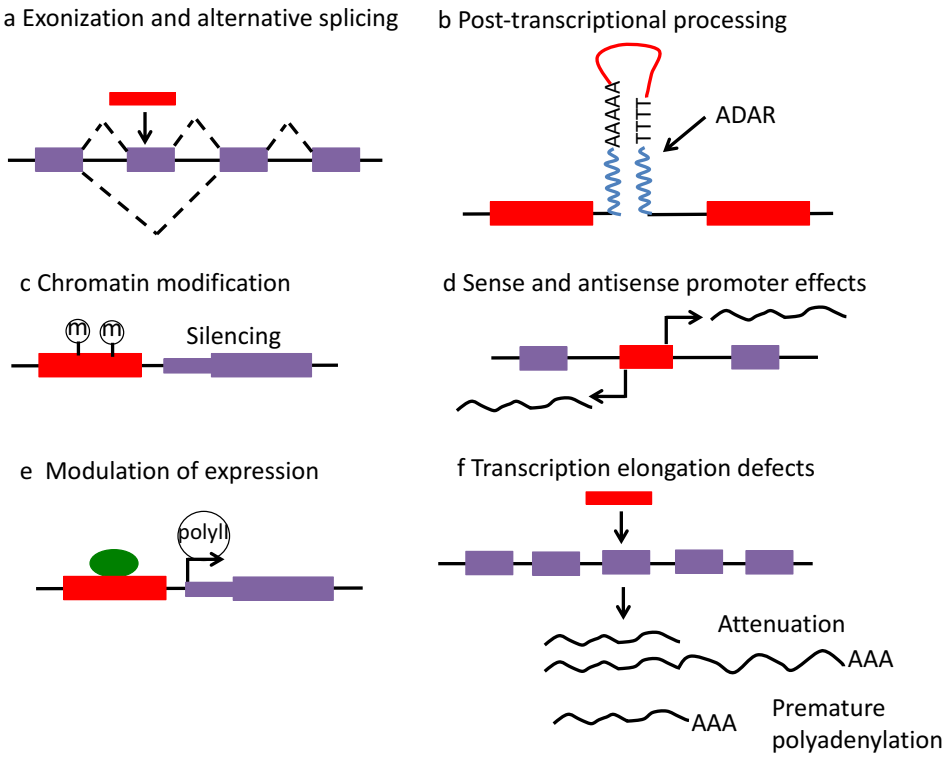


Figure 4: **Impact of retrotransposons on human gene expression.**

Research into the impact of repetitive elements on gene expression has found that genes containing LINE elements tend to have reduced gene expression, as the insertion of L1 elements into an intron can attenuate the target genes expression by premature truncation of RNAs (for

antisense insertions) or transcriptional elongation defects (for sense insertions) [47]. Moreover, L1 elements with a downstream antisense promoter (ASP) have been shown to act as alternative promoters for over 40 human protein-coding genes [50] [51].

SINEs also have an impact, as the insertion of *Alu* elements in or near a gene has the potential to influence the expression of that gene with different ways. For instance, *Alu* elements are relatively enriched in CpG Islands, which are highly correlated with chromosome demethylation [52]. In addition, the methylation of *Alu* elements varies in different tissues and appears to decrease in many tumour tissues [53]. Moreover, dozens of different transcription-factor binding-motifs have been predicted within *Alu* elements [54], and some of these binding sites are specific to certain *Alu* subfamilies, some are also enhanced by changes that occur in *Alu* elements post-insertion. *Alu* elements have also been found to contribute to an array of post-transcriptional processes, including providing polyadenylation sites [55], sites for alternative splicing [56] and sites for RNA editing that can influence the fate of RNA [57].

Additionally, in some mammals, TE insertion into genomes can drive expression in novel tissues, for example, LTR retroviruses are known to be most active in the placenta [58]. Furthermore, TEs have been shown to influence gene expression through non-coding RNAs, resulting in the reduction or silencing of gene expression. For instance, the expression of lincRNA was strongly correlated with HERVH transcriptional regulatory signals [59].

In addition, the non-random genomic distribution of human TEs, with their regulatory potential, may suggest the possibility that the TE environment of genes might affect the way they are expressed. According to previous research, conserved intronic sequences have been found to affect TE integration frequency over time. For example, *Alu* elements are under-represented in highly expressed introns, which has been explained because TEs in these intronic regions conflict with the need to reduce the energy costs of transcription [60]. However, TEs are generally abundant within intronic regions as well as in intergenic regions, indicating that differences from TE effects may depend on their location.

Evolution and gene expression levels

Gene expression changes are considered to underlie many of the phenotypic differences between species. A recent study based on large-scale analyses of gene expression found that the rate of gene expression evolution (the change of gene expression levels) varies in organs, lineages and chromosomes, due to evolutionary selective pressure [61]. In this paper, Brawand *et al.* (2011) showed that transcriptomes from nervous system tissues changed slowly, which was in contrast to testis. The sex chromosomes, especially the X chromosome, has evolved at a faster rate than autosomes. Necsulea *et al.* (2014) showed [62] that different mammalian lineages have different rates of expression evolution. For example, gene expression levels have evolved faster in primates than in rodents. However, this primate-specific acceleration of transcriptome evolution cannot be explained by mutation rate differences, given that rodents have much higher mutation rates than primates owing to their short generation time [63]. As a consequence of this elevated mutation rate, synonymous substitution rates were higher in the mouse lineage than in primates. Thus in the absence of natural selection, elevated mutation rates might increase the divergence of regulatory sequences and thus accelerate expression evolution in rodents.

The phylogeny of amniotes

Amniotes are a clade of tetrapod vertebrates comprised of the reptiles, birds, monotremes, marsupials and eutheria. The evolutionary tree below (Figure 5) shows the emergence of traits in the mammalian lineages. First, the amniotes split into sauropsids (leading to archosaurs and lepidosaurs) and synapsids (leading to mammal-like reptiles) about 315 Mya. Then these early mammals developed hair, homeothermy and lactation (red lines), as shown in Figure 5. Monotremes diverged from the prototherian mammal lineage around 166Mya [64], while the therian mammals split into marsupials and eutherians approximately 148 Mya [64] (based on placental cleavage). Therefore, the evolutionary specialisations of amniotes may help us better understand the biological processes of evolution.

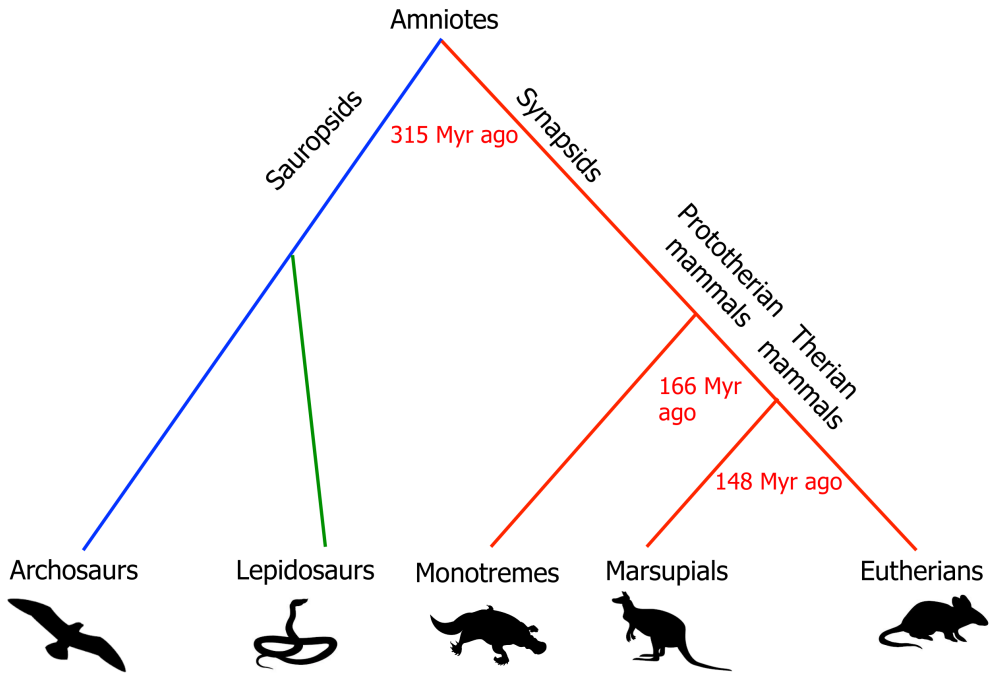


Figure 5: **The phylogeny tree of the amniotes evolution.**

Problems with TE annotation and identification

Repeats are computationally difficult to detect and annotate *ab initio* because of their abundance, varied features/sequence signatures, many length variants (truncated versions) and clade specificity.

Although repeats are abundant in the majority of genomes, Saha *et al.* [65] pointed out that the algorithms and computational tools for identifying and studying repeat sequences are relatively primitive compared to those being utilized to study genes. Furthermore, the majority of these tools cannot process a complete chromosome, let alone a whole genome; processing times may be long enough to be impractical; and sometimes there are difficulties in installing and using some of these tools. Therefore, annotating newly sequenced genomes requires tools to discover repeats *ab initio*.

Research questions:

In order to further characterize how transposable elements affect the genome and may affect phenotype, my research focuses on addressing the following questions:

1) What is the content and class composition of repetitive elements in amniotes (e.g. chicken, anole lizard, bearded dragon, platypus, echidna, opossum and human)?

2) Can we identify any newly developed classes of repetitive elements in these species?

3) What is the association between repetitive elements and gene expression evolution?

4) What classes of repetitive elements may significantly contribute to gene expression evolution?

Aims/objectives:

1) Update the repetitive element annotation across amniotes.

2) Identify new classes of repetitive elements from those species.

3) Analyse the association between repetitive elements and gene expression evolution from these species.

Conclusion

Previous research has shown that TEs play a major role in shaping vertebrate genomes, and that they can influence genome structure and evolution, providing novel promoters, and splice sites, and re-wiring developmental regulatory and transcription networks. A number of existing studies focused on TEs have been carried out in a restricted subset of closely related species; human, other primates and mouse, and highlight the importance of genome structure as a determinant of genomic regulation. The association between TEs and genome structure and gene expression in more widely divergent species, such as platypus, bearded dragon and

opossum is an area that requires additional investigation. Owing to the unique characteristics of amniotes with respect to genome evolution and gene expression, exploring the distribution of TEs in these species may help us better understand the mechanisms of genome evolution.

References

[1] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[2] Arian FA Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development*, 9(6):657–663, 1999.

[3] Barbara McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, 1950.

[4] Nicholas J Holton, Timothy JD Goodwin, Margaret I Butler, and Russell TM Poulter. An active retrotransposon in candida albicans. *Nucleic acids research*, 29(19):4014–4024, 2001.

[5] Haig H Kazazian. Mobile elements: drivers of genome evolution. *science*, 303(5664):1626–1632, 2004.

[6] Carlos Llorens, Alfonso Muñoz-Pomer, Lucia Bernad, Hector Botella, and Andrés Moya. Network dynamics of eukaryotic ltr retroelements beyond phylogenetic trees. *Biology direct*, 4(1):41, 2009.

[7] Thomas H Eickbush and Varuni K Jamburuthugoda. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus research*, 134(1-2):221–234, 2008.

[8] Gkikas Magiorkinis, Robert J Gifford, Aris Katzourakis, Joris De Ranter, and Robert Belshaw. Env-less endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of Sciences*, 109(19):7385–7390, 2012.

[9] Maxine F Singer. Sines and lines: highly repeated short and long interspersed sequences in mammalian genomes. *Cell*, 28(3):433–434, 1982.

[10] Aurélien J Doucet, Amy E Hulme, Elodie Sahinovic, Deanna A Kulpa, John B Moldovan, Huira C Kopera, Jyoti N Athanikar, Manel Hasnaoui, Alain Bucheton, John V Moran, et al. Characterization of line-1 ribonucleoprotein particles. *PLoS genetics*, 6(10):e1001150, 2010.

[11] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, 2009.

[12] International Chicken Genome Sequencing Consortium et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695, 2004.

[13] Wesley C Warren, David F Clayton, Hans Ellegren, Arthur P Arnold, LaDeana W Hillier, Axel Künstner, Steve Searle, Simon White, Albert J Vilella, Susan Fairley, et al. The genome of a songbird. *Nature*, 464(7289):757, 2010.

[14] Guojie Zhang, Cai Li, Qiye Li, Bo Li, Denis M Larkin, Chul Lee, Jay F Storz, Agostinho Antunes, Matthew J Greenwold, Robert W Meredith, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, 346(6215):1311–1320, 2014.

[15] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[16] Tarjei S Mikkelsen, Matthew J Wakefield, Bronwen Aken, Chris T Amemiya, Jean L Chang, Shannon Duke, Manuel Garber, Andrew J Gentles, Leo Goodstadt, Andreas Heger, et al. Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):167, 2007.

[17] Wesley C Warren, LaDeana W Hillier, Jennifer A Marshall Graves, Ewan Birney, Chris P Ponting, Frank Grützner, Katherine Belov, Webb Miller, Laura Clarke, Asif T Chinwalla, et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–183, 2008.

[18] Sergej Nowoshilow, Siegfried Schloissnig, Ji-Feng Fei, Andreas Dahl, Andy WC Pang, Martin Pippel, Sylke Winkler, Alex R Hastie, George Young, Juliana G Roscito, et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature*, 554(7690):50, 2018.

[19] Marc Tollis and Stéphane Boissinot. The transposable element profile of the anolis genome: How a lizard can provide insights into the evolution of vertebrate genome size and structure. *Mobile genetic elements*, 1(2):107–111, 2011.

[20] Jessica Alföldi, Federica Di Palma, Manfred Grabherr, Christina Williams, Lesheng Kong, Evan Mauceli, Pamela Russell, Craig B Lowe, Richard E Glor, Jacob D Jaffe, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, 477(7366):587, 2011.

[21] D Kordis. Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenetic and genome research*, 127(2-4):94–111, 2009.

[22] Richard E Green, Edward L Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W Vandewege, John A St John, Salvador Capella-Gutiérrez, Todd A

Castoe, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215):1254449, 2014.

[23] Todd A Castoe, AP Jason De Koning, Kathryn T Hall, Daren C Card, Drew R Schield, Matthew K Fujita, Robert P Ruggiero, Jack F Degner, Juan M Daza, Wanjun Gu, et al. The burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences*, 110(51):20645–20650, 2013.

[24] H Bradley Shaffer, Patrick Minx, Daniel E Warren, Andrew M Shedlock, Robert C Thomson, Nicole Valenzuela, John Abramyan, Chris T Amemiya, Daleen Badenhorst, Kyle K Biggar, et al. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome biology*, 14(3):R28, 2013.

[25] Andrew M Shedlock, Christopher W Botka, Shaying Zhao, Jyoti Shetty, Tingting Zhang, Jun S Liu, Patrick J Deschavanne, and Scott V Edwards. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences*, 104(8):2767–2772, 2007.

[26] Alexander Suh, Gennady Churakov, Meganathan P Ramakodi, Roy N Platt, Jerzy Jurka, Kenji K Kojima, Juan Caballero, Arian F Smit, Kent A Vliet, Federico G Hoffmann, et al. Multiple lineages of ancient cr1 retroposons shaped the early genome evolution of amniotes. *Genome biology and evolution*, 7(1):205–217, 2014.

[27] David L Adelson, Joy M Raison, and Robert C Edgar. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences*, 106(31):12855–12860, 2009.

[28] Jürgen Brosius and Stephen Jay Gould. On" genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other" junk dna". *Proceedings of the National Academy of Sciences*, 89(22):10706–10710, 1992.

[29] W Ford Doolittle and Carmen Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603, 1980.

[30] Patrik Medstrand, LN Van de Lagemaat, Catherine A Dunn, J-R Landry, Daniel Svenback, and Dixie L Mager. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenetic and genome research*, 110(1-4):342–352, 2005.

[31] Alistair P McGregor, Virginie Orgogozo, Isabelle Delon, Jennifer Zanet, Dayalan G Srinivasan, François Payre, and David L Stern. Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, 448(7153):587–590, 2007.

[32] Craig B Lowe and David Haussler. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS One*, 7(8):e43128, 2012.

[33] Vincent J Lynch, Robert D Leclerc, Gemma May, and Günter P Wagner. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature genetics*, 43(11):1154–1159, 2011.

[34] Marcelo A Nobrega, Ivan Ovcharenko, Veena Afzal, and Edward M Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413–413, 2003.

[35] Adam Woolfe, Martin Goodson, Debbie K Goode, Phil Snell, Gayle K McEwen, Tanya Vavouri, Sarah F Smith, Phil North, Heather Callaway, Krys Kelly, et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology*, 3(1):e7, 2004.

[36] Francis Poulin, Marcelo A Nobrega, Ingrid Plajzer-Frick, Amy Holt, Veena Afzal, Edward M Rubin, and Len A Pennacchio. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics*, 85(6):774–781, 2005.

[37] Jan Ole Kriegs, Gennady Churakov, Jerzy Jurka, Jürgen Brosius, and Jürgen Schmitz. Evolutionary history of 7sl rna-derived sines in supraprimates. *Trends in Genetics*, 23(4):158–161, 2007.

[38] Ryan E Mills, E Andrew Bennett, Rebecca C Iskow, Christopher T Luttig, Circe Tsui, W Stephen Pittard, and Scott E Devine. Recently mobilized transposons in the human and chimpanzee genomes. *The American Journal of human genetics*, 78(4):671–679, 2006.

[39] Dale J Hedges, Pauline A Callinan, Richard Cordaux, Jinchuan Xing, Erin Barnes, and Mark A Batzer. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome research*, 14(6):1068–1075, 2004.

[40] Reuben M Buckley, R Daniel Kortschak, Joy M Raison, and David L Adelson. Similar evolutionary trajectories for retrotransposon accumulation in mammals. *Genome biology and evolution*, 9(9):2336–2353, 2017.

[41] Sanjida H Rangwala, Lili Zhang, and Haig H Kazazian. Many line1 elements contribute to the transcriptome of human somatic cells. *Genome biology*, 10(9):R100, 2009.

[42] Brook Brouha, Joshua Schustak, Richard M Badge, Sheila Lutz-Prigge, Alexander H Farley, John V Moran, and Haig H Kazazian. Hot l1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, 100(9):5280–5285, 2003.

[43] Atma M Ivancevic, Ali M Walsh, R Daniel Kortschak, and David L Adelson. Jumping the fine line between species: horizontal transfer of transposable elements in animals catalyses genome evolution. *Bioessays*, 35(12):1071–1082, 2013.

[44] Sarah Schaack, Clément Gilbert, and Cédric Feschotte. Promiscuous dna: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in ecology & evolution*, 25(9):537–546, 2010.

[45] Joana C Silva, Elgion L Loreto, and Jonathan B Clark. Factors that affect the horizontal transfer of transposable elements. *Current issues in molecular biology*, 6:57–71, 2004.

[46] Ali Morton Walsh, R Daniel Kortschak, Michael G Gardner, Terry Bertozzi, and David L Adelson. Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences*, 110(3):1012–1016, 2013.

[47] Jeffrey S Han, Suzanne T Szak, and Jef D Boeke. Transcriptional disruption by the l1 retrotransposon and implications for mammalian transcriptomes. *Nature*, 429(6989):268–274, 2004.

[48] Mark A Batzer and Prescott L Deininger. Alu repeats and human genomic diversity. *Nature reviews genetics*, 3(5):370–379, 2002.

[49] Leonora Balaj, Ryan Lessard, Lixin Dai, Yoon-Jae Cho, Scott L Pomeroy, Xandra O Breakefield, and Johan Skog. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nature communications*, 2:180, 2011.

[50] Mart Speek. Antisense promoter of human l1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and cellular biology*, 21(6):1973–1985, 2001.

[51] Kert Mätlik, Kaja Redik, and Mart Speek. L1 antisense promoter drives tissue-specific transcription of human genes. *BioMed Research International*, 2006, 2006.

[52] Hehuang Xie, Min Wang, Maria de F Bonaldo, Christina Smith, Veena Rajaram, Stewart Goldman, Tadanori Tomita, and Marcelo B Soares. High-throughput sequence-based

epigenomic analysis of alu repeats in human cerebellum. *Nucleic acids research*, 37(13):4331–4340, 2009.

[53] Prescott Deininger. Alu elements: know the sines. *Genome biology*, 12(12):236, 2011.

[54] Paz Polak and Eytan Domany. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC genomics*, 7(1):133, 2006.

[55] Chongjian Chen, Takeshi Ara, and Daniel Gautheret. Using alu elements as polyadenylation sites: A case of retroposon exaptation. *Molecular biology and evolution*, 26(2):327–334, 2008.

[56] Shihao Shen, Lan Lin, James J Cai, Peng Jiang, Elizabeth J Kenkel, Mallory R Stroik, Seiko Sato, Beverly L Davidson, and Yi Xing. Widespread establishment and regulatory impact of alu exons in human genes. *Proceedings of the National Academy of Sciences*, 108(7):2837–2842, 2011.

[57] Ling-Ling Chen, Joshua N DeCerbo, and Gordon G Carmichael. Alu element-mediated gene silencing. *The EMBO journal*, 27(12):1694–1705, 2008.

[58] Ivan Bieche, Anne Laurent, Ingrid Laurendeau, Laurent Duret, Yves Giovangrandi, Jean-Louis Frendo, Martine Olivi, Jean-Luc Fausser, Daniele Evain-Brion, and Michel Vidaud. Placenta-specific insl4 expression is mediated by a human endogenous retrovirus element. *Biology of reproduction*, 68(4):1422–1429, 2003.

[59] David Kelley and John Rinn. Transposable elements reveal a stem cell-specific class of long noncoding rnas. *Genome biology*, 13(11):R107, 2012.

[60] Manuela Sironi, Giorgia Menozzi, Giacomo P Comi, Matteo Cereda, Rachele Cagliani, Nereo Bresolin, and Uberto Pozzoli. Gene function and expression level influence the

insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome biology*, 7(12):R120, 2006.

[61] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.

[62] Anamaria Necsulea and Henrik Kaessmann. Evolutionary dynamics of coding and non-coding transcriptomes. *Nature Reviews Genetics*, 15(11):734–748, 2014.

[63] Wen-Hsiung Li, Darrell L Ellsworth, Julia Krushkal, Benny H-J Chang, and David Hewett-Emmett. Rates of nucleotide substitution in primates and rodents and the generation–time effect hypothesis. *Molecular phylogenetics and evolution*, 5(1):182–187, 1996.

[64] Olaf RP Bininda-Emonds, Marcel Cardillo, Kate E Jones, Ross DE MacPhee, Robin MD Beck, Richard Grenyer, Samantha A Price, Rutger A Vos, John L Gittleman, and Andy Purvis. The delayed rise of present-day mammals. *Nature*, 446(7135):507–512, 2007.

[65] Surya Saha, Susan Bridges, Zenaida V Magbanua, and Daniel G Peterson. Computational approaches and tools used in identification of dispersed repetitive dna sequences. *Tropical Plant Biology*, 1(1):85–96, 2008.

Chapter 2

Superior *ab initio* Identification, Annotation and Characterisation of TEs and Segmental Duplications from Genome Assemblies

> *"Any fool can write code that a computer can understand. Good programmers write code that humans can understand."*
>
> — Martin Fowler

Thousands of genomes have been sequenced thanks to decreased cost and increased speed of DNA sequencing methods. Repeats including transposable elements (TEs) comprise significant portions of eukaryotic genomes. Species-specific TEs in newly sequenced genomes are likely to be unknown. Therefore, annotating newly sequenced genomes requires tools to discover TEs *ab initio*. However, the currently available *ab initio* tools have limitations concerning the size of the input sequence, sensitivities to major types of repeats, and consistency of performance. In this chapter, I will describe an *ab initio* approach used to identify and annotate TEs within multiple species, including both well-annotated and draft genomes.

Statement of Authorship

Title of Paper	Superior *ab initio* Identification, Annotation and Characterisation of TEs and Segmental Duplications from Genome Assemblies.
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Lu Zeng, R. Danial Kortshcak, Joy M. Raison, Terry Bertozzi, David L. Adelson. Superior *ab initio* Identification, Annotation and Characterisation of TEs and Segmental Duplications from Genome Assemblies. Research Article have been accepted by PLOS ONE.

Principal Author

Name of Principal Author (Candidate)	Lu Zeng		
Contribution to the Paper	Processed data, performed analysis, prepared figures and wrote the manuscript		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	20/02/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	R. Danial Kortschak		
Contribution to the Paper	Supervised the development of work, provided the code, assisted with analysis of data and assisted in writing the manuscript.		
Signature		Date	26/2/2018

Name of Co-Author	Joy M. Raison		
Contribution to the Paper	Assisted with development of data analysis techniques and edited manuscript.		
Signature		Date	22/2/18

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Terry Bertozzi		
Contribution to the Paper	Assisted with development of data analysis techniques and edited manuscript.		
Signature		Date	26/2/18

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised the development of work and assisted in analysing the results and writing the manuscript.		
Signature		Date	26/2/18

Superior *ab initio* Identification, Annotation and Characterisation of TEs and Segmental Duplications from Genome Assemblies.

Lu Zeng[1], R. Daniel Kortschak[1], Joy M. Raison[1], Terry Bertozzi[1,2], David L. Adelson[1*]

1 School of Biological Sciences, The University of Adelaide, Adelaide, SA 5005, Australia
2 Evolutionary Biology Unit, South Australian Museum, Adelaide, SA 5005, Australia

* david.adelson@adelaide.edu.au

Abstract

Transposable Elements (TEs) are mobile DNA sequences that make up significant fractions of amniote genomes. However, they are difficult to detect and annotate *ab initio* because of their variable features, lengths and clade-specific variants. We have addressed this problem by refining and developing a Comprehensive *ab initio* Repeat Pipeline (CARP) to identify and cluster TEs and other repetitive sequences in genome assemblies. The pipeline begins with a pairwise alignment using krishna, a custom aligner. Single linkage clustering is then carried out to produce families of repetitive elements. Consensus sequences are then filtered for protein coding genes and then annotated using Repbase and a custom library of retrovirus and reverse transcriptase sequences. This process yields three types of family: fully annotated, partially annotated and unannotated. Fully annotated families reflect recently diverged/young known TEs present in Repbase. The remaining two types of families contain a mixture of novel TEs and segmental duplications. These can be resolved by aligning these consensus sequences back to the genome to assess copy number vs. length distribution. Our pipeline has three significant advantages compared to other methods for *ab initio* repeat identification: 1) we generate not only consensus sequences, but keep the genomic intervals for the original aligned sequences, allowing straightforward analysis of evolutionary dynamics, 2) consensus sequences represent low-divergence, recently/currently active TE families, 3) segmental duplications are annotated as a useful by-product. We have compared our *ab initio* repeat annotations for 6 genome assemblies to other methods and demonstrate that CARP compares favourably with RepeatModeler, the most widely used repeat annotation package.

Introduction

Thousands of genomes have been sequenced thanks to decreased cost and increased speed of DNA sequencing methods. The explosion of genome sequences has expanded our knowledge of repetitive DNA, which is an important component of the genomes of almost all eukaryotes. Repetitive DNA is made up of sequences that have been duplicated. Some repetitive elements are able to replicate to new genomic locations and are referred to as transposable elements (TEs). TEs are known to account for a significant proportion of genome sequences in eukaryotes, varying from a few percent to the majority of the genome. For example, around 50% of the human [1] and 85% of the maize genome are TEs [2]. Therefore, it is important to have an efficient and accurate *ab initio* method of identifying and annotating repeats in newly sequenced genomes.

Repetitive DNA sequences can be divided into three major categories: tandem repeats, segmental duplications and transposable elements. Tandem repeats are repeated DNA sequences that are directly adjacent to each other and account for 3% of the human genome [3].

Segmental duplications (SDs, also termed "low-copy repeats") are DNA sequences of variable sequence length (ranging from 1kb to 400kb) and a high level of sequence identity. SDs are identified from pairwise local alignments generated with BLAST using arbitrary criteria (>90% id, >1000bp length) [4]. Because SD identification is based on local alignments, repeat masked genome sequences are used as input to remove the enormous number of alignments produced by TEs that would overwhelm the SD output. This means that repeat identification and annotation is currently required before SDs can be identified.

Transposable elements are the most prevalent repetitive sequences in eukaryotic genomes, and fall into two major classes: those moving via direct cut and paste of their DNA sequences (DNA transposons) and those moving/replicating via a copy and paste mechanism with an RNA intermediate (retrotransposons). DNA transposons encode a transposase gene that is flanked by two *Terminal Inverted Repeats* (TIRs) [5]. The transposase recognizes these TIRs to excise the transposon DNA, which is then inserted into a new genomic location by cut and paste mobilization [6].

Retrotransposons can be subdivided into two groups: those with long terminal repeats (LTRs), and those without LTRs (non-LTR). Endogenous retroviruses (ERVs) are domesticated remnants of retroviral infection and full-length ERVs encode an array of proteins (*gag*, *pol*, and *env*) flanked by LTRs [7]. The *env* protein allows ERVs to transfer to other organisms by infection [8] and thus ERVs can be acquired from the environment. LTR retrotransposons are the dominant retrotransposons in plants and are less abundant in mammals [9]. Similar to ERVs, LTR retrotransposons contain two long-terminal repeats that flank a 5-7kb long internal protein-coding domain [10] containing two open reading frames (ORFs): *gag* and *pol*. The *gag* ORF encodes the structural protein that makes up a virus-like particle (VLP) [11]. The *pol* ORF encodes an enzyme needed for replication that contains protease (PR), integrase (IN), reverse transcriptase (RT), and RNase H (RH) domains required for reverse transcription and integration. Promoter and transcription termination signals are present in the LTRs that are divided into three functional areas: U3, R and U5. U3 contains the enhancer and promoter sequences that drive viral transcription [11]. However, due to the lack of *env* protein, LTR retrotransposons are not infectious; they are obligate intracellular elements [12].

Non-LTR retrotransposons include two sub-types: autonomous long interspersed elements (LINEs), and non-autonomous short interspersed elements (SINEs), that are dependent on LINEs for their replication [3]. Typical insertions of non-LTR retrotransposons are flanked by target site duplications, which result from micro-homology based repair during the insertion process [13].

LINEs contribute significantly to eukaryotic genomes. Full-length LINEs are around 6kb long and usually contain two ORFs flanked by 5′ and 3′ untranslated regions (UTRs). LINE 5′ UTRs possess an internal RNA polymerase II promoter, which allows them to be transcribed [1]. ORF1 can vary significantly from species to species, and can encode proteins with different characteristics [14]. ORF2 is similar across all LINEs and encodes a protein with endonuclease and reverse-transcriptase activities required for replication [14].

SINEs are much shorter; usually less than 500 base pairs. The 5′ region contains an internal RNA polymerase III promoter and the 3′ end contains an oligo dA-rich tail. *Alu* elements have no ORFs, therefore they have no coding capacity and are non-autonomous TEs. Because they share functional sequences at their 3′ with LINEs,

they borrow the retrotransposition molecular machinery encoded by LINEs that bind to their 3′ end [1].

Repeats are computationally difficult to detect and annotate *ab initio* because of their abundance, varied features/sequence signatures, many length variants (truncated versions) and clade specificity. Many computational tools have been developed to detect TEs, and the most commonly used approaches can be divided into three categories:

1) Library-based methods (e.g. RepeatMasker [15]), that use sequence alignment to search a genome for homologs of known repeats from a database such as Repbase [16], Repbase is a manually curated repeat library of species-specific and pan-species TEs, and cannot be used to identify segmental duplications.

2) Signature-based methods, that rely on the fact that each class of TE has a set of unique sequence features such as target site duplications, a poly-A tail, terminal inverted repeats, etc... These methods search for the sequence signatures of the repeat class of interest (e.g. LTR_STRUC [17]). However, because repeat types are so varied, this method is usually only able to identify specific types of TE.

3) *Ab initio* consensus methods, four examples here are RepeatModeler (`http://www.repeatmasker.org/RepeatModeler/`), REPET [18], Red [19] and PILER [20]. RepeatModeler (RMD) is a *de novo* package that has been widely used for repeat identification and modeling that combines different programs: RepeatMasker, RepeatScout [21], RECON [22] and TRF (Tandem Repeat Finder) [23]. RepeatMasker identifies and masks interspersed repeats using curated libraries of consensus sequences supported by Dfam; Dfam contains entries corresponding to all Repbase TE entries, and each Dfam entry is represented by a profile hidden Markov model. RECON evaluates pair-wise similarities to build repeat consensus sequences. RepeatScout identifies and uses highly over-represented k-mers as seeds that are extended to produce multiple sequence alignments. However, RMD doesn't identify the individual sequences used to derive the consensus sequences; making it impossible to confirm or assess the accuracy of the consensus sequences, or to directly analyse the repeat instances in the genome they are derived from.

Red is an *ab initio* tool for discovering repetitive elements in a genome. Red utilizes a Hidden Markov Model dependent on labeled training data, i.e. it is an instance of supervised learning. Red identifies candidate repetitive regions using adjusted counts of k-mers, score smoothing with a Gaussian mask and the second derivative test to find local maxima [19]. Red can detect both transposons and simple repeats. However, it only generates genome coordinates for repeats, without any annotation. Red output is therefore not useful for analysing repeat content or transposon evolution.

PILER can identify and cluster repeats based on pairwise whole-genome alignments. In contrast to previous methods that attempt to explain all the off-diagonal local alignments or hits, it focuses on identifying subsets of hits that form a pattern characteristic of a given type of repeat. PILER was orignally designed to use PALS to generate pairwise alignment; however, PALS cannot handle concurrent jobs and it was built for a 32-bit processor architecture, which makes it relatively time consuming and seriously limits PILER applicability to small genomes. Although any local aligner can be used to replace PALS, this requires attention to required alignment parameters, and hits need to be converted to PILER-compatible GFF format.

REPET is a package that requires a local aligner, three clustering tools (RECON, PILER and GROUPER [24]) and a knowledge/library based annotation pipeline [25]. REPET produces a very comprehensive output of repeat annotations, but excludes segmental duplications, is complex, requires genome annotation of gene models and is computationally expensive.

In order to address these limitations, we have created a comprehensive *ab initio* repeat pipeline (CARP) for identifying species-specific TE elements with high

sensitivity and accuracy that deals with both TEs and segmental duplications. Our method also provides a full audit trail that links identified repeat sequences (and their genome intervals) to their families and consensus sequences. This permits direct evolutionary analysis of highly similar TE families.

Methods

For a diagrammatic overview of our method for *de novo* discovery and annotation of repetitive elements from genome sequences see (Figure 1).

Datasets

Six genomes were used in this study, 2 reptiles (anolis, *Anolis carolinensis* and bearded dragon, *Pogona vitticeps*), 1 bird (chicken, *Gallus gallus*), 1 monotreme (platypus, *Ornithorhynchus anatinus*, 1 marsupial (opossum, *Monodelphis domestica*) and 1 eutherian mammalian (human, *Homo sapiens*). All genomes are publicly available from the National Center for Biotechnology Information (NCBI). Supporting Information S1 Table lists the systematic name, common name, version, source and submitter for each genome assembly. Supporting Information S2 Table shows the total genome sequence length and scaffold/contig N50 values, giving an approximation of the assembly quality. Supporting Information S3 Table compares the different sequencing technologies and methods.

Comprehensive *ab initio* Repeat Pipeline (CARP)

Repeats were identified using a pipeline comprised of krishna/igor [26], MUSCLE (v3.8.31) [27]. Krishna/igor is an improved version of PALS/PILER implemented in Go (https://golang.org/) that can find dispersed repeat families. A dispersed repeat family has members that are typically separated in the genome, i.e. that are rarely or never found in tandem, and are usually mobile elements such as retrotransposons [20]. Genome sequences were pairwise aligned using krishna (`https://github.com/biogo/examples/krishna`) with default parameters set at 94% sequence identity (-dpid) and a minimum alignment length (-dplen) of 250bp for most cases, except bearded dragon and chicken, which used -dpid 90% and -dplen 200bp. The resulting alignment intervals were then used as input for igor to define families of repeat sequences using the default parameters with single-linkage clustering, which is a clustering method that combines clusters containing elements that are linked by alignment. Igor output was used as input for seqer to generate repeat consensus sequences for each cluster/family based on MUSCLE alignments. Only family members within 95% of the length of the longest family member were aligned, and to avoid consensus sequence expansion due to indels in the global alignment, a maximum of 100 randomly chosen sequences/family were included in the alignment. This process yielded three types of family: fully annotated, partially annotated and unannotated.

Identifiable repeat consensus sequences were annotated by using CENSOR [28] with the Repbase 'Vertebrate' library (downloaded on 1st March, 2016, includes 41,908 sequences). Further annotation of consensus sequences was based on WU-BLAST (v2.0) (`https://www.advbiocomp.com/blast/obsolete/`) [29]/NCBI-BLAST (v2.2.27) [30] alignment against a comprehensive retroviral and retrotransposon protein database assembled from the NCBI [31], and against Swiss-Prot [32] to identify known protein-coding genes from large gene families inappropriately included in the repeat set. Consensus sequences identified as either simple sequence repeats (SSRs) or protein-coding sequences (Evalue < 0.00001) were removed from the consensus set.

After acquiring all the annotated repeat consensus sequences, these annotated consensus sequences were then combined with the Repbase 'Vertebrate' library and CENSOR was used to annotate all repeat intervals in the source genome. Supporting Information S4 Table represents the summary of time consumed for each analysis step. For additional details of time consumption and memory use for each step, see Supporting Information S1 Appendix.

Method Evaluation

RepeatModeler (version 1.0.8) was used to evaluate the performance of CARP by applying it to the same seven genomes with default parameters, with WU-BLAST used as the alignment engine. A combination of the repeat consensus sequences generated by RepeatModeler and Repbase 'Vertebrate' library was also fed into CENSOR to annotate the repeat content for each genome.

Identification Of Novel Repeat Sequences From Tested Genomes

In order to explore the unclassified consensus sequences generated by CARP, we extracted all unclassified repeat sequences from the seven genomes, and the R package ggplot2 was used to visualise their length distribution with respect to copy number.

For high copy number (>2,000 copies), a coverage plot was used to investigate the positional distribution of genomic sequence fragments with respect to the unclassified consensus sequences. BLASTN [30] and CENSOR were further used to characterise the consensus sequences from the coverage peaks of 5 unclassified consensus sequence examples found in the bearded dragon coverage plot.

Human (GRCh37) segmental duplication coordinates were also downloaded (`http://humanparalogy.gs.washington.edu/build37/build37.htm`) and BedTools [33] was used to merge the overlapping intervals from this data. We then used the human unclassified consensus sequences generated from both our *ab initio* method and RepeatModeler as libraries to run CENSOR against the merged segmental duplication data.

Dendrogram construction from platypus nucleotide L2 sequences

Full-length platypus L2 consensus sequences (2-4kb) generated from CARP and RMD were extracted respectively, as well as the genome intervals that linked to the L2 consensus sequences from CARP. We then globally aligned the resulting sequences using MUSCLE (-maxiters 2). Alignments were trimmed with Gblocks [34] to remove large gaps (default parameters, allowed gap postions: with half). FastTree (v2.1.8) [35] was used to infer a maximum likelihood phylogeny from the global alignment, using a generalized time-reversible model (-gtr). Archaeopteryx v0.9901 beta was used to visualise the tree, including 94 genome intervals from CARP, 12 L2 consensus sequences from CARP and 7 consensus sequences from RMD.

Classification of potentially active L2 elements

USEARCH [36] was then used to scan for open reading frames (ORFs) in those full-length L2 consensus sequences that were at least 60% of the expected length ($\geqslant$1.5kb nucleotide sequence for ORF2p, complete with start and stop codons and no inactivating mutations). After translation, ORF2p candidates were checked for similarity to known domains using HMM-HMM comparison [37] against the Pfam28.0 database [38] as of May 2015 (includes 16,230 families). ORF2p containing RT domains

were extracted using the envelope coordinates from the HMMer domain hits table (–domtblout), with a minumum length of 200 amino acids.

Results

Consensus Generation

We identified and annotated repeats from seven genomes using both CARP and RepeatModeler. Because ancient transposable elements are highly diverged and already well described, we have implemented CARP to identify and annotate slightly diverged (recent) repetitive elements. CARP is based on whole-genome pairwise local alignment (default 94% identity), followed by clustering and consensus generation from clusters. This means all consensus sequences generated by CARP can be traced back to their input sequences and the original genomic sequence intervals of the input sequences. This provides an audit trail and the ability to easily carry out evolutionary and phylogenetic analysis of recently diverged, and hence recently active TEs. Because the initial clusters may contain gene families with many paralogs, we cleaned the consensus sequences by aligning them to Swiss-Prot and to a custom database assembled from retroviral and reverse transcriptase (RT) sequences from NCBI. We then removed consensus sequences that align to *bona fide* protein coding genes that do not annotate as retroviral/RT. Cleaned consensus sequences were then annotated with CENSOR using known TE reference sequences from Repbase. This resulted in three types of annotation: 1) well annotated, almost full length alignment to a Repbase reference sequence, 2) partially annotated, partial alignment with one or more Repbase sequences and 3) no significant alignment to a Repbase reference sequence. Partially annotated and unannotated consensus sequences were combined to produce the unclassified consensus repeat set.

CARP generated numerous consensus sequences (see Table 1 and S5 Table), because TEs, particularly LINEs, are often 5′ truncated, generating many insertion length variants and because consensus generation is based on alignment pairs that are single-linkage clustered with a length constraint (within 95% of the longest family member length). By comparing the repeat consensus sequences generated from CARP and RMD, we can see that CARP identified many more repeat sub-families, in contrast to RMD, which only generated a small number of broad consensus sequences. The latter are useful for masking, but are not as useful for studying TE evolution.

Table 1. Summary of consensus sequence libraries generated by CARP and RMD.

	Chicken		**Bearded dragon**		**Anolis**		**Platypus**		**Opossum**		**Human**	
	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD
Well-annotated	8,898	131	19,810	511	32,658	600	9,476	240	49,548	591	30,114	428
Unclassified	12,140	19	112,337	523	45,201	375	165,231	74	25,221	78	23,199	4
Total	21,038	150	132,147	1034	77,859	975	174,707	314	74,769	669	53,313	432

Consensus Classification

CARP generated annotated consensus sequences for all major TE types (except for SINEs in the chicken), whereas few SINE consensus sequences were produced by RepeatModeler in the species we tested (see Table 2 and S6 Table). Based on this result, CARP was more sensitive for detecting SINEs compared to RMD. CARP generated many more consensus sequences than RMD and this is a function of the

single linkage clustering used to identify families. Because many LINE insertions are 5' truncated, leading to variable insertion sizes with a common 3' end, the requirement for family members to be at least 95% as long as the longest family member means that many clusters are created across the insertion size continuum.

Table 2. Comparison of the total number of specific TE types in each method.

	Chicken		Bearded dragon		Anolis		Platypus		Opossum		Human	
	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD
SINE	0	2	1,613	44	2,177	30	2,292	41	596	73	13,165	23
LINE	6,186	51	16,068	248	18,014	270	6,290	122	25,150	244	10,832	137
LTR	2,405	54	385	55	4,784	76	263	27	23,195	212	5,635	164
DNA	102	17	1,725	146	7,619	202	46	30	600	55	239	91
Others	205	7	19	18	64	22	585	20	7	7	243	13
Total	8,898	131	19,810	511	32,658	600	9,476	240	49,548	591	30,114	428

Genome Repeat Content

CENSOR was used to annotate the repeat content in our data set of seven species because it uses minimal post-alignment processing of hits (see Table 3). In order to get a comprehensive annotation of repeats, we used a combination of the Repbase 'Vertebrate' library and repeat consensus sequences generated from CARP or RMD. Because CENSOR annotates based on the best hit, combining our consensus sequences with Repbase sequences allows annotation of genomic intervals most similar to either recent/less diverged repeats or Repbase repeats. As seen from Table 3, CARP performed consistently well in identifying and annotating repeats across all seven species (more detail in S7-S13 Tables).

Compared to RMD, CARP identified approximately the same amount of sequence made up of interspersed repeats in all seven species. However, CARP identified far more of all seven genomes as derived from unclassified repeats. Because unclassified repeats are defined as not being classifiable using Repbase, these repeats must either be novel transposable elements, or repeated sequences that are not transposable. In Table 3 we have labeled the unclassified repeat contribution to the genomes as segmental duplications based on their properties (see below).

Segmental Duplications

Because segmental duplications are generally present at low copy number, we examined the relationship between copy number and consensus length for our unclassified consensus sequences. We plotted the log_{10}-transformed copy number of the unclassified CENSOR hits against the log_{10}-transformed length for our unclassified consensus sequences from all seven genomes (see Figure 2). For both RMD and CARP unclassified sequences, copy number of hits increased with length due to a small tail of long, high hit number sequences. This can be seen from the regression line (S14 Table). However, virtually all CARP unclassified sequences were present at much lower hit copy number, a strong indication of segmental duplication. The small number of high hit copy number (>2000) CARP unclassified sequences were examined for the presence of either novel TEs or partial TEs.

In order to determine if the small number of unclassified CENSOR hits with copy numbers >2000 were novel or partially annotated TEs, we used coverage plots for the CARP unclassified consensus sequences to look for high copy number subsequences with

Table 3. Comparison of repeat annotation for CARP and RMD. Summary of specific repeat content from CENSOR output, using a combined library of Repbase 'Vertebrate' with CARP or RMD consensus libraries .
IR = Interspersed Repeats
SD = Segmental Duplications

	Chicken		Bearded dragon		Anolis		Platypus		Opossum		Human	
	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD
SINE	0.09	0.09	1.72	2.83	4.02	4.14	19.51	19.51	10.44	10.67	11.25	11.41
LINE	7.73	8.31	11.61	12.55	14.65	15.23	20.40	20.84	28.21	28.83	18.88	18.97
LTR	3.37	3.62	2.38	3.12	5.98	6.26	1.34	1.32	10.62	10.29	9.13	9.49
DNA	2.60	2.08	3.55	6.62	12.84	15.13	1.72	1.67	3.06	2.27	4.49	4.34
Other	1.52	2.22	1.39	1.44	1.51	1.65	2.89	3.10	1.72	1.47	2.04	1.93
IR	15.31	16.32	20.65	26.56	39.00	42.41	45.86	46.44	54.05	53.53	45.79	46.14
Potential SD	1.99	0.17	22.22	7.92	12.02	4.84	12.60	0.58	3.93	0.87	2.94	0.00
Total	17.30	16.49	42.87	34.48	51.02	47.25	58.46	47.02	57.98	54.40	48.73	46.14

TE properties. Figure 3 shows the top 5 high copy number CARP unclassified consensus sequences from bearded dragon as an example. BLASTN and CENSOR annotations were also used to characterize these consensus sequences in terms of TE or gene model homology. From Figure 3 we can see that coverage plots for high copy number CARP unclassified CENSOR hits were of two types: those incorporating high copy subsequences (Figure 3A,C,E) and those with uniform high coverage (Figure 3B,D). Close examination of the high copy hit subsequences from Figure 3A,C,E show that known TE annotation cannot explain the high copy number subsequences detected in these families. Because these three consensus sequences were derived from families with a small number of members, the observed high copy subsequences may indicate similarity to unclassified TEs or TE fragments that are present as part of a small number of highly conserved segmental duplications. For the uniform high coverage family 0309690 (Figure 3B), CENSOR annotated one end as the 5′ end of a DNA transposon (Mariner-3N1), and the other end as the 3′ end of the same DNA transposon, likely indicating a novel variant of a known DNA transposon. For the uniform high coverage family 137078, there is no known TE annotation, only annotation for a part of GPR34, a probable G-protein coupled receptor gene.

Based on the above results, we conclude that the vast majority of unclassified consensus sequences represent segmental duplications. We have therefore labeled these annotations accordingly in Table 3. In our final annotation, significant fractions of the genomes from our seven test species were annotated as SD, particularly in bearded dragon (22.22%), anolis (12.02%) and platypus (12.60%) (see Table 3).

Because the human genome has the best SD annotation of our seven species, we compared segmental duplication coordinates downloaded from the human 'Segmental Duplication Database' to our CARP unclassified CENSOR hits. Approximately 70% of human SD overlapped with CENSOR hits from CARP unclassified consensus sequences, confirming our conclusion above. Only 0.2% of human SD overlapped with RMD unclassified CENSOR hits.

CARP classification of TEs allows insight into TE evolutionary dynamics

Because CARP enabled us to identify and classify recently diverged repeats, we were able to determine whether those repeats were consistent with recent TE activity/family

expansion. We used the platypus to illustrate this. L2 and its non-autonomous SINE companion, mammalian-wide interspersed repeat (MIR, MON-1 in monotremes), are the most abundant and active repeats in monotremes (see S11 Table). This is in contrast to metatheria and eutheria (marsupials and placentals) where they are inactive due to extinction 60-100 Myr ago.

L2s were defined as potentially active if they contained an intact ORF2 (regardless of the state of ORF1), as this meant that they were capable of either autonomous retrotransposition [39] and/or mobilisation of SINEs [40]. CARP identified numerous long L2 elements (2-4kb) in the platypus genome. More than 43% (41/94) of these were potentially active based on the above criteria (Figure 4A) and some clusters of potentially active elements at the tips of short branches, were consistent with "hot" or hyperactive elements. This differed significantly from the RMD result, which generated only seven long consensus sequences (Figure 4C), none of them containing an intact ORF2.

It is worth noting that the Repbase annotation for L2s puts the full-length platypus L2 consensus sequences at 5kb long. However, based on both the CARP and RMD identification outputs, L2 elements in platypus were significantly shorter, at 3kb, with the longest one we could find only 3,110bp in length.

Discussion

Design considerations for bioinformatics pipelines or packages to identify and annotate repetitive sequences in DNA reflect the (sometimes unstated) goals of their developers. We have chosen to prioritise the identification of recent, slightly divergent repeats, to do so with a minimal number of tools and dependencies and to allow users the flexibility of choosing their own annotation tools (ie RepeatMasker or CENSOR). Ancestral, or previously characterised repeats can easily be detected using existing tools, but identifying novel repetitive elements, such as clade specific SINEs requires *ab initio* identification. It is also our experience that researchers sequencing a new genome usually want to identify repeats and segmental duplications early, and independently of gene model prediction. CARP is based on PALS and PILER, but improved and re-implemented in Go as krishna and igor. Our pipeline boils down to five simple steps: 1) find repeats using a pair-wise all vs all local alignment in your genome of choice, 2) use single linkage clustering with a length constraint to create repeat families from the alignments, 3) generate consensus sequences from repeat families and annotate them using RepBase and reverse transcriptase sequences and TE sequences from NCBI, 4) filter out protein coding genes by alignment to Swiss-Prot and 5) combine the *ab initio* library with RepBase to annotate both TEs and candidate segmental duplications.

At present RMD is the most widely used *ab initio* TE identification package, but it has limitations, particularly for users interested in the evolution of TEs. It only provides broad consensus sequences and does not allow one to determine what sequences contributed to a consensus. REPET can provide the sequences/genome intervals used to generate the consensus sequences from PILER families, but not for GROUPER or RECON families. Neither RMD nor REPET removes families obtained from gene families as does CARP. REPET will use gene models to filter out gene repeats, but if no gene model intervals are available this is not an option.

Neither RMD nor REPET are designed to detect segmental duplications. REPET in particular is designed to remove low copy number families from the analysis in order to avoid having segmental duplications in the final consensus set. At present, segmental duplications are detected using all vs all pair-wise alignments of TE repeat masked genomes, because TE repeats generate a huge number of alignments that mask the *bona fide* segmental duplications. This masking of TEs also reduces the sensitivity of existing

segmental duplication approaches as TEs are a significant component of segmental duplication sequences. CARP generates consensus sequences from low copy repeats (segmental duplications) without masking, which improves the sensitivity of segmental duplication detection. When we compared our segmental duplication annotation to what has been reported for these seven species, we found that our method detected more candidate segmental duplications in the anolis (4.9%) [41] (Table 3) and the opossum (1.7%) (Table 3) [42].

Finally, the platypus genome is made up of almost 21% LINE L2 sequences, which is an extraordinarily high percentage. Such a high percentage of a single repeat type usually means that there are many actively retrotransposing elements in the genome. As part of CARP's standard output, we were able to identify 41 potentially active, L2 elements in the platypus genome with minimal additional analysis (Figure 4).

Conclusion

Here we introduce a simple and flexible *ab initio* repeat identification and annotation method (CARP) that annotates TEs and candidate segmental duplications. We applied CARP to seven animal genomes and demonstrated that it performs as well or better than RepeatModeller, the most commonly used *ab initio* TE annotation package.

Limitation: Our approach is limited by memory requirements and runtime. However, as hardware improves and becomes less expensive, these limitations will become less of an issue.

Supporting information

S1 Fig. Coverage plot of high copy number unclassified repeats in the anolis genome. Shows the coverage plot for the top 12 highest copy number (>2,000 copies) unclassified consensus sequences in the anole genome.

S2 Fig. Coverage plot of high copy number unclassified repeats in the opossum genome. Shows the coverage plot for the top 21 highest copy number (>2,000 copies) unclassified consensus sequences in the opossum genome.

S3 Fig. Coverage plot of high copy number unclassified repeats in the human genome. Shows the coverage plot for the highest copy number (>2,000 copies) unclassified consensus sequence in the human genome.

S1 Appendix. CARP documentation. Gives a detailed account of how to use our *ab initio* method to identify and annotate TEs from a genome assembly, including the benchmarks used for the seven species in this report.

S1 Table. Genome dataset. Shows the systematic name, common name, genome version, source and submitter for all the genomes tested for our *ab initio* method.

S2 Table. Assembly statistics. Shows the systematic name, total sequence length, scaffold N50, contig N50 and assembly level.

S3 Table. Assembly method and coverage. Shows the systematic name, assembly method, sequencing techonology and estimated genome coverage for the seven genomes in this study.

S4 Table. Benchmarks for each method. Here we show the compute time used for the seven tested species with CARP and RMD.

S5 Table. Summary of library lengths generated from two methods. Total length (bp) of consensus sequence libraries generated by CARP and RMD.

S6 Table. Summary of specific TE length generated from two methods. Comparison of the total consensus sequence lengths (bp) of specific TE types generated by CARP and RMD.

S7 Table. Repeat content in the chicken genome. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the chicken genome.

S8 Table. Repeat content in the anolis genome. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the anolis genome.

S9 Table. Repeat content in the bearded dragon genome. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the bearded dragon genome.

S10 Table. Repeat content in the platypus genome. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the platypus genome.

S11 Table. Repeat content in the opossum genome. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the opossum genome.

S12 Table. Repeat content in the human genome. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the human genome.

S13 Table. Linear regression for unknown sequence copy number against length. Shows the estimate value, standard error, t-value, p-value and significance codes from linear regression analysis. Significance asterisks follow the conventions of R, i.e. ***, **, *, ., for p-values below 0.001, 0.01, 0.05 and 0.1 respectively.

Acknowledgments

We would like thank James Galbraith for taking the time to read the manuscript in full and offering helpful comments. Finally, this paper would not be possible without the invaluable insights and writing help from Atma Ivancevic, brainstorming from Reuben Buckley, and extraordinary IT support from Matt Westlake.

Availability and requirements

The Go source code are available on github: `github.com/biogo/examples/krishna`

References

1. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nature reviews Genetics. 2009;10(10):691.
2. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science. 2009;326(5956):1112–1115.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. 2001;.
4. Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human genome. TRENDS in Genetics. 2001;17(11):661–669.
5. Muñoz-López M, García-Pérez JL. DNA transposons: nature and applications in genomics. Current genomics. 2010;11(2):115–128.
6. Skipper KA, Andersen PR, Sharma N, Mikkelsen JG. DNA transposon-based gene vehicles-scenes from an evolutionary drive. Journal of biomedical science. 2013;20(1):92.
7. Khodosevich K, Lebedev Y, Sverdlov E. Endogenous retroviruses and human evolution. Comparative and functional genomics. 2002;3(6):494–498.
8. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, Van de Lagemaat LN, Mager DL. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS genetics. 2006;2(1):e2.
9. Lee SI, Kim NS. Transposable elements and genome size variations in plants. Genomics & informatics. 2014;12(3):87–97.
10. Holton NJ, Goodwin TJ, Butler MI, Poulter RT. An active retrotransposon in Candida albicans. Nucleic acids research. 2001;29(19):4014–4024.
11. Matthews GD, Goodwin T, Butler MI, Berryman TA, Poulter R. pCal, a highly unusual Ty1/copia retrotransposon from the pathogenic yeast Candida albicans. Journal of bacteriology. 1997;179(22):7118–7128.
12. Nefedova L, Kim A. Molecular phylogeny and systematics of Drosophila retrotransposons and retroviruses. Molecular biology. 2009;43(5):747.
13. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993;72(4):595–605.
14. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. Genome Biol Evol. 2016;8(11):3301–3322. doi:10.1093/gbe/evw243.
15. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0.; 2013-2015.
16. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research. 2005;110(1-4):462–467.
17. McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003;19(3):362–367.

18. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011;6(1):e16526. doi:10.1371/journal.pone.0016526.

19. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC bioinformatics. 2015;16(1):227.

20. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21(suppl_1):i152–i158.

21. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(suppl_1):i351–i358.

22. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome research. 2002;12(8):1269–1276.

23. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research. 1999;27(2):573.

24. Quesneville H, Nouaud D, Anxolabéhère D. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. J Mol Evol. 2003;57 Suppl 1:S50–9. doi:10.1007/s00239-003-0007-2.

25. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1(2):166–75. doi:10.1371/journal.pcbi.0010022.

26. Kortschak RD, Adelson DL. bíogo: a simple high-performance bioinformatics toolkit for the Go language. bioRxiv. 2014; p. 005033.

27. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

28. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC bioinformatics. 2006;7(1):474.

29. Gish W. Wu-blast; 1996.

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403–410.

31. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research. 2006;35(suppl_1):D61–D65.

32. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research. 2003;31(1):365–370.

33. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842.

34. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution. 2000;17(4):540–552.

35. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution. 2009;26(7):1641–1650.

36. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–2461.

37. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011;39(suppl_2):W29–W37.

38. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic acids research. 2004;32(suppl_1):D138–D141.

39. Heras S, Thomas M, Garcia-Canadas M, De Felipe P, Garcia-Perez J, Ryan M, et al. L1Tc non-LTR retrotransposons from Trypanosoma cruzi contain a functional viral-like self-cleaving 2A sequence in frame with the active proteins they encode. Cellular and Molecular Life Sciences CMLS. 2006;63(12):1449–1460.

40. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nature genetics. 2003;35(1):41.

41. Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 2011;477(7366):587.

42. Samollow PB. The opossum genome: insights and opportunities from an alternative mammal. Genome research. 2008;18(8):1199–1215.

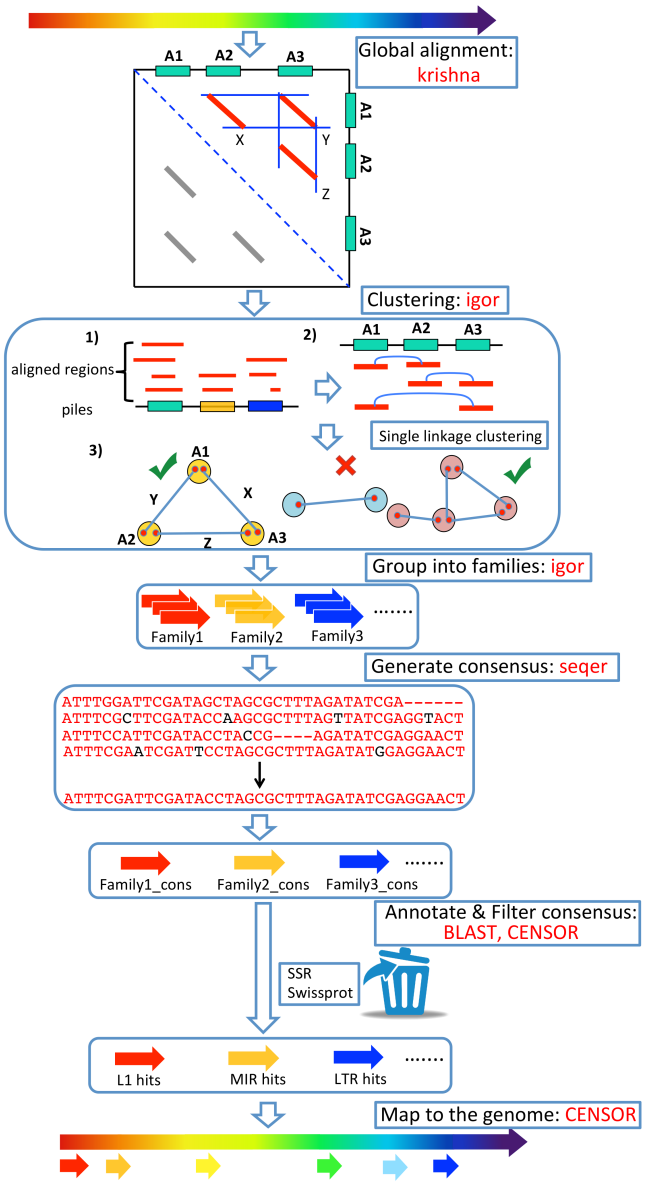


Fig 1. Comprehensive *Ab initio* Repeat Pipeline (CARP). Figure shows the detailed steps for CARP. Repetitive DNA is identified by all vs all pairwise alignment using krishna. Single linkage clustering is then carried out to produce families of repetitive sequences that are globally aligned to generate a consensus sequence for each family. Consensus sequences are filtered for non-TE protein coding genes and then annotated using Repbase and a custom library of retrovirus and reverse transcriptase sequences. The annotated consensus sequences are then used to annotate the genome. This is required to identify repeats with less than the threshold identity used for alignment that are overlooked during the initial pairwise alignment step.

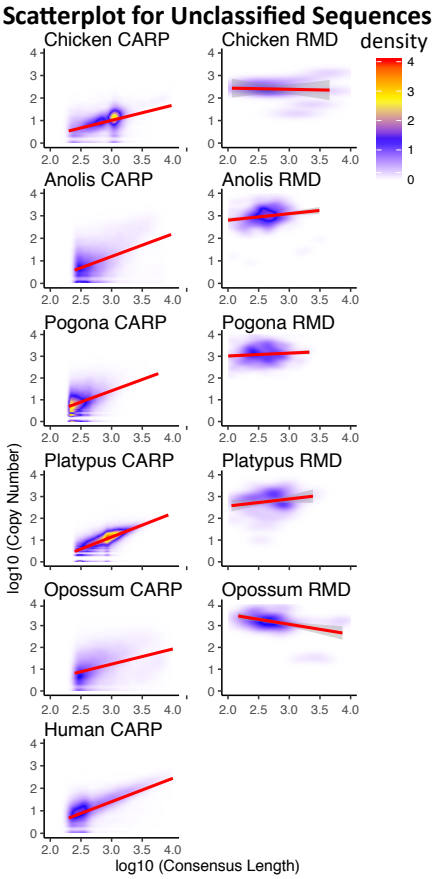


Fig 2. Scatter plot of unclassified sequence copy number *versus* length. Plots show the copy number of hits of unclassified sequences annotated using CENSOR and combined libraries, with respect to their length. Both copy number and length were log_{10}-transformed. Red regions on the plot indicate high density, while blue regions indicate low density. Linear regression lines are plotted in red, with STANDARD ERROR represented by the gray shadow around the lines.

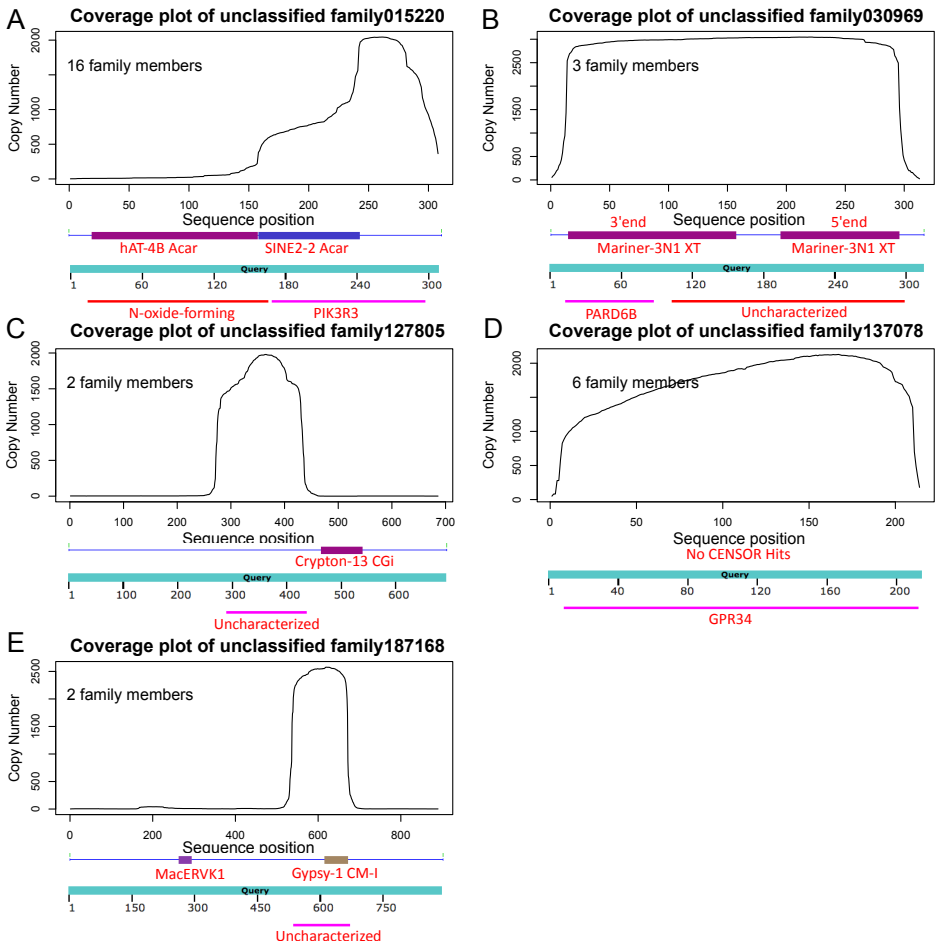


Fig 3. Coverage plot of the top 5 high hit copy number CARP unclassified consensus sequences from the bearded dragon. A) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 015220; B) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 0309690; C) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 127805; D) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 137078; E) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 187168. The number of family members identified by krishna/igor used for consensus sequence generation is shown in the upper left corner of each panel.

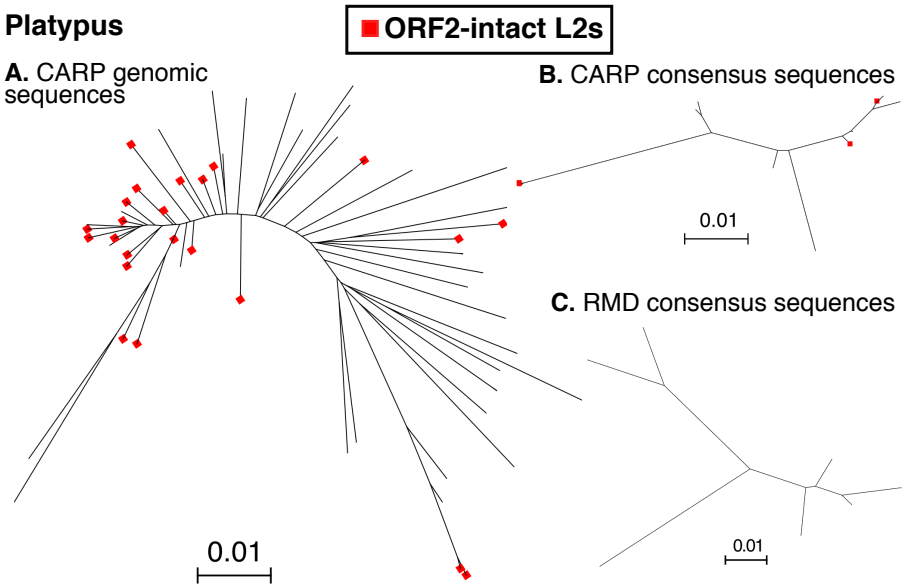


Fig 4. Phylogenetic analysis of L2 elements in the platypus genome. Figure shows the dendrograms of full-length L2 elements in the platypus genome. Panel A) long L2 sequences from the platypus genome. Panel B) Long L2 CARP consensus sequences from platypus. Panel C) Long L2 RMD consensus sequences from platypus. Sequences were aligned with MUSCLE, trees inferred with FastTree and visualized with Archaeopteryx. ORF2-instact L2s are shown with a red dot at the tip of the branch.

Chapter 3

Ab initio identification and annotation of DNA repeats in the tuatara genome

> *"Science is not a thing. It's a verb. It's a way of thinking about things. It's a way of looking for natural explanations for all phenomena."*
>
> — Michael Shermer

The *Sphenodon punctatus* (tuatara) forms the deepest branch within lepidosauria, the sister taxon of birds, crocodilians and turtles. So far, the anole lizard is the only comprehensively studied representative of lepidosaurs and contains virtually no ancient TE copies due to rapid DNA sequence turnover. In-depth analysis of the tuatara repetitive landscape therefore promises to yield more detailed insights into the genome organization of the amniote ancestor. I am interested in investigating how transposable elements shape the tuatara genome, what the differences of transposable elements in the tuatara genome compared to other lepidosaurs, and how these TEs drives the evolution of tuatara. This manuscript will be merged into the Tuatara Genome Consortium paper, and is being prepared for submission to *Nature*.

Statement of Authorship

Title of Paper	*Ab initio* identification and annotation DNA repeats in the tuatara genome
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Lu Zeng, R. Daniel Kortschak, Joy M. Raison, Terry Bertozzi, David L. Adelson (2018). *Ab initio* identification and annotation DNA repeats in the tuatara genome. Prepared for submission as a Scientific Report to Nature.

Principal Author

Name of Principal Author (Candidate)	Lu Zeng		
Contribution to the Paper	Processed data, performed analysis, prepared figures and wrote the manuscript		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	20/02/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	R. Danial Kortschak		
Contribution to the Paper	Supervised the development of work, provided the code, assisted with analysis of data and assisted in writing the manuscript.		
Signature		Date	26/2/2018

Name of Co-Author	Joy M. Raison		
Contribution to the Paper	Assisted with development of data analysis techniques and edited manuscript		
Signature		Date	22/2/18

Name of Co-Author	Terry Bertozzi
Contribution to the Paper	Assisted with development of data analysis techniques and edited manuscript.
Signature	Date 26/2/18.

Name of Co-Author	David L. Adelson
Contribution to the Paper	Supervised the development of work and assisted in analysing the results and writing the manuscript.
Signature	Date 26/2/18

Ab initio identification and annotation of DNA repeats in the tuatara genome

Lu Zeng[1], R. Daniel Kortschak[1],
Joy M. Raison[1], Terry Bertozzi[2], David L. Adelson[1*]

[1]School of Biological Sciences, The University of Adelaide, SA 5005, Australia
[2]Evolutionary Biology Unit, South Australian Museum, SA 5005, Australia

[*]Corresponding author: david.adelson@adelaide.edu.au

Abstract

We used combined approaches, including *ab initio* repeat identification and annotation methods to investigate the repeat content of the tuatara genome. Our analysis revealed that the tuatara genome contains an unusually large proportion (12.5%) of non-LTR LINE 2 retrotransposons (Table 1), even though the total transposable element content (30%) is similar to other reptiles. We found two main sub-families of LINE L2 in the Tuatara, one similar to lizard L2s and the other similar to platypus L2s (Figure 2). The latter provide evidence for a putative horizontal transfer event of L2s between lizards and monotremes. Tuatara SINE elements are present at one tenth of their proportional level compared to anole lizard (0.475% *vs* 4.019%) yet are of recent origin. While there are differences in LINE *vs* SINE expansions in Tuatara compared to Anole, the overall contribution to genome size is proportionally similar in both species. Finally, an extraordinarily high proportion (33%) of the tuatara genome originates from low copy number segmental duplications, with 6.7% of these of recent origin based on their high level of sequence identity (>94% identity), which is more than seen in other vertebrates. The tuatara genome is 2.4x longer than the anole genome and this difference appears to be driven disproportionately by segmental duplications, many of which are of recent origin.

Introduction

Transposable elements (TEs) are genomic parasites that are distributed all across the tree of life and are largely responsible for genome size differences among cellular organisms [1]. Their remarkable diversity has been classified into several main groups based on their distinct mechanisms of propagation [2]: DNA transposons move via a 'cut-and-paste' mechanism, whereas retrotransposons have an RNA intermediate and proliferate via 'copy-and-paste'. Retrotransposons are further subdivided into long terminal repeat (LTR) retrotransposons that closely resemble retroviruses, and non-LTR retrotransposons. Among the latter, long interspersed elements (LINEs) are parasitized by non-autonomous elements, so-called short interspersed elements (SINEs).

Birds and mammals are very different amniotes in terms of genome size and repetitive element organization. Avian genomes are relatively small with generally $<$10% TEs [3] [4] [5], whereas mammalian genomes are about three times as large and often contain a TE density of 50% [6] [7] [8]. However, a recent study found that the axolotl genome has an even greater TE density at 65.6% for a total of 18.6 Gb of repetitive sequence[9]. With the exception of the anole lizard that exhibits a highly diverse landscape of TE activity [10] [11], the hitherto analyzed amniote genomes are largely dominated by activity of a single LINE superfamily, namely L1 in placental and marsupial mammals [6] [7], and L2 in monotreme mammals [8], with L2 as the most ancient family of retrotransposons in the human genome [6]. Furthermore, CR1 LINEs are the major "genome component" [12] in birds [3] [4] [5], crocodilians [13], snakes [14], and turtles [15]. Given the observation that ancient CR1 lineages are also present in mammals, it has been suggested that the dominance of CR1 in sauropsids resembles the genome organization of the amniote ancestor [16] [17] [18].

Based on previous literature, active LINEs are usually defined as full-length between 3-7kb long containing a 5′-untranslated region (5′-UTR) with an internal promoter; two open reading frames (ORF1 and ORF2) separated by an intergenic region, and a 3′ UTR containing

a poly A tail [19]. All known LINE ORF2 encode a 150-kDa protein which contains an apurinic endonuclease (APE) and reverse transcriptase (RT) domains that provide the enzymatic activities for cDNA synthesis and host genome nicking during the replication cycle [20]. LINE elements with an intact ORF2p can retain the ability to mobilise SINEs within the genomes [21].

Horizontal transfer (HT) is the transmission of genetic material by means other than parent-to-offspring: a phenomenon primarily considered in prokaryotic context. However, given a vector of transfer (e.g. virus, parasite), retrotransposons have the innate ability to jump between species as they do within genomes [22] [23]. Previous studies have investigated the possibility of HT in retrotransposons, including CR1s and RTEs [24] [25] [26].

Sphenodon punctatus (tuatara) forms the deepest branch within lepidosauria, the sister taxon of birds, crocodilians and turtles [27]. So far, the anole lizard is the only comprehensively studied representative of lepidosaurs and contains virtually no ancient TE copies due to rapid DNA sequence turnover [10] [11] [28]. In-depth analysis of the tuatara repetitive landscape therefore promises to yield more detailed insights into the genome organization of the amniote ancestor.

In this study, we describe an *ab initio* repeat identification and annotation method to investigate the repeat content of the tuatara genome. In order to evaluate the performance of our *ab initio* method, we have compared the repeat libraries generated from our method and from RepeatModeler (RMD) [29]. Our analysis revealed that most repeats in the tuatara genome are non-LTR LINE L2 retrotransposons. We also investigated the evolutionary relationships between tuatara L2 sequences and those from other vertebrates using both the full nucleotide sequences and the Reverse Transcriptase (RT) domains of these L2 sequences. We found two main sub-families of L2 in the tuatara, one similar to lizard L2s and the other similar to Platypus L2s. The latter provide potential evidence for a horizontal transfer event of L2s between lizards and monotremes or between an as yet undefined organism and both tuatara and monotremes.

Finally, there were a large number of repeated sequences from our *ab initio* method and these are likely segmental duplications in the tuatara genome. There were, however, some highly repeated sub-sequences in these duplications that could represent potential novel SINEs.

Results

Repeat coverage in the tuatara genome

Using the *ab initio* method, 60% of the tuatara genome was annotated as repetitive (Table 1), and 28% of the genome was comprised of known repeats. This indicated that the tuatara genome was significantly enriched with repeats compared to other reptile genomes.

One interesting observation was that the fraction of non-LTR retrotransposons in the tuatara genome was much higher (18.7%) than in placental mammals, with L2 the dominant LINE element. We estimated that there are 1.6 million L2 sequences (full length and fragments) in the tuatara genome, and the longest L2 sequences we found were between 2-4kb in length. The second most abundant repeat type was CR1 (chicken repeat 1), which comprised 3% of the tuatara genome.

In contrast to our method, RepeatModeler (RMD) identified approximately 50% of the tuatara genome as repetitive, and 31% of this was annotated based on Repbase data. L2 and CR1 were also the two dominant non-LTR retrotransposons, but at a lower level for L2 compared with our method. RMD identified 8% of the tuatara genome as DNA transposons, mainly from the DNA/hAT (*hobo/AC/Tam3*) superfamily (4.9%) and the Harbinger superfamily (3.1%).

Table 1: Copy number and fraction of tuatara genome covered by interspersed repeats.

	ab initio library			RMD library		
Group	Number	Total bp	Percentage	Number	Total bp	Percentage
Non-LTR retrotransposons						
LINE L2	1617,025	533,695,311	12.493	51,6398	181,856,555	4.257
LINE CR1	608,039	129,476,630	3.031	513,323	148,669,465	3.480
LINE RTE	176,236	36,629,395	0.858	192,878	92,350,494	2.162
LINE L1	242,327	34,056,859	0.797	837,73	34,808,672	0.815
LINE L3	107,843	25,425,588	0.595	8,769	1,678,297	0.039
SINE	188,583	20,270,432	0.475	154,567	21,950,928	0.514
Others	170,168	18,451,758	0.432	222,465	21,032,662	0.713
	3,110,221	798,005,973	18.681	1,692,173	502,347,073	11.980
ERVs						
ERV1	114,470	7458957	0.175	61,520	24,396,523	0.571
SloEFV	496	389,867	0.009	NA	NA	NA
Others	108,195	11,682,102	0.273	26,280	11,360,641	0.266
	223,161	19,530,926	0.457	87,808	35,757,164	0.837
DNA transposons						
hAT	450,228	62,937,914	1.473	695,777	209,398,652	4.902
Harbinger	195,754	26,461,644	0.619	292,843	13,391,0630	3.135
Others	341,056	31,185,671	0.730	35,023	9,974,492	0.233
	987,038	120,585,229	2.822	1,023,643	357,167,302	8.495
LTR						
DIRS	275,147	71,266,408	1.668	160,508	85,671,820	2.005
Gypsy	514,850	63,228,939	1.480	379,412	201,432,961	4.715
LTR others	72,278	19,933,116	0.467	170,927	70,550,760	1.652
	862,275	154,428,463	3.615	710,847	357,655,541	8.372
Others	1,182,484	104,435,710	2.445	235,998	91,596,806	1.790
Well-annotated	6,365,179	1,196,986,301	28.021	3,750,469	1,344,523,886	31.474
Unknown	6,798,135	1,418,242,189	33.200	3,102,725	706,552,451	16.540
Total	13,163,314	2,615,228,490	61.221	6,853,194	2,051,076,337	48.014

Classification of L2 elements in the tuatara genome

Based on both repeat annotation methods, L2 elements were found to be the most abundant repeat type in the tuatara genome.

In the RMD repeat library, only 6 L2 sequences were longer than 2kb, with 2 of them approximately 3kb long. While in our *ab initio* library, 49 consensus sequences were longer than 2kb with one over 3kb long. Figure 1 shows the sequence similarity of L2 elements (2-4kb) from these two methods. RMD L2 sequences were found to cluster with our L2 consensus sequences, and were most similar to turtle and platypus L2s based on CENSOR annotation. In addition, L2 consensus sequences from our method that were annotated as most similar to platypus L2, were also found always clustered with turtle L2 based on CENSOR annotation. However, platypus L2 from the Repbase library were classified as CR1. This is one example of the confusing state of repeat annotations and nomenclature. The remaining L2 consensus sequences (including the longest L2) were found to be similar to anole L2 sequences based on Repbase annotation.

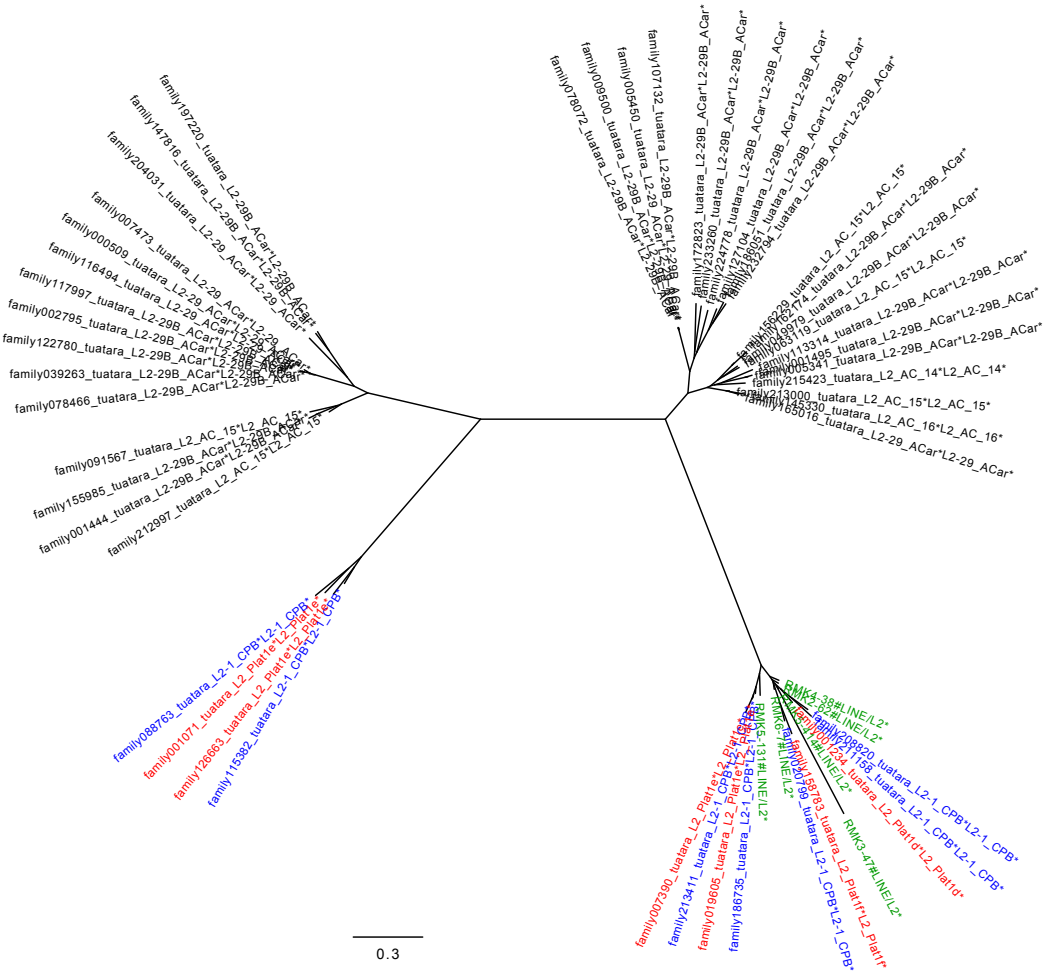


Figure 1: **Sequence similarity analysis based on RMD L2 (2-4kb) and *ab initio* L2 (2-4kb) consensus sequences.** Maximum likelihood dendrogram inferred using FastTree based on global alignment of tuatara L2 nucleotide sequences generated from both our *ab initio* method and RMD. Sequences were aligned with MUSCLE, and visualised with FigTree. Dark blue labels are tuatara L2 sequences similar to turtle, red labels are tuatara L2 similar to platypus, black labels are L2 sequences similar to anole lizard, green labels are RMD tuatara sequences.

Tuatara L2 do not cluster with chicken CR1

As mentioned above, platypus L2 sequences are categorised as belonging to the CR1 clade according to Repbase. Therefore, we used full-length chicken CR1 sequences to analyse the overall clustering pattern of tuatara L2 sequences (Figure 2). With the inclusion of chicken CR1 sequences, tuatara L2 sequences split into 4-5 groups; two groups contain repeats similar to platypus L2, turtle L2 and RMD L2, the other three groups are made up of L2 sequences that

are most similar to anole lizard. The full-length chicken CR1 elements did not cluster with the L2 consensus sequences, indicating that all of the L2 sequences we have identified are not CR1 like, in spite of the RepBase annotation.

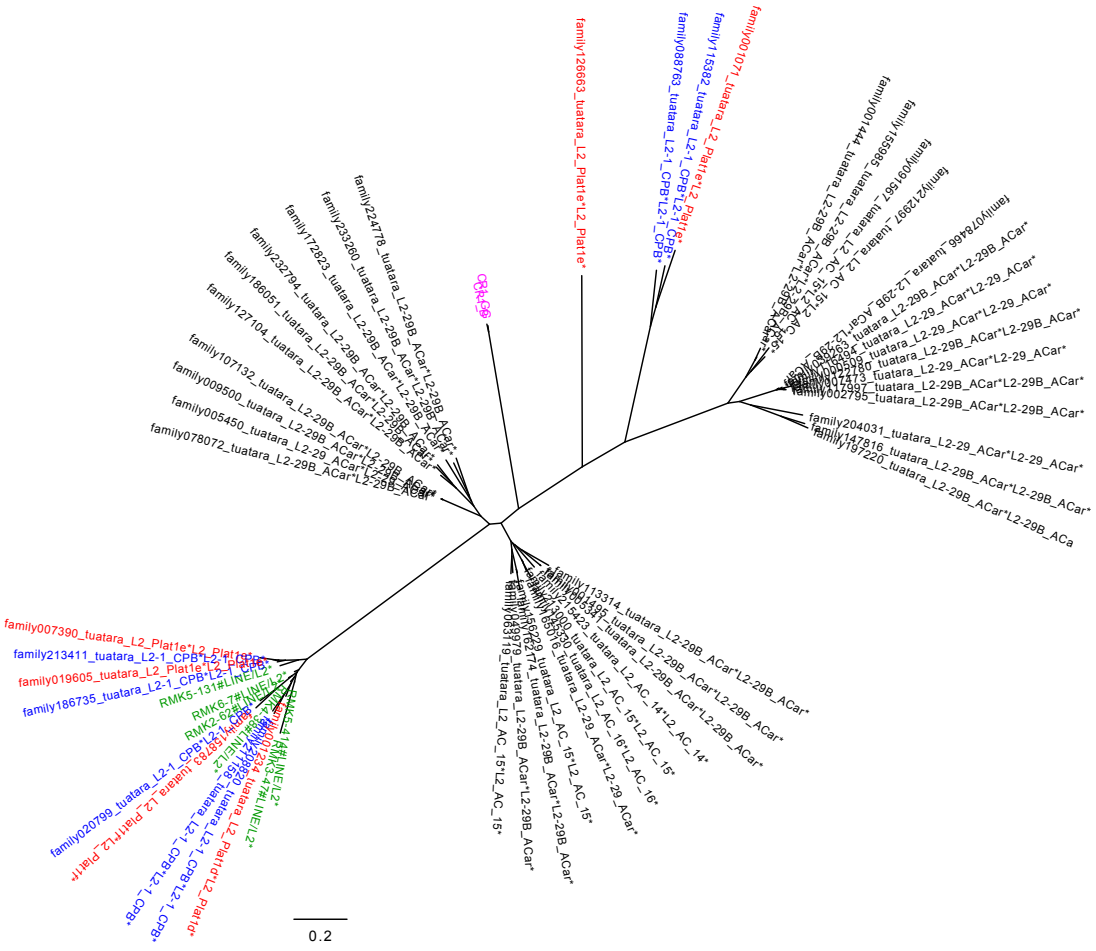


Figure 2: **Tuatara L2 and chicken CR1 do not cluster together.** Maximum likelihood dendrogram inferred using FastTree based on global alignment of tuatara L2 nucleotide sequences and chicken CR1. Sequences were aligned with MUSCLE, and visualised with FigTree. All sequences were generated from the *ab initio* method, except green labelled sequences, that are from RepeatModeler. Dark blue labels are tuatara L2 sequences similar to turtle, red labels are tuatara L2 similar to platypus, black labels are L2 sequences similar to anole lizard, pink labels are chicken CR1 from Repbase.

Phylogenetic analysis of tuatara L2 compared to other vertebrates

We then globally aligned the L2 tuatara sequences with L2 sequences from other vertebrates. The resulting maximum likelihood based phylogenetic analysis clearly showed that tuatara L2 sequences are divided into two groups (Figure 3), one group is more similar to bearded dragon and anole L2 (L2 from the *ab initio* method), while the other group clustered with turtle and crocodile L2 (including L2 from RMD and the *ab initio* method). Platypus like L2 consensus sequences identified from the *ab initio* method consistently clustered with platypus L2 from RepBase. The presence of two lineages of L2 (reptile vs monotreme) in tuatara is not characteristic of other reptiles and may be the result of incomplete lineage sorting of L2 in tuatara or raises the possibility of horizontal transfer of L2 between monotremes and the tuatara lineage.

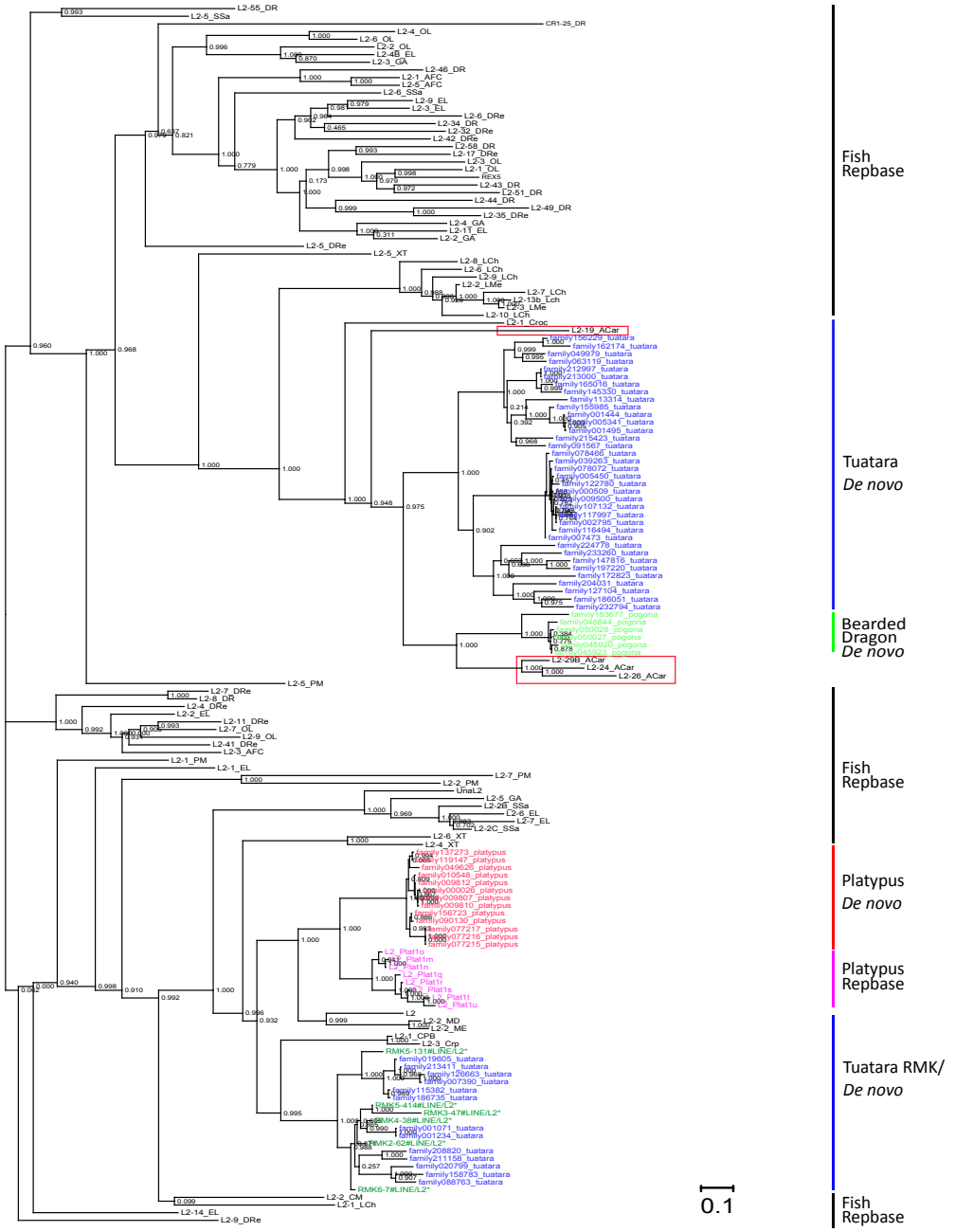


Figure 3: **Phylogenetic analysis of vertebrate L2.** Maximum likelihood dendrogram inferred using FastTree based on global alignment of L2 nucleotide sequences. L2 sequences were either extracted from species full genome assemblies (tuatara, platypus and bearded dragon) or sourced from RepBase. Sequences were aligned with MUSCLE, and visualised with Archaeopteryx. Branches are coloured to indicate the original L2 source. Dark blue labels are tuatara L2 sequences generated using the *ab initio* method, dark green labels are tuatara L2 sequences from RepeatModeler, pink labels are platypus sequences from RepBase, red labels are platypus sequences from the *ab initio* method and light green labels are bearded dragon *ab initio* sequences. Red boxes in the upper part of the tree show anole sequences.

Tuatara L2 may still be active

In order to investigate if L2 elements are still active in the tuatara genome, we carried out open reading frame (ORF) analysis of the long tuatara L2 consensus sequences (2-4kb) and their corresponding genomic sequences to identify potentially active L2 elements.

With respect to the *ab initio* consensus sequences, we found that 49 of the tuatara consensus sequences were longer than 2kb, with 1 of them longer than 3kb. We also found that 36 (73%) of these consensus sequences appeared to contain intact ORF2p (longer than 500 amino acids and contained a RVT_1 reverse transcriptase motif). While in tuatara genomic sequences we found that 711 of the tuatara genome fragments were longer than 2kb, with four of them longer than 3kb. Significantly, 541 (76%) of these genomic sequences appeared to contain intact ORF2p. See Table 2. This strongly suggests that L2 elements may still be active in the tuatara genome.

Table 2: Copy number and fraction of the tuatara genome covered by interspersed repeats.

	3-4kb		2-3kb	
	Number full length	% active L2	Number full length	% active L2
RMD L2 consensus	2	100	4	50
ab initio L2 consensus	1	100	48	73
ab initio L2 genome	4	100	711	76

We also carried out global alignments of >500aa long ORF2p sequences from tuatara and RepBase consensus sequences that had RT domains >200aa in length (see methods). The phylogenetic analysis of RT families (Figure 4) clearly illustrated differences between L2 groups. This figure shows that tuatara L2 RT domains differed from anole lizard and split into two clusters, one cluster was closer to platypus, while the other cluster was closer to bearded dragon. The L2 RT domain from crocodile was separate from all other species. Although the L2 nucleotide phylogeny tree showed that tuatara L2 shared high similarity with anole L2, when comparing sequences based on the RT domain, they were quite different, with anole RT domains

clearly separated from other species. This result is consistent with either incomplete lineage sorting of two L2 families or horizontal transfer of L2 sequences to or from a monotreme. Both of these families may still be active.

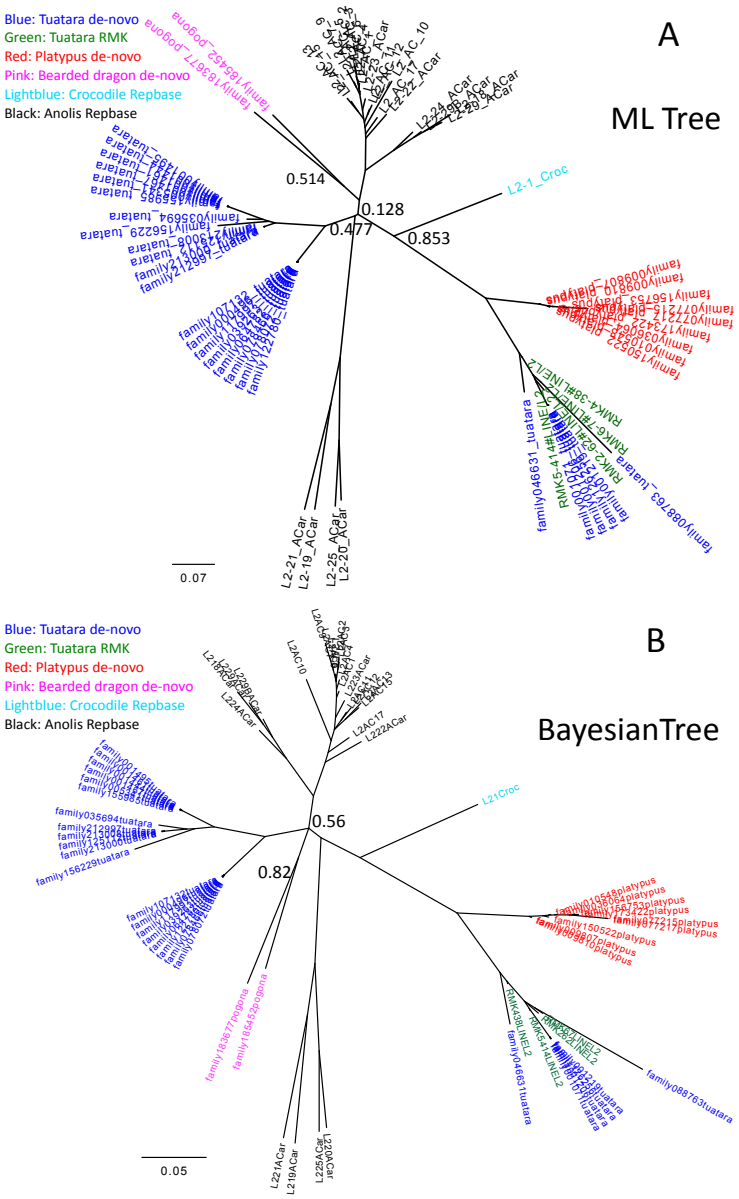


Figure 4: **Phylogenetic analysis based on the RT domain.** A) Maximum likelihood dendrogram inferred using FastTree from L2 RT domain multiple alignment. Nucleotide sequences were aligned with MUSCLE, and the phylogeny visualised with FigTree. Branches are coloured to indicate the original L2 source: dark blue labels are tuatara L2 RT domain generated from the *ab initio* method, dark green labels are tuatara L2 RT domains from RepeatModeler, red branches are platypus L2 RT domains, light green branches are L2 RT domains from bearded dragon and black branches were L2 RT domains from Anolis. B) Bayesian dendrogram inferred using MrBayes from L2 RT domain multiple alignment. Nucleotide sequences were aligned with MUSCLE, and the phylogeny visualised with FigTree. Branch labeling same as panel A. Only support values lower than 0.9 are shown in this figure, support values from panel A used a bootstrapping approach, while support values from panel B were posterior probability.

Possible horizontal transfer of L2 between monotreme and tuatara

Based on both nucleotide and RT domain phylogenies, L2 elements consistently split into two groups, one most similar to lizard L2 (43 consensus sequences), the other most similar to platypus L2 (17 consensus sequences), indicating a possible horizontal transfer event between monotreme and tuatara. The platypus and platypus like L2 sequences are far more similar to each other than to the anole L2 sequences. We used these sequences to generate custom L2 libraries for repeat annotation (see below).

We developed a method based on reciprocal best hits for identifying potential horizontally transferred sequences. First, CENSOR was used to find hits that annotated as platypus-like tuatara L2 or non platypus-like tuatara L2 in five reptile genomes (anole lizard, alligator, crocodile, turtle and bearded dragon) and one monotreme genome (platypus) using the custom L2 libraries. Second, CENSOR hit sequences were extracted and used as BLASTN queries to find reciprocal best hits in the custom libraries. This allowed us to determine the reliability of the original annotation of L2 hits and determine the L2 family they are most likely to belong to.

Based on the initial CENSOR output (Table 3), platypus-like tuatara L2 elements were found to be enriched in the platypus genome, while non-platypus tuatara L2 elements were abundant in both the anole lizard and bearded dragon genomes. Validation of these hits using reciprocal BLASTN shows that although many fragments in reptiles annotated as platypus-like tuatara L2 from CENSOR, they did not validate as reciprocal hits (Table 4). This is most likely a product of different parameter settings for BLASTN used by CENSOR and the more stringent settings used in the reciprocal BLASTN search. However, 15/17 platypus-like L2 tuatara consensus sequences were validated as being most similar to platypus L2 based on their reciprocal best hits. Similarly, 41/43 reptile like tuatara L2 consensus sequences were validated as non platypus-like based on the reciprocal best hit results.

Table 3: **CENSOR output**: Summary of non-redundant CENSOR hit genome intervals used as queries for reciprocal BLASTN alignment. Number of consensus sequences in custom library (in parentheses).
PT L2 = Platypus-like tuatara L2;
NPT L2 = Non platypus-like tuatara L2;
int = Intervals;
con = Consensus.

CENSOR library	Anolis		Alligator		Crocodile		Bearded dragon		Turtle		Platypus	
	int	(con)	int	(con)	int	(con)	int	(con)	int	(con)	int	(con)
PT L2 (17)	395	(17)	1,183	(17)	9,893	(17)	1,044	(17)	2,333	(17)	60,668	(17)
NPT L2 (43)	3,009	(43)	231	(43)	2,504	(43)	6,387	(43)	576	(43)	105	(41)

Table 4: **BLASTN output**: Summary of the BLASTN outputs for the reciprocal alignments of the intervals from Table 3 against platypus-like tuatara L2 and non platypus-like tuatara L2 consensus sequences. Number of validated consensus sequences (in parentheses).

L2 Type	Anolis		Alligator		Crocodile		Bearded dragon		Turtle		Platypus	
	int	(con)	int	(con)	int	(con)	int	(con)	int	(con)	int	(con)
PT L2 (17)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	**1,381**	**(15)**
NPT L2 (43)	556	(19)	6	(4)	2	(2)	458	(25)	0	(0)	0	(0)

Furthermore, in order to estimate the relative ages of platypus-like tuatara L2 and non platypus-like tuatara L2, super consensus sequences were generated for each L2 type (see Methods), RepeatMasker [30] was then used to calculate the divergence rate between super consensus and corresponding tuatara L2 elements (e.g. platypus-like tuatara L2 super consensus against platypus-like tuatara L2 elements). Figure 5 clearly shows that compared to non platypus-like tuatara L2, platypus-like tuatara L2 has a much lower sequence substitution level, which indicates that the platypus-like tuatara L2 sequences are of more recent origin, and may have resulted from horizontal transfer from platypus, or some third, as yet unidentified species, into the tuatara genome.

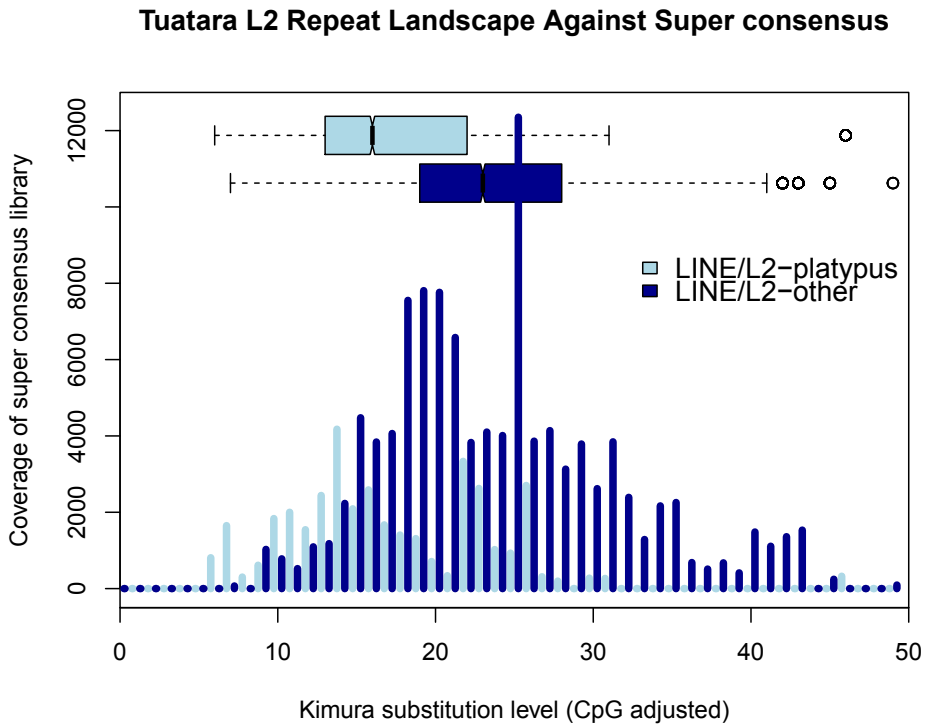


Figure 5: **Kimura substitution level of tuatara L2 elements.** The divergence rate of platypus-like tuatara L2 and non platypus-like tuatara L2 was calculated using the Kimura 2-Parameter divergence metric, and adjusted for 'GC' content. Dark blue bars show the sequence substitution rate of non platypus-like tuatara L2 against non platypus-like L2 super consensus, while light blue bars show the sequence substitution rate of platypus-like tuatara L2 against platypus-like L2 super consensus. Bars are paired for each substitution level. Boxplot shows the upper and lower quartiles and the mean value of each tuatara L2 type with respect to their divergence rate.

Un-annotated consensus sequences are probably segmental duplications

Since 211,910 of the *ab initio* consensus sequences were classified as un-annotated repeats (unable to be annotated as transposable elements or gene families) and 33% of the tuatara genome was annotated as un-annotated repeats, we examined the characteristics of these un-annotated consensus sequences. Specifically, we looked at their length distribution and their copy number in the tuatara genome (Figure 6). Figure 6A shows that most of the tuatara contigs are shorter than 5kb, the shortest contig is 880bp and the longest 29,979,683bp. Figure 6 shows that 97% of the un-annotated repeat consensus sequences were shorter than 2kb, and the longest

un-annotated consensus sequence was 12,707bp. In terms of copy number, only 1% of the un-annotated consensus sequences were present in more than 1,000 copies. Only 3 of them were present at more than 10,000 copies (Figure 6D in red box; of the longest un-annotated consensus sequences, two were 1-2kb in length. Because most of the un-annotated consensus sequences are short and present at low copy number, they are most likely the results of segmental duplications.

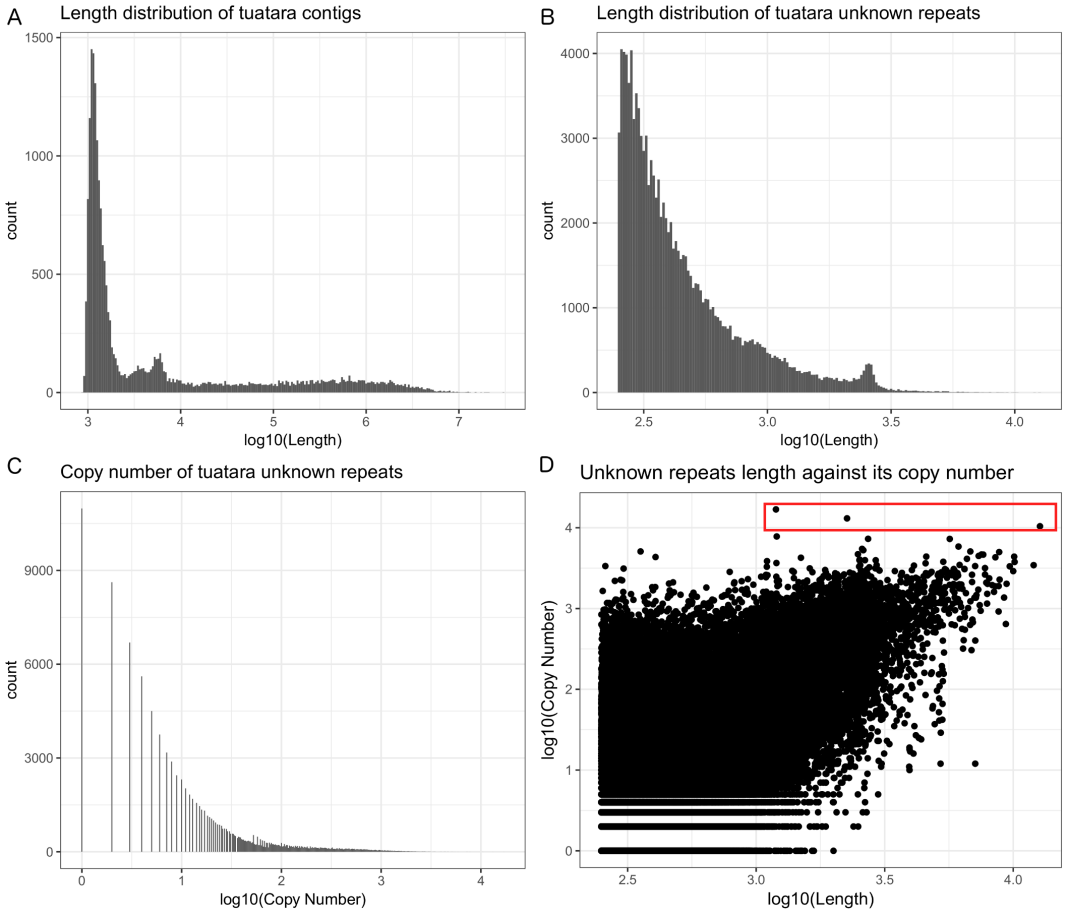


Figure 6: **Characteristics of un-annotated consensus sequences in the tuatara genome.** A) Length distribution of tuatara contigs, transformed length with log10; B) Length distribution of tuatara un-annotated repeat consensus sequences generated from the *ab initio* method, transformed length with log10; C) Copy number distribution of un-annotated repeat consensus sequences, transformed copy number with log10; D) Scatter plot of length and copy number of un-annotated repeat consensus sequences.

In order to further analyse the un-annotated repeats with high copy number in the tuatara

genome (>10,000), a coverage plot was used to visualise the copy number of genomic sequences similar to the high copy consensus (Figure 7). Figure 7 clearly shows that in unclassified repeat family 110313 (1,194bp) and family 141729 (2,257bp), significant copy number enrichment was observed in a 400bp region (>10,000 copy number), while in family 205820 (12,707bp), we observed multiple peak regions with copy numbers less than 1,500.

The consensus sequences for these three families were obtained using single linkage clustering of pairwise alignments, the low base line coverage level across the whole consensus sequences indicates the low copy number of the genomic sequences used to construct the consensus. Therefore, they are more likely segmental duplications. The high copy number peaks indicate the presence of highly repetitive subsequences in the consensus.

CENSOR and BLASTN were used to further characterise the consensus sub-sequences from each peak. Figure 7 contains outputs from both CENSOR and BLASTN that were used to annotate the peaks. In family 110313, CENSOR annotated one end as the $5'$ end of a DNA transposon (DNAX1_ML), and the other end as the $3'$ end of turtle CR1 (CR1-10 CPB). Interestingly, a promoter region was found in the $5'$ end using Promoter2.0 (PolII promoter prediction tool), and the promoter located in the region that annotated as the $5'$ end of a DNA transposon. This subsequence is thus a good candidate for a potential novel SINE dependent on CR1 for retrotransposition (see below). In family 141729, CENSOR also annotated the $5'$ end of the coverage peak as the $5'$ end of DNAX1_ML, but there was no RepBase annotation of the $3'$ end, and no potential promoter. BLASTN annotation showed that both peak regions are weakly similar to tuatara DMRT1 non-coding regions. However, this annotation may result from a bias in the BLAST database, as DMRT1 is the fifth longest tuatara sequence in the NCBI nr database, which only contains 2660 tuatara entries. For family 205820, we observed what seemed to be a random combination of different repeat classes. As a final test to determine if any of the high copy subsequences might be SINEs, we compared all peak subsequences with manually curated tuatara SINEs (provided by Alex Suh, Uppsala University) using BLASTN

(-word_size=9, -outfmt=6), but did not detect any significant similarities.

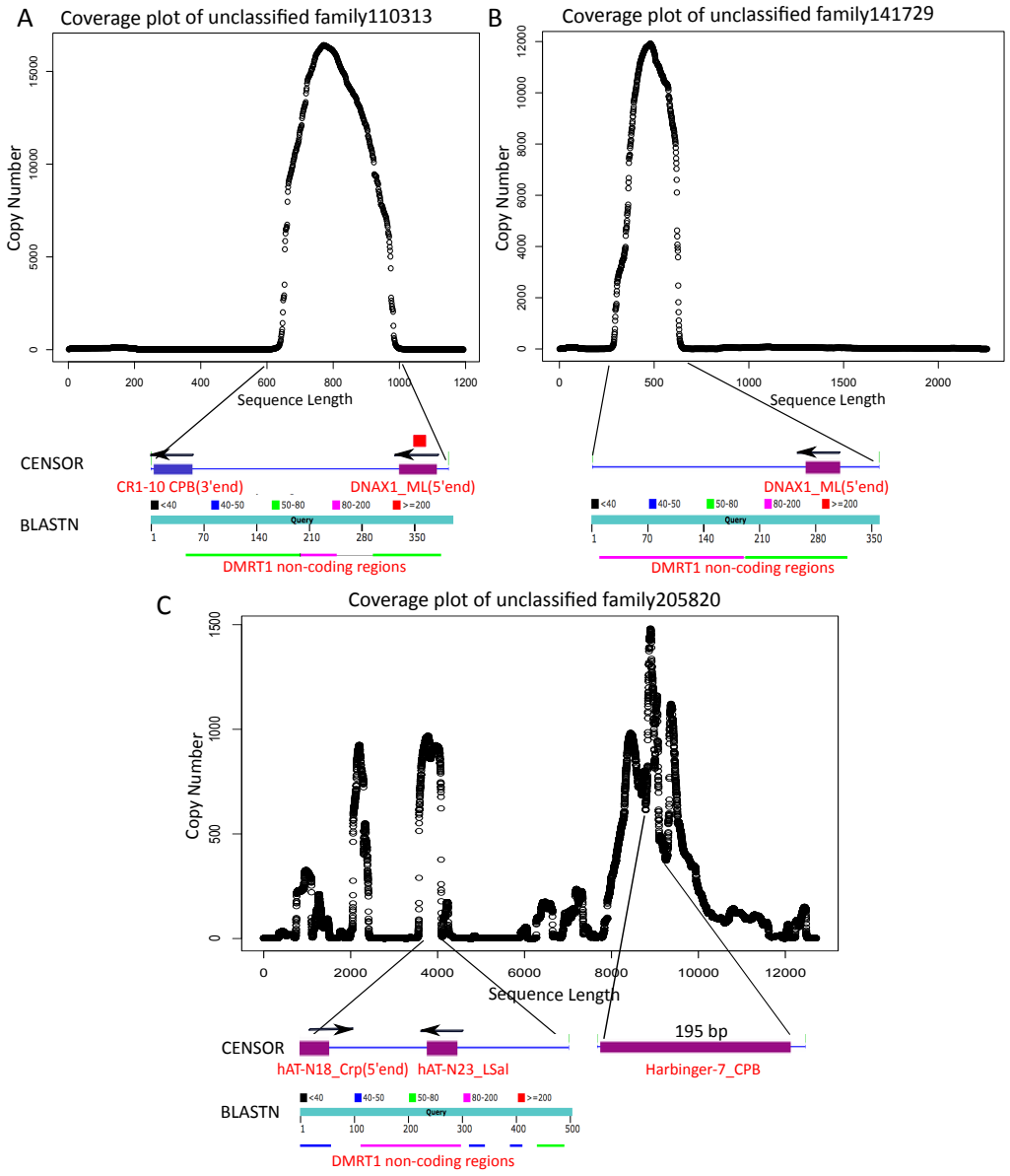


Figure 7: **Coverage plots of the top 3 high copy number un-annotated repeats in the tuatara genome.** A) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 113013, the red box represents the potential promoter mentioned in the text; B) CENSOR and BLASTN annotation of the peak coverage region in unclassified family 141729; C) CENSOR and BLASTN annotation of selected peak coverage regions in unclassified family 205820. Black arrows represent the strand orientation of sequences based on CENSOR annotation.

According to previous research, amniote CR1 subfamilies exhibit an 8-nt microsatellite

motif at their 3′ ends, and a hairpin structure with inverted repeat motifs (IR1 and IR2) (Figure 8A) [18]. We manually analysed the family 110313 to check if the CR1 motif was found in this sub-sequence. Surprisingly, this sub-sequence fulfilled five of the characteristics needed to be a CR1-like repeat (Figure 8B), indicating that this sub-sequence may be a CR1-mobilised SINE. The sequence below the peak in red shows one IR1, two IR2 and a conserved microsatellite region (Figure 8B). Moreover, the peak region in family 110313 and peak region in family 141729 shared high similarity (E-value: 8e-83), except for the fact that family 141729 lacks the 3′ end found in family 110313 (Figure 8C).

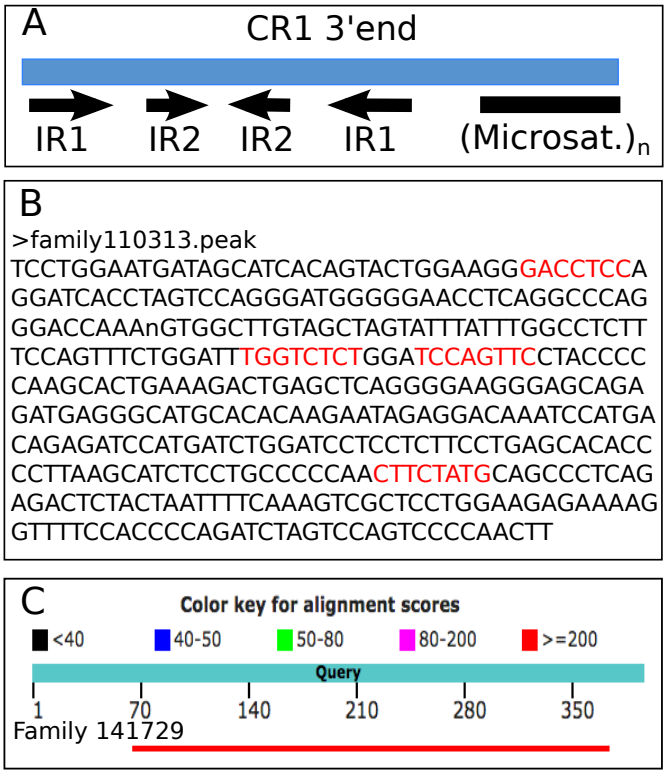


Figure 8: **Example of candidate novel CR1-SINE (family 110313).** A) Structure of conserved CR1 3ends; B) Family 113103 contains IR1, IR2, IR2 and microsatellite (red word in order) of CR1 conserved sequence; C) BLASTN output shows high sequence similarity between family 110313 high copy number sub-sequence and family 141729 high copy number sub-sequence.

We also investigated the sequence similarity of family 113103 and family 141729 compared to tuatara SINE elements (16 sequences generated using Alex Suh's SINEs), Figure 9 shows that

the two unclassified repeats have a similar degree of similarity compared to the 16 tuatara SINE elements, and the tight length distributions combined with high degrees of sequence identity are indicative of recent repeat expansion events. SINEs can expand exponentially, so a tight distribution of sequence identities is expected for these elements, particularly if they have expanded recently. If families 113103 and 141729 were unable to self-amplify ie require an external promoter, their rate of expansion would be linear rather than exponential and we would expect to see a broader distribution of sequence identities for such a scenario. Based on these observations and arguments it appears that families 113103 and 141729 behave in a manner consistent with SINEs and are therefore candidate novel SINE elements.

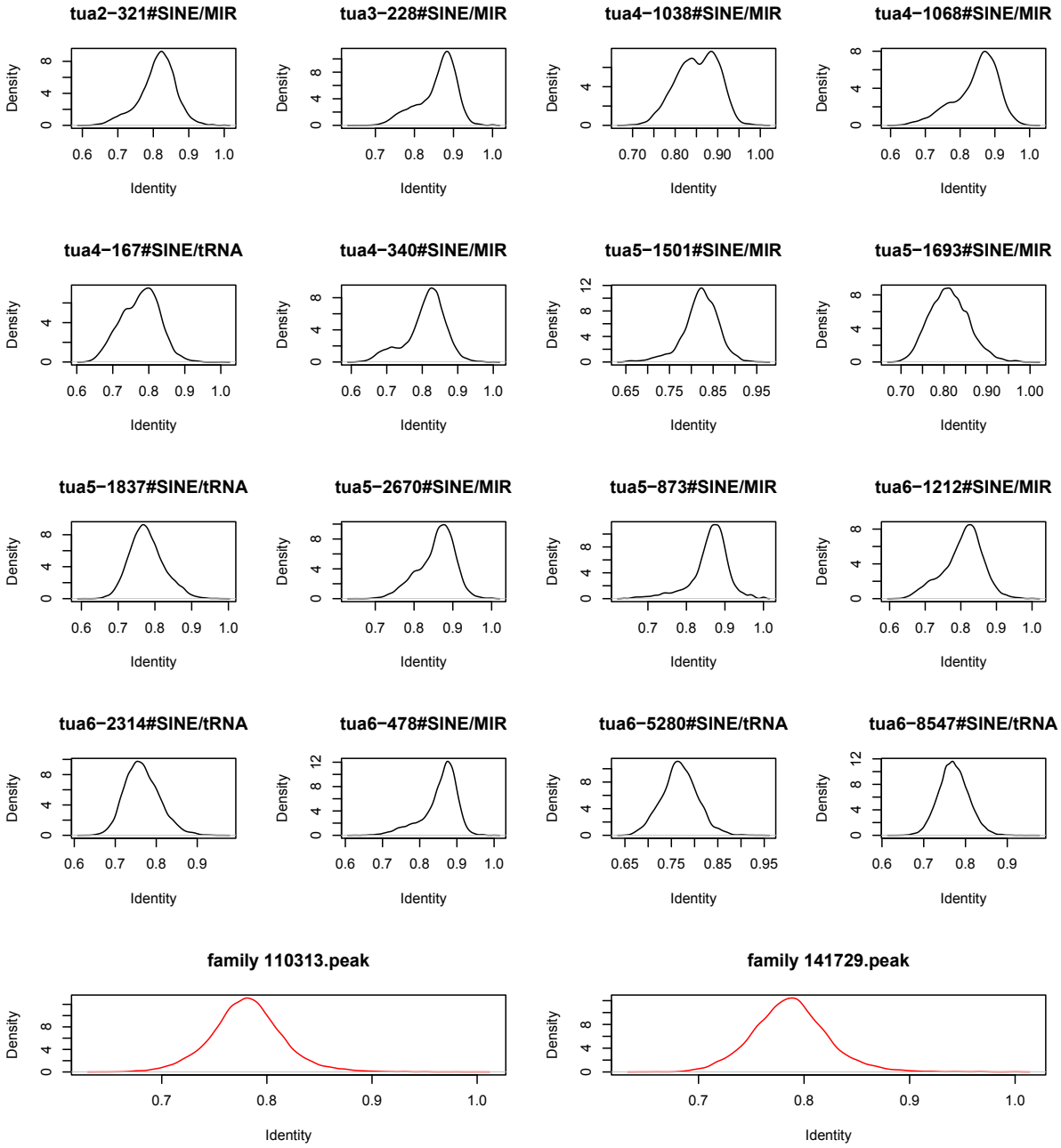


Figure 9: **Divergence of tuatara SINE elements and two unclassified repeats.** Density plot shows the sequence identities from CENSOR output for each consensus sequences. The two bottom plots with red lines show high copy number un-annotated repeats from the *ab initio* method.

We extended the analysis of high copy number un-annotated consensus sequences to a

further 21 un-annotated sequences (Figure 10). The one result of interest was family 186589, which showed high similarity at its 5′ end with family 113103 and family 141729 (E-value:1e-63 and 3e-65) (Figure 11); three of these unknown repeats were partially annotated as DNA transposons at their 5′ ends. However the remaining peak regions showed no significant hits from CENSOR, BLASTN and BLASTX output, and had no shared sequence similarity, which further supports the idea that most of the un-annotated repetitive sequences we found are from segmental duplications.

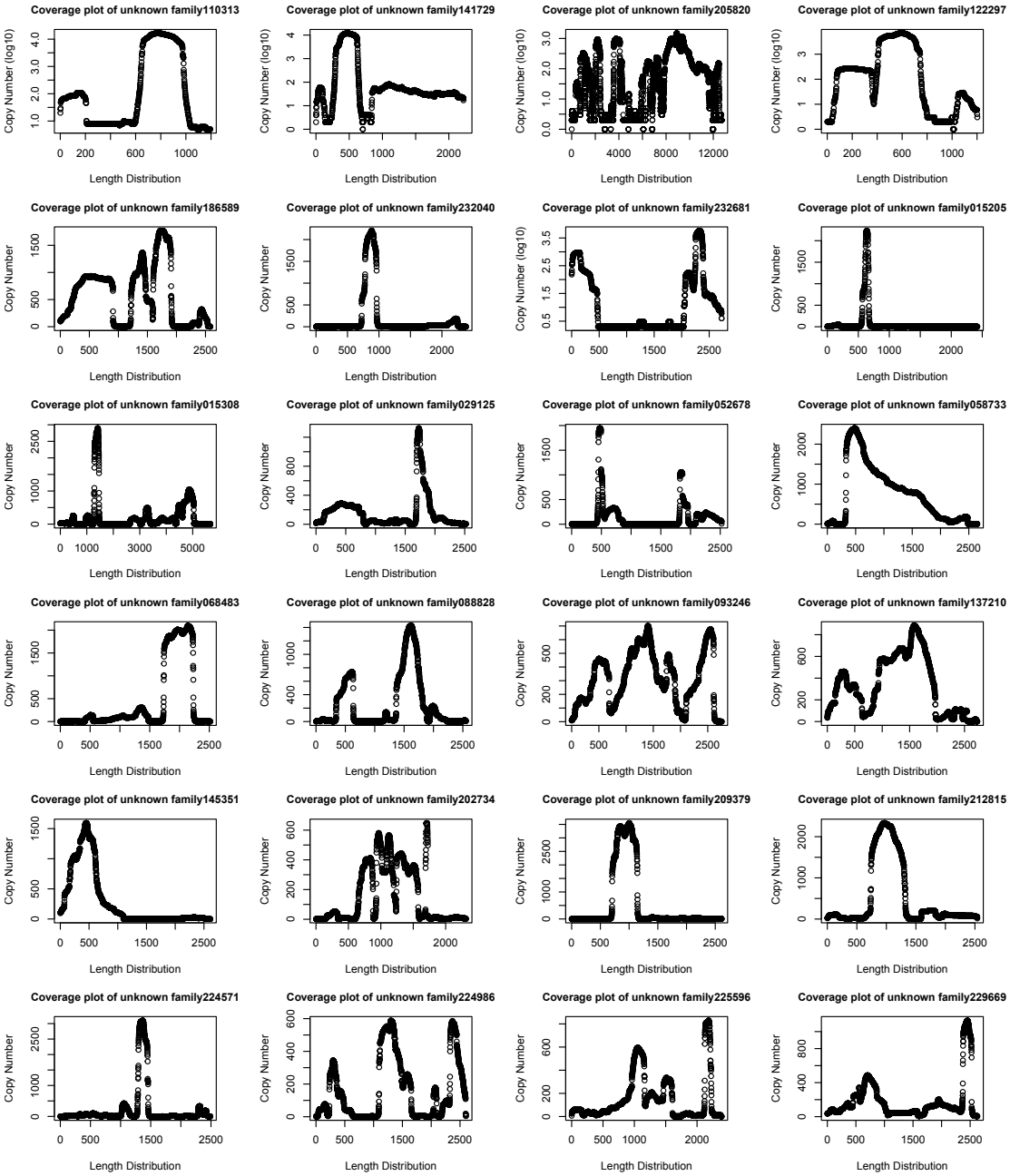


Figure 10: **Coverage plots of high copy number un-annotated repeats in the tuatara genome .** Overall coverage plots of 24 high copy number un-annotated sequences (including the 3 highest un-annotated families).

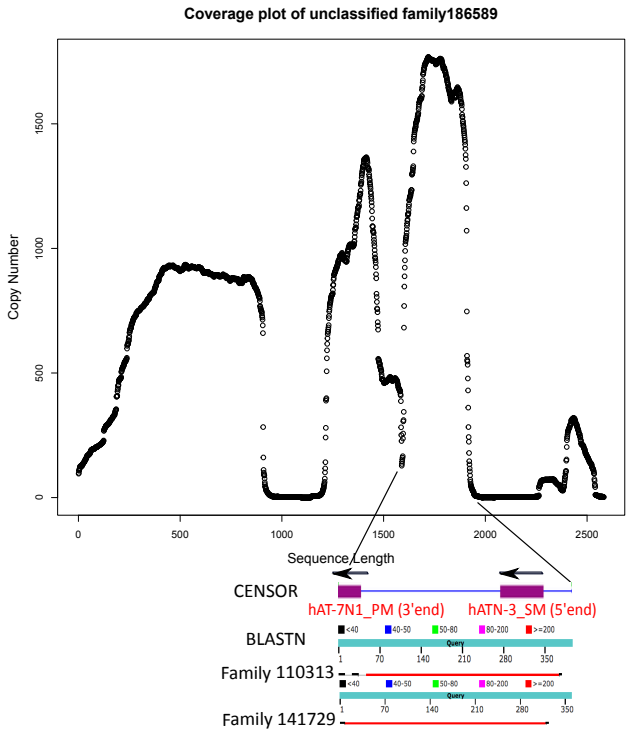


Figure 11: **Coverage plot of un-annotated family 186589 in the tuatara genome.** CENSOR and BLASTN annotation of the peak coverage region in unclassified family 186589. Black arrows represent the strand orientation of sequences based on CENSOR annotation

Discussion

Segmental duplications in the tuatara genome

Compared to RepeatModeler, our method was able to classify more than 200 times as many repetitive consensus sequences, with a broad length distribution, from 250bp to 31,536bp (Table 6). Furthermore, the *ab initio* method allowed us to identify probable segmental duplications, as 99% the un-annotated repeat sequences from our method were present at fewer than 1,000 copies in the tuatara genome (Figure 6). It is worth noting that because our method identifies similar sequences with low divergence that many of the presumptive segmental duplications we found are of relatively recent origin.

Segmental duplications are usually defined as being >1kb long and >90% identical[31]

and have proven to be relevant to disease [32] and integral to studies on genome evolution [33, 34]. According to previous research, segmental duplications account for 2.1% [35], 4.9% [11], 3.7% [36] and 6.5% [4] respectively of the Chinese alligator, green anole lizard, chicken and zebra finch genomes. While we have not characterised segmental duplications using the commonly used parameters that identify recent duplications, we found that 6.7% of the tuatara genome is present as duplicated sequences >250bp and >94% identical. This indicates that tuatara has a greater prevalence of SD compared to other vertebrates. Furthermore, the low copy number un-annotated duplications represented 33.2% of the tuatara genome based on CENSOR analysis, suggesting the presence of many older SD. It is worthwhile noting that a comparable analysis of the anole lizard genome identified a total of 12% of the genome as SD [37]. The presence of so many SD in a vertebrate is unusual and may reflect differing drivers of genome expansion/turnover in tuatara (Figure 6-11).

Significance of LINE retrotransposons in the tuatara genome

The retrotransposon content of reptiles and birds is lower than in mammals, and the 28.2% repeat content of tuatara is similar to previous studies in archosaurs showing that repetitive elements accounted for 23.4%, 27.2% [38], 10% [15], 30.4% [11], 8.5% [36] and 9.8% [4] of genome sequence in alligator, crocodile, western painted turtle, green anole lizard, chicken and zebra finch, respectively. While CR1 elements are the dominant LINEs in these genomes, L2 non-LTR LINEs are the dominant autonomous retrotransposon class in tuatara, accounting for 12% of the genome (Table 1). L2 are regarded as ancient repeats and most individual L2 sequences tend to be non-autonomous and truncated [39]. Less than 3% of the human genome is annotated as L2 (L2 are molecular fossils in eutheria) [40], but in platypus, L2 elements are still active and they account for 19% of the genome [8]. Because our method identifies weakly divergent repetitive elements (newly inserted elements), and because most L2 elements we found in the tuatara genome could not be identified using the RepBase library, the

L2 sequences we found are most likely tuatara specific TE of recent origin and are probably still active (Table 2).

Possible HT event identified by analysis of tuatara and platypus L2

Horizontal transfer describes the non parent-to-offspring transmission of genetic material between individuals: a phenomenon primarily considered in a prokaryotic context. However, with assistance from a suitable vector (e.g. parasites, viruses), retrotransposons have the ability to jump between species as they do within a genome [41]. Relatively few studies have demonstrated HT of retrotransposons, including CR1s, RTEs and L1s [24, 42, 26, 43], with no studies reporting HT for L2s. Our phylogenetic trees demonstrate that a particular class of tuatara L2 elements may have arisen from a transfer of platypus-like L2s into tuatara (Figures 1, 3 and 4), this same class of L2 is absent from other lepidosaurs (Table 3 and Table 4).

According to the geographical distribution of fossils, monotremes are believed to have evolved in a region of Gondwanaland corresponding to Australia and Western New Guinea [44] and diverged from therian mammals about 163 to 186 million years [45]. In contrast, tuatara diverged from Squamata about 220 million years [46] when sphenodons had a very wide geographic distribution [47] before being restricted to New Zealand [48]. The identification of platypus-like L2 elements in tuatara indicates that L2 retrotransposons may have either been transferred between a monotreme and sphenodon or into both from a third, as yet unknown species [23]. The relatively low divergence of the Platypus like L2 in sphenodon suggests that such an HT event must have occurred after geographic separation of sphenodon and monotremes, implicating likely involvement of a third species/vector in this potential HT transfer (Figure 12).

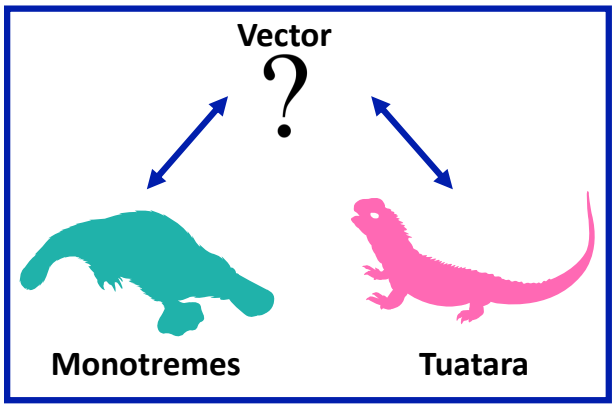


Figure 12: **Possible HT transfer of L2 retrotransposons via a third species/vector.** The L2 elements common to monotremes and tuatara suggest a potential recent HT event, facilitated by a third species.

In conclusion, while tuatara has a transposable element content similar to other reptiles it is dominated by L2 instead of CR1 LINEs and shows evidence of recent HT and expansion of platypus-like L2 elements. It also has a far higher segmental duplication content compared to other vertebrate genomes. The unusual transposable element composition and segmental duplication prevalence in the tuatara genome may indicate that drivers of genome size and complexity in tuatara differ compared to mammals, birds and other reptiles and may be more representative of the genome of the common ancestor for theria and archosaurs.

Materials and Methods

Reference genome: repeat identification, annotation

De novo repeat identification and annotation

The *Sphenodon punctatus* (tuatara) draft genome assembly (Table 5) was used for repeat identification. All tuatara genome scaffolds were pairwise aligned using Krishna (https://github.com/biogo/examples/krishna) [49, 50] with parameters set for 94% sequence identity (-dpid) and a minimum length (-dplen) of 250bp. The resulting alignment intervals were then used as input for igor [49, 50] to define families of repeat sequences using the default parameters. Igor output was used as input for seqer [49, 50] in order to generate repeat

consensus sequences for each cluster/family based on MUSCLE (v3.8.31) alignments [51]. Only family members within 95% of the length of the longest family member were aligned and to avoid consensus sequence expansion due to indels in the global alignment, a maximum of 100 randomly chosen sequences/family were included in the alignment.

Table 5: **Tuatara genome dataset**. Shows the systematic name, common name, genome size, scaffold N50, assembly level, assembly method, sequencing techology and genome coverage.

Systematic Name	Common Name	Total Sequence Length	Scaffold N50	Assembly Level	Assembly Method	Sequencing Technology	Genome Coverage
Sphenodon punctatus	tuatara	4,272,217,537	3,052,611	Scaffold	Allpaths-LG, HiRise	Illumina	44X

Identifiable repeat consensus sequences were annotated using CENSOR [52] with the Repbase 'Vertebrate' library (downloaded on 1st march, 2016, includes 41,908 sequences). Further annotation of consensus sequences was based on WU-BLAST alignment against a comprehensive retroviral and retrotransposon protein database assembled from the National Center for Biotechnology Information [53], and against swissprot to identify known protein-coding genes from large gene families inappropriately included in the repeat set. Consensus sequences identified as either simple sequence repeats (SSRs) or protein-coding sequences, but not similar to retrotransposon or endogenous retrovirus protein-coding sequences, were removed from the consensus set. After annotation of these tuatara repeat consensus sequences, CENSOR was used to map these sequences back to the tuatara genome in combination with the Repbase 'Vertebrate' library. Table 6 shows the comparison of repeat libraries generated using RepeatModeler and the *ab initio* method.

Analysis of L2 elements in the tuatara genome

In order to investigate the similarity of tuatara L2 sequences generated from the *ab initio* method and RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), we extracted long L2

Table 6: Summary of *Sphenodon punctatus* repeat library metrics

	RepeatModeler Library	—*ab initio* Library
No. of Consensuses	1,149	229,311
Total length (MB)	1.1	182
Well-annotated Consensuses	446	17,401
Unknown Consensuses	703	211,910
Min./Max. Length (bp)	151/8,753	250/31,536

consensus sequences (2-4kb) from both libraries. MUSCLE was used to carry out global alignments between the two repeat sequence sets. FastTree (v2.1.8) [54] was used to infer a maximum likelihood phylogeny from the alignment output. FigTree (v1.4.2) was used to visualise and annotate the tree using repeat class labels.

Resolving L2 classification

Many of the tuatara L2 consensus sequences were annotated as being most similar to platypus L2. This introduced an annotation problem, since according to Repbase, almost all platypus L2 consensus sequences were annotated as belonging to the CR1 clade, rather than the L2 clade. In order to resolve the annotation, we needed to determine which clade platypus L2 really belonged to. Full-length consensus sequences of chicken CR1 elements were extracted from Repbase, and MUSCLE was then used to carry out global alignments between CR1 elements and the 2-4kb long L2 consensus sequences from tuatara. Alignment output was used to construct a maximum likelihood phylogeny using FastTree. FigTree was used to visualise and annotate the tree using repeat class labels.

Dendrogram construction from L2 nucleotide sequence alignments

In order to determine the evolutionary position of tuatara L2 within vertebrates, we used the *ab initio* method to identify platypus and bearded dragon. In this fashion we were able to acquire comparable L2 consensus sequences from these two species. L2 consensus

sequences approximately 2-4kb long from platypus, bearded dragon and tuatara identified using the *ab initio* method and tuatara L2 sequences that ranged from 2-4kb long were extracted from the RMD output. Finally, L2 consensus sequences (2-4kb) were extracted from the Repbase 'Vertebrate' library. We globally aligned the resulting 159 consensus sequences using MUSCLE. FastTree was used to infer a maximum likelihood phylogeny from the global alignment. Archaeopteryx v0.9901 beta was used to visualise and annotate the tree using repeat class labels.

Phylogenetic analysis of L2 elements using RT domain sequences

All L2 sequences were extracted from the *ab initio* consensus sequence libraries (platypus, tuatara, bearded dragon), and anole lizard, turtle and crocodile L2 consensus sequences were extracted from the Repbase library. USEARCH [55] was then used to scan for open reading frames in L2 consensus sequences that were at least 60% of the expected length ($>$ 1.5kb for ORF2p). After translation, ORF2p candidates were checked for similarity to known domains using HMM-HMM comparison [56] against the Pfam28.0 database [57] as of May 2015 (includes 16,230 families). ORF2p containing RT domains were extracted using the envelope coordinates from the HMMer domain hits table (–domtblout) [56], with a minimum length of 200 amino acids. Nucleotide sequences that contained RT domains were extracted and assembled into one file (a total of 66 sequences).

Two methods were tested to describe the evolutionary dynamics of potentially active L2 elements. First, 66 RT domain nucleotide sequences within ORF2p were aligned with muscle, then FastTree was used to infer a maximum likelihood phylogeny (-nt, -gtr). Another RT domain Phylogeny was inferred by using MrBayes (lse nst=6) [58] from MUSCLE alignment output. Both of the methods used a GTR model and FigTree was used to visualise and annotate the tree using repeat class labels.

Potential horizontal transfer of L2 elements between tuatara and platypus

CENSOR was used with a custom library of platypus-like and non platypus-like L2 consensus sequences from tuatara to find similar sequences (both full length and fragments) in five reptile genomes (anole lizard, crocodile, alligator and bearded dragon) and one monotreme genome (platypus). To confirm the validity of hits, each hit was extracted as a nucleotide sequence and aligned with BLASTN (default parameter, E-value $<$ 1e-5) against the platypus-like tuatara L2 and non platypus-like tuatara L2 consensus sequences. Hits smaller than 50bp were discarded.

MUSCLE and PILER were used to build super consensus for platypus-like tuatara L2 and non platypus-like tuatara L2. RepeatMasker was used to align each super consensus against corresponding tuatara L2 elements, in order to calculate divergence rate using Kimura 2-parameter divergence metric, adjusted for 'GC' content.

Identification of novel repeat sequences from the tuatara genome

In order to explore the unclassified consensus sequences from the *ab initio* method, 211,910 un-annotated sequences were extracted, and the R package ggplot2 [59] was used to visualise their length distribution against copy number.

For high copy number (10,000 copies) families, a coverage plot was used to investigate the positional distribution of genomic sequence fragments with respect to these un-annotated sequences. BLASTN and CENSOR were further used to characterise the consensus sub-sequences from the coverage peaks of un-annotated sequences found in the coverage plots. We scanned for potential promoters present in coverage peaks using an RNA PolII promoter prediction tool (Promoter 2.0) [60]. However, since the internal promoters found in SINEs are RNA PolIII promoters, this is only a preliminary finding. Fragments of 16 SINE and un-annotated repeats identified from CENSOR output were extracted and sequence identity curves were plotted using R.

The criteria we used to identify novel repeats from un-annotated repeat consensus sequences

were as follows: the sequence must have copy number greater than 10,000; the sequence should be of a reasonable length for a repeat (200bp-6kb); the sequence cannot be identified as a repeat based on similarity to the Repbase library, and it should not be similar to any protein coding sequences (the sequences would then likely to belong to a gene family, or be part of a conserved domain).

References

[1] Tyler A Elliott and T Ryan Gregory. Do larger genomes contain more diverse transposable elements? *BMC evolutionary biology*, 15(1):69, 2015.

[2] Haig H Kazazian. Mobile elements: drivers of genome evolution. *science*, 303(5664):1626–1632, 2004.

[3] International Chicken Genome Sequencing Consortium et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695, 2004.

[4] Wesley C Warren, David F Clayton, Hans Ellegren, Arthur P Arnold, LaDeana W Hillier, Axel Künstner, Steve Searle, Simon White, Albert J Vilella, Susan Fairley, et al. The genome of a songbird. *Nature*, 464(7289):757, 2010.

[5] Guojie Zhang, Cai Li, Qiye Li, Bo Li, Denis M Larkin, Chul Lee, Jay F Storz, Agostinho Antunes, Matthew J Greenwold, Robert W Meredith, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, 346(6215):1311–1320, 2014.

[6] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[7] Tarjei S Mikkelsen, Matthew J Wakefield, Bronwen Aken, Chris T Amemiya, Jean L Chang, Shannon Duke, Manuel Garber, Andrew J Gentles, Leo Goodstadt, Andreas Heger, et al. Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):167, 2007.

[8] Wesley C Warren, LaDeana W Hillier, Jennifer A Marshall Graves, Ewan Birney, Chris P Ponting, Frank Grützner, Katherine Belov, Webb Miller, Laura Clarke, Asif T Chinwalla, et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175, 2008.

[9] Sergej Nowoshilow, Siegfried Schloissnig, Ji-Feng Fei, Andreas Dahl, Andy WC Pang, Martin Pippel, Sylke Winkler, Alex R Hastie, George Young, Juliana G Roscito, et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature*, 554(7690):50, 2018.

[10] Marc Tollis and Stéphane Boissinot. The transposable element profile of the anolis genome: How a lizard can provide insights into the evolution of vertebrate genome size and structure. *Mobile genetic elements*, 1(2):107–111, 2011.

[11] Jessica Alföldi, Federica Di Palma, Manfred Grabherr, Christina Williams, Lesheng Kong, Evan Mauceli, Pamela Russell, Craig B Lowe, Richard E Glor, Jacob D Jaffe, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, 477(7366):587, 2011.

[12] D Kordis. Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenetic and genome research*, 127(2-4):94–111, 2009.

[13] Richard E Green, Edward L Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W Vandewege, John A St John, Salvador Capella-Gutiérrez, Todd A

Castoe, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215):1254449, 2014.

[14] Todd A Castoe, AP Jason De Koning, Kathryn T Hall, Daren C Card, Drew R Schield, Matthew K Fujita, Robert P Ruggiero, Jack F Degner, Juan M Daza, Wanjun Gu, et al. The burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences*, 110(51):20645–20650, 2013.

[15] H Bradley Shaffer, Patrick Minx, Daniel E Warren, Andrew M Shedlock, Robert C Thomson, Nicole Valenzuela, John Abramyan, Chris T Amemiya, Daleen Badenhorst, Kyle K Biggar, et al. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome biology*, 14(3):R28, 2013.

[16] Andrew M Shedlock, Christopher W Botka, Shaying Zhao, Jyoti Shetty, Tingting Zhang, Jun S Liu, Patrick J Deschavanne, and Scott V Edwards. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences*, 104(8):2767–2772, 2007.

[17] Alexander Suh, Gennady Churakov, Meganathan P Ramakodi, Roy N Platt, Jerzy Jurka, Kenji K Kojima, Juan Caballero, Arian F Smit, Kent A Vliet, Federico G Hoffmann, et al. Multiple lineages of ancient cr1 retroposons shaped the early genome evolution of amniotes. *Genome biology and evolution*, 7(1):205–217, 2014.

[18] Alexander Suh. The specific requirements for cr1 retrotransposition explain the scarcity of retrogenes in birds. *Journal of molecular evolution*, 81(1-2):18–20, 2015.

[19] Anthony V Furano. The biological properties and evolutionary dynamics of mammalian line-1 retrotransposons. 2000.

[20] Cushla J Metcalfe and Didier Casane. Modular organization and reticulate evolution of the orf1 of jockey superfamily transposable elements. *Mobile DNA*, 5(1):19, 2014.

[21] Prescott L Deininger, John V Moran, Mark A Batzer, and Haig H Kazazian. Mobile elements and mammalian genome evolution. *Current opinion in genetics & development*, 13(6):651–658, 2003.

[22] Sarah Schaack, Clément Gilbert, and Cédric Feschotte. Promiscuous dna: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in ecology & evolution*, 25(9):537–546, 2010.

[23] Atma M Ivancevic, Ali M Walsh, R Daniel Kortschak, and David L Adelson. Jumping the fine line between species: horizontal transfer of transposable elements in animals catalyses genome evolution. *Bioessays*, 35(12):1071–1082, 2013.

[24] Ali Morton Walsh, R Daniel Kortschak, Michael G Gardner, Terry Bertozzi, and David L Adelson. Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences*, 110(3):1012–1016, 2013.

[25] Alexander Suh, Christopher C Witt, Juliana Menger, Keren R Sadanandan, Lars Podsiadlowski, Michael Gerth, Anne Weigert, Jimmy A McGuire, Joann Mudge, Scott V Edwards, et al. Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. *Nature communications*, 7:11396, 2016.

[26] Irina Sormacheva, Georgiy Smyshlyaev, Vladimir Mayorov, Alexander Blinov, Anton Novikov, and Olga Novikova. Vertical evolution and horizontal transfer of cr1 non-ltr retrotransposons and tc1/mariner dna transposons in lepidoptera species. *Molecular biology and evolution*, 29(12):3685–3702, 2012.

[27] Andrew M Shedlock and Scott V Edwards. Amniotes (amniota). *The timetree of life*, 375:379, 2009.

[28] Peter A Novick, Holly Basta, Mark Floumanhaft, Marcella A McClure, and Stéphane Boissinot. The evolutionary dynamics of autonomous non-ltr retrotransposons in the lizard anolis carolinensis shows more similarity to fish than mammals. *Molecular biology and evolution*, 26(8):1811–1822, 2009.

[29] AFA Smit, R Hubley, and P Green. Repeatmodeler open-1.0. 2008-2010. *Access date Dec*, 2014.

[30] AFA Smit, R Hubley, and P Green. Repeatmasker open-4.0. 2013–2015. *Institute for Systems Biology. http://repeatmasker. org*, 2015.

[31] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research*, 11(6):1005–1017, Jun 2001.

[32] Bernard Conrad and Stylianos E Antonarakis. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu. Rev. Genomics Hum. Genet.*, 8:17–35, 2007.

[33] Andrew J Sharp, Devin P Locke, Sean D McGrath, Ze Cheng, Jeffrey A Bailey, Rhea U Vallente, Lisa M Pertz, Royden A Clark, Stuart Schwartz, Rick Segraves, et al. Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*, 77(1):78–88, 2005.

[34] Sarah A Teichmann and M Madan Babu. Gene regulatory network growth by duplication. *Nature genetics*, 36(5):492–496, 2004.

[35] Qiu-Hong Wan, Sheng-Kai Pan, Li Hu, Ying Zhu, Peng-Wei Xu, Jin-Quan Xia, Hui Chen, Gen-Yun He, Jing He, Xiao-Wei Ni, et al. Genome analysis and signature discovery for diving and sensory properties of the endangered chinese alligator. *Cell research*, 23(9):1091–1105, 2013.

[36] LaDeana W Hillier, Webb Miller, Ewan Birney, Wesley Warren, Ross C Hardison, Chris P Ponting, Peer Bork, David W Burt, Martien AM Groenen, Mary E Delany, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, 2004.

[37] Lu Zeng, R Daniel Kortschak, Joy M Raison, Terry Bertozzi, and David L Adelson. Superior ab initio identification, annotation and characterisation of tes and segmental duplications from genome assemblies. *bioRxiv*, page 190694, 2017.

[38] John A St John, Edward L Braun, Sally R Isberg, Lee G Miles, Amanda Y Chong, Jaime Gongora, Pauline Dalzell, Christopher Moran, Bertrand Bed'Hom, Arkhat Abzhanov, et al. Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome biology*, 13(1):415, 2012.

[39] Nika Lovšin, Franc Gubenšek, and Dušan Kordi. Evolutionary dynamics in a novel l2 clade of non-ltr retrotransposons in deuterostomia. *Molecular biology and evolution*, 18(12):2213–2224, 2001.

[40] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[41] Atma M. Ivancevic, Ali M. Walsh, R. Daniel Kortschak, and David L. Adelson. Jumping the fine line between species: horizontal transfer of transposable elements in animals catalyses genome evolution. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 35(12):1071–1082, Dec 2013.

[42] Vera Župunski, Franc Gubenšek, and Dušan Kordis. Evolutionary dynamics and evolutionary history in the rte clade of non-ltr retrotransposons. *Molecular Biology and Evolution*, 18(10):1849–1863, 2001.

[43] Atma Ivancevic, Daniel Kortschak, Terry Bertozzi, and David Adelson. Re-evaluating inheritance in genome evolution: widespread transfer of lines between species. *bioRxiv*, 2017.

[44] Kristofer M Helgen, Roberto Portela Miguez, James Kohen, and Lauren Helgen. Twentieth century occurrence of the long-beaked echidna zaglossus bruijnii in the kimberley region of australia. *ZooKeys*, 255:103, 2012.

[45] Michael Messer, Anthony S Weiss, Denis C Shaw, and Michael Westerman. Evolution of the monotremes: phylogenetic relationship to marsupials and eutherians, and estimation of divergence dates based on α-lactalbumin amino acid sequences. *Journal of Mammalian Evolution*, 5(1):95–105, 1998.

[46] Robert L Carroll. *Vertebrate paleontology and evolution*. Freeman, 1988.

[47] Marc EH Jones, Alan JD Tennyson, Jennifer P Worthy, Susan E Evans, and Trevor H Worthy. A sphenodontine (rhynchocephalia) from the miocene of new zealand and palaeobiogeography of the tuatara (sphenodon). *Proceedings of the Royal Society of London B: Biological Sciences*, pages rspb–2008, 2009.

[48] Ian F Spellerberg and John WD Sawyer. *An introduction to applied biogeography*. Cambridge University Press, 1999.

[49] Robert C Edgar and Eugene W Myers. Piler: identification and classification of genomic repeats. *Bioinformatics*, 21(suppl 1):i152–i158, 2005.

[50] R Daniel Kortschak and David L Adelson. bíogo: a simple high-performance bioinformatics toolkit for the go language. *bioRxiv*, page 005033, 2014.

[51] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[52] Oleksiy Kohany, Andrew J Gentles, Lukasz Hankus, and Jerzy Jurka. Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor. *BMC bioinformatics*, 7(1):1, 2006.

[53] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl 1):D5–D12, 2007.

[54] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

[55] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[56] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, page gkr367, 2011.

[57] Marco Punta, Penny C Coggill, Ruth Y Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, et al. The pfam protein families database. *Nucleic acids research*, page gkr1065, 2011.

[58] Fredrik Ronquist and John P Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

[59] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

[60] Steen Knudsen. Promoter2. 0: for the recognition of polii promoter sequences. *Bioinformatics*, 15(5):356–361, 1999.

Chapter 4

The impact of transposable elements on gene expression during the evolution of amniotes

"Biology is the science, evolution is the concept that makes biology unique." — Jared Diamond

Transposable elements (TEs) have commonly been viewed as "selfish parasites", whose persistence in the genome is best explained by their success as replicating units, rather than any benefit they might bestow on the host. However, recently findings have found that TEs can affect nearby gene activity, either directly by disrupting regulatory sequences or indirectly through the host mechanisms used to prevent TE proliferation. In particular, they seem to play a role in transcriptional regulation by providing genes with promoters and enhancers in the human genome. In this chapter, I have used four specific TEs in six genomes to examine what extent their insertions have contributed to the gene expression during the evolution of amniotes. The manuscript has been submitted to *Mobile DNA*, formatted according to the guidelines of a BMC research article.

Statement of Authorship

Title of Paper	The impact of transposable elements on gene expression during the evolution of amniotes
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Lu Zeng, Stephen M. Pederson, R. Danial Kortschak, David L. Adelson (2018). The impact of transposable elements on gene expression during the evolution of amniotes. Prepared for submission as a research article to Mobile DNA.

Principal Author

Name of Principal Author (Candidate)	Lu Zeng
Contribution to the Paper	Performed analysis, interpreted the results and wrote the manuscript.
Overall percentage (%)	85%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 26/02/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Stephen M. Pederson
Contribution to the Paper	Assisted with development of data analysis techniques and edited manuscript
Signature	Date 26/2/18

Name of Co-Author	R. Daniel Kortshak
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.
Signature	Date 26/2/2018

Please cut and paste additional co-author panels here as required.

Name of Co-Author	David L. Adelson
Contribution to the Paper	Supervised the development of work, assisted with analysis of data and assisted in writing the manuscript.
Signature	Date 26/2/18

RESEARCH

Transposable elements and gene expression during the evolution of amniotes

Lu Zeng[1], Stephen M. Pederson[2], R.Daniel Kortschak[1] and David L. Adelson[1*]

*Correspondence:
david.adelson@adelaide.edu.au
[1]School of Biological Sciences, The University of Adelaide, North Terrace, 5005 Adelaide, Australia
Full list of author information is available at the end of the article

Abstract

Background: Transposable elements (TEs) are primarily responsible for the changes in genome sequences that occur over time within and between species. TEs themselves evolve, with clade specific LTR/ERV, LINEs and SINEs responsible for the bulk of species specific genomic features. Because TEs can contain regulatory motifs, they can be exapted as regulators of gene expression. While TE insertions can provide evolutionary novelty for the regulation of gene expression, their overall impact on the evolution of gene expression is unclear. Previous investigators have shown that tissue specific gene expression in amniotes is more similar across species than within species, supporting the existence of conserved developmental gene regulation. In order to understand how species specific TE insertions might affect the evolution/conservation of gene expression, we have looked at the association of gene expression in six tissues with TE insertions in six representative amniote genomes (human, opossum, platypus, anole lizard, bearded dragon and chicken).

Results: We have used a novel bootstrapping approach to minimise the conflation of effects of repeat types on gene expression. We compared the expression of orthologs containing different types of recent TE insertions to orthologs that contained older TE insertions and found significant differences in gene expression associated with TE insertions. Likewise, we compared the expression of non-ortholog genes containing different types of recent TE insertions to non-orthologs with older TE insertions and found significant differences in gene expression associated with TE insertions. As expected TEs were associated with species-specific changes in gene expression, but the magnitude and direction of change of expression changes were unexpected. Overall, orthologs containing clade specific TEs were associated with lower gene expression, while in non-orthologs, non clade-specific TEs were associated with higher gene expression. Exceptions were SINE elements in human and chicken, which had an opposite association with gene expression compared to other species.

Conclusions: Our observed species-specific associations of TEs with gene expression support a role for TEs in speciation/response to selection by species. TEs do not exhibit consistent associations with gene expression and observed associations can vary depending on the age of TE insertions. Based on these observations, it would be prudent to refrain from extrapolating these and previously reported associations to distantly related species.

Keywords: Transposon; Gene expression; Amniotes; Evolution; Retrotransposon

List of Abbreviations

TE(s): Transposable element(s); **SINE(s):** Short INterspersed Element(s); **polIII:** polymerase III; **LINE(s):** Long INterspersed Element(s); **ORFs:** Open Reading Frames; **RT:** Reverse Transcriptase; **EN:** Endonuclease; **LTR:** Long terminal repeat; **ERV:** Endogenous retroviruses; **TIRs:** Terminal Inverted Repeats; **ssTE(s):** species-specific TE(s); **nsTE(s):** non-species specific TE(s); ∅**TE**: genes containing no TEs; **TPM:** Transcripts Per Million; **UPGMA:** Unweighted Pair Group Method with Arithmetic Mean; **ward.D2:** Ward's minimum variance; ***vs*:** versus.

Introduction

Transposable Elements (TEs) have been shown to alter gene regulation and drive genome evolution [1] [2] [3] [4]. TEs can exert these effects on genes by altering chromatin structure, providing novel promoters or insulators, novel splice sites or other post-transcriptional modifications to re-wire transcriptional networks important in development and reproduction [2] [5]. TEs that land in introns can become "exonized" or spliced into mRNA of the gene into which they have inserted, often introducing stop codons into mRNA that can lead to nonsense-mediated mRNA decay, serving to control gene expression [6] [7].

Short INterspersed Elements (SINEs) are non-autonomous TEs ancestrally related to functionally important RNAs, such as tRNA, 5S rRNA and 7SL RNA that replicate by retrotransposition. SINEs possess an internal promoter that can be recognized and transcribed by the RNA polymerase III (polIII) enzyme complex, and are usually present in a monomeric or tandem dimeric structure [8]. Monomeric tRNA-related SINE families are present in the genomes of species from all major eukaryotic lineages and this structure is, by far, the most frequent. These elements are composed of a 5' tRNA-related region and a central region of unknown origin, followed by a stretch of homopolymeric adenosine residues or other simple repeats [9] [10]. In contrast to the very widespread phylogenetic distribution of tRNA derived SINEs, 7SL-derived SINEs have been found only in mammals [8]. They are composed of a 7SL-derived

region followed by a poly(A) tail and can be either monomeric (B1 family) or dimeric (Alu family) [11] [12]. 5S rRNA-derived SINEs were found in fishes (SINE3) but were likely active in the common ancestor of vertebrates [13] [14]. They are with a 5S-related region (instead of a tRNA-related region), followed by a central region of unknown origin and 3′-terminal repeats [13]. SINE RNAs have also been shown to possess the potential to regulate gene expression at the post-transcriptional level, for example, Alu RNAs an modulate protein translation, influence on RNA editing and mRNA splicing [15].

Long INterspersed Elements (LINEs) are autonomously replicating TEs that replicate through an RNA intermediate that is reverse transcribed back into the genome at a new location. LINEs contain an internal DNA Polymerase II promoter and either one or two Open Reading Frames (ORFs) that contain a Reverse Transcriptase (RT) domain and an Endonuclease (EN) domain. L1 family repeats show a stronger negative correlation with expression levels than the gene length [16], and the presence of L1 sequences within genes can lower transcriptional activity [17].

Long terminal repeat (LTR) retrotransposons are a group of TE, that are flanked by long terminal repeats and contain two ORFs: *gag* and *pol*. The *gag* ORF encodes the structural protein that makes up a virus-like particle [18]. The pol ORF encodes an enzyme needed for replication that contains protease, integrase, reverse transcriptase, and RNase H domains required for reverse transcription and integration. LTRs can also act as alternative promoters to provide new tissue-specificity, act as the major promoters, or exert only minor effects [19]. Many endogenous retroviruses (ERV) contain sequences that can serve as transcriptional start sites or as cis-acting regulatory elements in the host genomes [20].

DNA transposons encode a transposase gene that is flanked by two Terminal Inverted Repeats (TIRs) [21]. The transposase recognizes these TIRs to excise the transposon DNA, which is then inserted into a new genomic location by cut and paste mobilization[22]. DNA transposons can

inactivate or alter the expression of genes by insertion within introns, exons or regulatory region [1] [21].

There is a growing realization that many TEs are highly conserved among distantly related taxonomic groups, suggesting their biological value to the genome. In this report, we describe the association of clade specific TEs with gene expression in long diverged amniotes (Figure 1A) in order to determine how much these TEs might have altered the regulation of gene expression in six tissues during the evolution of these species.

Methods

Expression data

RNA-seq expression data were available for six species (Table 1), belonging to the five main amniote lineages (eutherian: human; marsupial: gray short-tailed opossum; monotreme: platypus; lepidosaur: green anole lizard, bearded dragon; archosaur: chicken) from four somatic (brain, heart, liver, kidney) and two reproductive tissues (testis, ovary)(Gene Expression Omnibus accession numbers GSE30352 [23] and GSE97367 [24], BioProject number PRJEB5206 [25]).

Trim_galore (v0.4.2)(–clip_R1 5; –three_prime_clip_R1 5) [26] was used for adapter trimming and quality control. Adapter-trimmed RNA-seq reads were aligned to the reference genomes (Ensembl release 74) with RSEM (v1.3.0) [27] using Bowtie2 (v2.2.9) [28] with default parameters as the alignment tool. Gene expression was estimated as TPM (Transcripts Per Million). A complete list of accessions can be found in Table S1, Additional file 1.

Genomic data

For chicken, anole lizard, platypus, opossum and human, gene annotations were download from Ensembl release 74. For bearded dragon, RefSeq assembly GCF_900067755.1 was used for analysis. Complete information on genomes used can be found in Table S2, Additional file 1.

Ortholog definition

Gene orthologies were downloaded from Ensembl release 74. Amniote orthologs were defined as single-copy orthologous genes conserved in all 6 amniote species. Reciprocal best hits were used to extract orthologous genes between bearded dragon and other five species by using BLASTN [29]. A total number of 6,595 orthologous genes were extracted from the six species.

TE annotation

TEs were annotated by using CARP: a *ab initio* method [30]. Recently inserted, low divergence TE referred to hereafter as species-specific TE (ssTE) were defined as having ≥94% sequence identity. They were extracted from CARP output, which identifies and annotates TEs that have ≥94% sequence identity. Older TEs were defined as the remaining TE insertions in the genome and are referred to as non-species specific TE (nsTE).

The weighted bootstrap procedure for assessing association of gene expression and TEs

Many genes contain multiple transposable elements, with only a minority of genes containing a single TE. In order to assess any effects on transcription due to the presence of a single TE, a weighted bootstrap approach was devised. For a given TE type within each individual gene, the frequencies of co-occurring TE types and combinations of TE types were noted. Uniform sampling probabilities were then used for the set of genes containing a specific TE type (test sample), whilst sampling weights were assigned to genes lacking the specific TE type based on TE composition (reference sample) (See detail in Table S3-6, Additional file 1). Gene length was divided into 10 bins and these were included as an additional category when defining sampling weights. This ensured that two gene sets were obtained for each bootstrap iteration, which were matched in length and TE composition with the sole difference being the presence of the specific TE type. The median difference in expression level, as measured by log2(TPM), and the difference in the proportions of genes detected as expressed were then used as the variables of interest in the bootstrap procedure. The bootstrap was performed on sets of 1,000 genes

(except for ortholog genes containing non-recent species specific SINE elements in platypus) for 5,000 iterations. Samples that could not meet the minimum number of 600 genes were not used. When comparing expression levels, genes with zero read counts were omitted prior to bootstrapping. In order to compensate for multiple testing considerations, confidence intervals were obtained across the m=nTissues*nElements tests at the level 1-α/m, which is equivalent to the Bonferroni correction, giving confidence intervals that controlled the family-wise error rate at the level α=0.05. Approximate two-sided p-values were also calculated by finding the point at which each confidence interval crossed zero, and additional significance was determined by estimating the FDR on these sets of p-values using the Benjamini-Hochberg method.

Results

Mammalian gene expression phylogenies

To obtain an initial overview of gene expression patterns, we evaluated the similarity of ortholog gene expression in 6 tissues (heart, brain, kidney, liver, testis and ovary), from both males and females in our 6 species. These RNA-seq samples were assembled from three different studies (Table 1, further detail can be seen in Additional file Table S1) [24] [23] [25].

Two hierarchical clustering methods were used to investigate the conservation of expression signatures in these six species within six tissues. 1 - Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and 2 - Ward's minimum variance (ward.D2) hierarchical clustering.

While mostly similar, the two methods did give slightly different clustering results (Figure 2). Generally, gene expression clustered according to tissue with three exceptions. The first exception was bearded dragon heart expression clustered using Ward's method, where heart samples clustered with kidney and liver samples. The second exception was for platypus testis expression clustered using UPGMA, where testis expression clustered with ovary. The third exception was more widespread, and found with both clustering methods; kidney and liver samples only clustered by tissue for human and opossum and were found together more often in species-specific clusters for the other species.

Comparison of gene expression for genes on the basis of their TE content

There were two aspects of the data that affected our analysis. First, because the vast majority of genes contain TEs, it was impossible to compare expression of genes with TEs against genes without TEs, as there were too few of the latter. So we designed our comparisons as shown in Figure 3. Second, most genes contain multiple TE types. In order to minimize the conflation of co-occuring TEs, a weighted bootstrap approach was used in this study. The idea is simple, if we want to investigate the association between a SINE insertion and gene expression, first we randomly select 1,000 genes that contain a SINE element, and then compare their expression level to 1,000 randomly selected genes that do not contain any SINEs. We repeat this process 5,000 times in order to generate enough observations for statistical analysis.

Ortholog expression is associated with with TE type

For our specific analyses, BedTools was used to get the intersection between TE types and 6,595 orthologous genes (including 1kb upstream and 1kb downstream regions) within our six species (chicken, anole lizard, bearded dragon, platypus, opossum and human). The boostrap approach as described above was then applied to this data in order to investigate the association between orthologous gene expression and TE insertions. TEs were split into two groups: recently inserted, low divergence TEs, referred to as species-specific TEs (ssTEs, see methods for detail) and more divergent TEs, referred to as non-species specific TEs (nsTEs). Genes containing no TEs are referred to as ∅TE. The two TE groups were further broken down into four TE classes: DNA transposon, ERV/LTR, LINE or SINE.

Because purifying selection is likely to be more common on orthologs, and since tissue specificity of ortholog expression was largely conserved (Figure 2), we looked first at the association ortholog expression with TE insertions. We compared expression for orthologs containing ssTE against orthologs containing nsTE + ∅TE and expression of orthologs containing nsTE against orthologs containing ssTE + ∅TE (Figure 4) and (Figures S1 and S2, Additional file 1). We found that ssTEs (ERV/LTR, LINE and SINE) were associated with lower gene expression in

orthologs, especially in anole lizard, bearded dragon and human. The exceptions to this negative association were in the human and chicken genome, where recent insertions of SINEs were found associated with higher gene expression in testis and brain.

For orthologs containing nsTE (LINE or SINE) (Figure 4, additional figure S1 and S3) we observed primarily a positive association with gene expression in contrast to the trend seen with ssTEs. The exceptions to this positive association were again found in the human and chicken genomes. Particularly in the chicken genome, where the insertion of non-species specific SINEs were associated with lower ortholog gene expression in multiple tissues.

Overall, species specific TE insertions in orthologs were mainly associated with lower gene expression, while non-species specific TE insertions were mainly associated with higher gene expression. This is true for ERV/LTR in anole lizard, bearded dragon and human, LINE and tRNA derived SINE insertions in anole, bearded dragon, platypus and human. There are some exceptions, notably for chicken orthologs with nsTE insertions which showed an association with decreased gene expression. Perhaps the most interesting observation was that the magnitude of the effect on gene expression was quite pronounced, ranging between about -30% to +40% changes in median gene expression values (Table S7, Additional file 1).

Non-ortholog gene expression is associated with TE type

In order to explore the association of TEs in a more general context, we then expanded our analysis from orthologous genes to non-orthologous genes.

As described above BedTools was used to get the intersection between TE types and non-orthologous genes, and the bootstrap approach was used to compare expression for non-orthologs containing ssTE against non-orthologs containing nsTE + ∅TE and expression of non-orthologs containing nsTE against non-orthologs containing ssTE + ∅TE (Figure 4) and (Figures S4 and S5, Additional file 1).

Similar to orthologs, ssTE insertions in non-orthologs showed a negative association with gene expression. This can be observed in ERV/LTR, LINE and SINE in anole lizard and bearded

dragon. In the chicken, older SINE insertions in non-orthologs were negatively associated with gene expression. In contrast to the anole lizard and bearded dragon, where recent ERV/LTR, LINE and SINE insertions were associated with lower gene expression, human (7SL derived) SINE insertions in non-orthologs were strongly associated with higher gene expression. The magnitude of the association of TEs with gene expression was even more pronounced in these comparisons, ranging from about -40% to +2.8x (Figure 4) and (Figures S4 and S6, Table S7, Additional file 1).

Discussion

Tissue *vs* species clustering of ortholog gene expression had previously been reported using PCA based analysis and used to support the notion that conservation of developmental gene expression programs results in tissue specific gene expression clustering [23] [31] [32]. These results have been reported for single experiments. We did not see quite as compelling tissue clustering of gene expression using PCA on data from aggregated experiments (Figure S7, Additional file 1). However we did see largely similar results when we applied hierarchical clustering methods across the aggregated data (Figure 2). However, in contrast with previous studies, we found liver and kidney gene expression clustered more by species. We attribute this to species specific metabolic adaptations responding to more pronounced environmental selection. We therefore expected to see species specific TE insertions associated with species specific changes in gene expression. For recent species specfic SINE, ERV/LTR and LINE insertions this is precisely what we found. However, we found no tissue specific patterns of association of gene expression with TEs.

We expected species specific TE insertions to be associated with changed gene expression, as they would both alter the spacing of pre-existing regulatory motifs and potentially contribute new regulatory motifs [5] [33] [34]. Because random changes in complex systems usually break things, we expected recent TE insertions to be associated with lower gene expression. While this expectation was largely met, there were some significant exceptions, such as human SINE,

which were associated with increased gene expression (see discussion below). Conversely, it has been shown that older TE insertions contribute to re-wiring of transcriptional networks [35] [36] and thus would have had time to be exapted as enhancers and might be associated with increased gene expression. Previous studies have found that differential decay of ancestral TE sequences across species may result in species-specific transcription factor binding sites [37]. This expectation was also met for human ERV/LTR. However to our surprise, older TE SINE insertions in the chicken were associated with decreased gene expression.

We expected the magnitude of changes in gene expression associated with TE insertions to be modest, however our analysis showed that TE insertions were associated with large changes in gene expression. Based on the median value of changed gene expression from our bootstrap analysis, most statistically significant log2 transformed changes in gene expression associated with TE were smaller than -0.5 and many were greater than 1.0, indicating a range of -40% to +100% change in median gene expression.

Species-specific TE, behaved differently depending on insertion age and species. The most striking example of this was seen in human with recent SINE insertions associated with increased gene expression and older SINE associated with decreased gene expression. This is consistent with observations that Alu elements have been exapted as transcription factor binding sites, and highly and broadly expressed housekeeping genes are enriched for Alus [38] [39] [40]. This was in contrast to an opposite relationship with LINE insertion age and expression change in human, but consistent with previously reported accumulation differences for SINE and LINE insertions in mammalian regulatory regions/open chromatin [41]. Furthermore, LINEs behave similarly in reptiles and human, with new LINEs associated with lower gene expression and older LINEs associated with higher gene expression. This suggests similar constraints on accumulation of TE in lizards and mammals. Finally, TEs had the fewest associations with gene expression in opossum and platypus. This might indicate that these two species are better at repressing TE activity than human, lizards and chicken.

Conclusions

The large changes in gene expression associated with TEs, and the species specific associations of TEs with gene expression support a role for TEs in speciation/response to selection by species. TE types do not exhibit consistent associations with gene expression and observed associations can vary depending on the age of TE insertions. Based on these observations, it would be prudent to refrain from extrapolating these and previously reported associations to distantly related species.

Acknowledgements

We would like to thank Terry Bertozzi and Catisha Coburn for taking the time to read the manuscript in full and offer helpful comments. This paper would not be possible without the helpful discussion with Zhipeng Qu, and extraodinary IT support from Matt Westlake.

Funding

Availability of data and materials

All data analyzed in this study are accessed from the public source NCBI (`https://www.ncbi.nlm.nih.gov/`) and detail can be seen in supplementary file. Repeat annotations and TPM values for each species will be provided to academic researchers upon request.

Author's contributions

L. Z., S.M.P, R.D.K, and D.L.A. designed research; L.Z. and S.M.P. performed research; and L.Z., R.D.K., and D.L.A wrote the paper.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Biological Sciences, The University of Adelaide, North Terrace, 5005 Adelaide, Australia. [2]Bioinformatics Hub, The University of Adelaide, North Terrace, 5005 Adelaide, Australia.

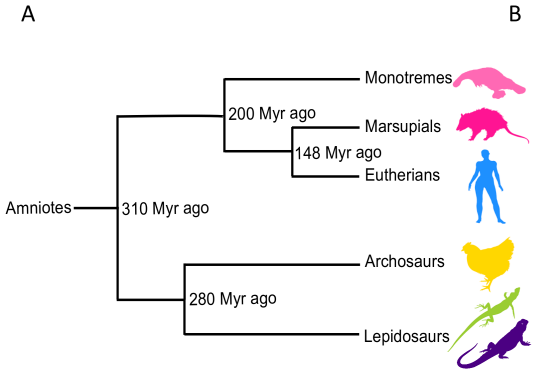
References

1. Kazazian, H.H.: Mobile elements: drivers of genome evolution. science **303**(5664), 1626–1632 (2004)
2. Cordaux, R., Batzer, M.A.: The impact of retrotransposons on human genome evolution. Nature Reviews Genetics **10**(10), 691 (2009)
3. Ponicsan, S.L., Kugel, J.F., Goodrich, J.A.: Genomic gems: Sine rnas regulate mrna production. Current opinion in genetics & development **20**(2), 149–155 (2010)

4. Buckley, R.M., Adelson, D.L.: Mammalian genome evolution as a result of epigenetic regulation of transposable elements. Biomolecular concepts **5**(3), 183–194 (2014)
5. Feschotte, C.: Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics **9**(5), 397 (2008)
6. Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., Ares, M.: Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes & development **21**(6), 708–718 (2007)
7. Attig, J., de los Mozos, I.R., Haberman, N., Wang, Z., Emmett, W., Zarnack, K., König, J., Ule, J.: Splicing repression allows the gradual emergence of new alu-exons in primate evolution. elife **5** (2016)
8. Pelissier, T., Bousquet-Antonelli, C., Lavie, L., Deragon, J.-M.: Synthesis and processing of trna-related sine transcripts in arabidopsis thaliana. Nucleic acids research **32**(13), 3957–3966 (2004)
9. Okada, N.: Sines. Current opinion in genetics & development **1**(4), 498–504 (1991)
10. Okada, N., Hamada, M., Ogiwara, I., Ohshima, K.: Sines and lines share common 3′ sequences: a review. Gene **205**(1), 229–243 (1997)
11. Labuda, D., Sinnett, D., Richer, C., Deragon, J.-M., Striker, G.: Evolution of mouse b1 repeats: 7sl rna folding pattern conserved. Journal of molecular evolution **32**(5), 405–414 (1991)
12. Sinnett, D., Richer, C., Deragon, J.-M., Labuda, D.: Alu rna secondary structure consists of two independent 7 sl rna-like folding units. Journal of Biological Chemistry **266**(14), 8675–8678 (1991)
13. Kapitonov, V.V., Jurka, J.: A novel class of sine elements derived from 5s rrna. Molecular biology and evolution **20**(5), 694–702 (2003)
14. Nishihara, H., Smit, A.F., Okada, N.: Functional noncoding sequences derived from sines in the mammalian genome. Genome research **16**(7), 864–874 (2006)
15. Häsler, J., Strub, K.: Alu elements as regulators of gene expression. Nucleic Acids Research **34**(19), 5491–5497 (2006)
16. Jjingo, D., Huda, A., Gundapuneni, M., Mariño-Ramírez, L., Jordan, I.K.: Effect of the transposable element environment of human genes on gene length and expression. Genome biology and evolution **3**, 259–271 (2011)
17. Han, J.S., Szak, S.T., Boeke, J.D.: Transcriptional disruption by the l1 retrotransposon and implications for mammalian transcriptomes. Nature **429**(6989), 268 (2004)
18. Havecker, E.R., Gao, X., Voytas, D.F.: The diversity of ltr retrotransposons. Genome biology **5**(6), 225 (2004)
19. Cohen, C.J., Lock, W.M., Mager, D.L.: Endogenous retroviral ltrs as promoters for human genes: a critical assessment. Gene **448**(2), 105–114 (2009)
20. Jern, P., Coffin, J.M.: Effects of retroviruses on host genome function. Annual review of genetics **42**, 709–732 (2008)
21. Muñoz-López, M., García-Pérez, J.L.: Dna transposons: nature and applications in genomics. Current genomics **11**(2), 115–128 (2010)
22. Skipper, K.A., Andersen, P.R., Sharma, N., Mikkelsen, J.G.: Dna transposon-based gene vehicles-scenes from an evolutionary drive. Journal of biomedical science **20**(1), 92 (2013)
23. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., *et al.*: The evolution of gene expression levels in mammalian organs. Nature **478**(7369), 343–348 (2011)
24. Marin, R., Cortez, D., Lamanna, F., Pradeepa, M.M., Leushkin, E., Julien, P., Liechti, A., Halbert, J., Brüning, T., Mössinger, K., *et al.*: Convergent origination of a drosophila-like dosage compensation mechanism in a reptile lineage. Genome research **27**(12), 1974–1987 (2017)
25. Georges, A., Li, Q., Lian, J., O'Meally, D., Deakin, J., Wang, Z., Zhang, P., Fujita, M., Patel, H.R., Holleley, C.E., *et al.*: High-coverage sequencing and annotated assembly of the genome of the australian dragon lizard pogona vitticeps. Gigascience

4(1), 45 (2015)

26. Krueger, F.: Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files (2015)
27. Li, B., Dewey, C.N.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics **12**(1), 323 (2011)
28. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. Nature methods **9**(4), 357 (2012)
29. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)
30. Zeng, L., Kortschak, R.D., Raison, J.M., Bertozzi, T., Adelson, D.L.: Superior ab initio identification, annotation and characterisation of tes and segmental duplications from genome assemblies. bioRxiv, 190694 (2017)
31. Sudmant, P.H., Alexis, M.S., Burge, C.B.: Meta-analysis of rna-seq expression data across species, tissues and studies. Genome biology **16**(1), 287 (2015)
32. Merkin, J., Russell, C., Chen, P., Burge, C.B.: Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science **338**(6114), 1593–1599 (2012)
33. Slotkin, R.K., Martienssen, R.: Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics **8**(4), 272 (2007)
34. Bourque, G.: Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current opinion in genetics & development **19**(6), 607–612 (2009)
35. Lynch, V.J., Nnamani, M.C., Kapusta, A., Brayer, K., Plaza, S.L., Mazur, E.C., Emera, D., Sheikh, S.Z., Grützner, F., Bauersachs, S., *et al.*: Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell reports **10**(4), 551–561 (2015)
36. Rebollo, R., Romanish, M.T., Mager, D.L.: Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annual review of genetics **46**, 21–42 (2012)
37. Chuong, E.B., Elde, N.C., Feschotte, C.: Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics **18**(2), 71 (2017)
38. Polak, P., Domany, E.: Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC genomics **7**(1), 133 (2006)
39. Piedrafita, F.J., Molander, R.B., Vansant, G., Orlova, E.A., Pfahl, M., Reynolds, W.F.: An alu element in the myeloperoxidase promoter contains a composite sp1-thyroid hormone-retinoic acid response element. Journal of Biological Chemistry **271**(24), 14412–14420 (1996)
40. Eller, C.D., Regelson, M., Merriman, B., Nelson, S., Horvath, S., Marahrens, Y.: Repetitive sequence environment distinguishes housekeeping genes. Gene **390**(1), 153–165 (2007)
41. Buckley, R.M., Kortschak, R.D., Raison, J.M., Adelson, D.L.: Similar evolutionary trajectories for retrotransposon accumulation in mammals. Genome biology and evolution **9**(9), 2336–2353 (2017)



Genome size (GB)	TE Content (%)	Number of genes
2.0	58.47	21,698
3.5	57.98	21,327
3.0	48.73	20,338
1.0	17.29	18,346
1.8	51.02	18,595
1.7	42.87	18,907

Figure 1: **Divergence times and genome statistics of the major amniote lineages.** A) The silhouettes indicate species used in this study: *Ornithorhynchus anatinus* (platypus), *Monodelphis domestica* (opossum), *Homo sapiens* (human), *Gallus gallus* (chicken), *Anolis carolinensis* (anole lizard) and *Pogona vitticeps* (bearded dragon) (from top to bottom). Time since main speciation events obtained from TimeTree (www.timetree.org) are indicated (millions of years ago, Myr ago) [24]; B) Genome statistics.

Table 1: Summary of datasets and tissue samples analyzed in this study.

Dataset(s)	Tissues	Species
Marin	brain, heart, kidney, liver, ovary, testes	chicken, anole, platypus, opossum, *human
Brawand	brain, heart, kidney, liver, testes	chicken, anole, platypus, opossum, human
Private	brain, heart, kidney, liver, ovary, testes	bearded dragon

* Human samples in this set do not include ovary tissue.

Additional Files

Additional file 1 — Supplementary Information

Additional file contains supplementary figures and tables as referred to in the main body of the paper.

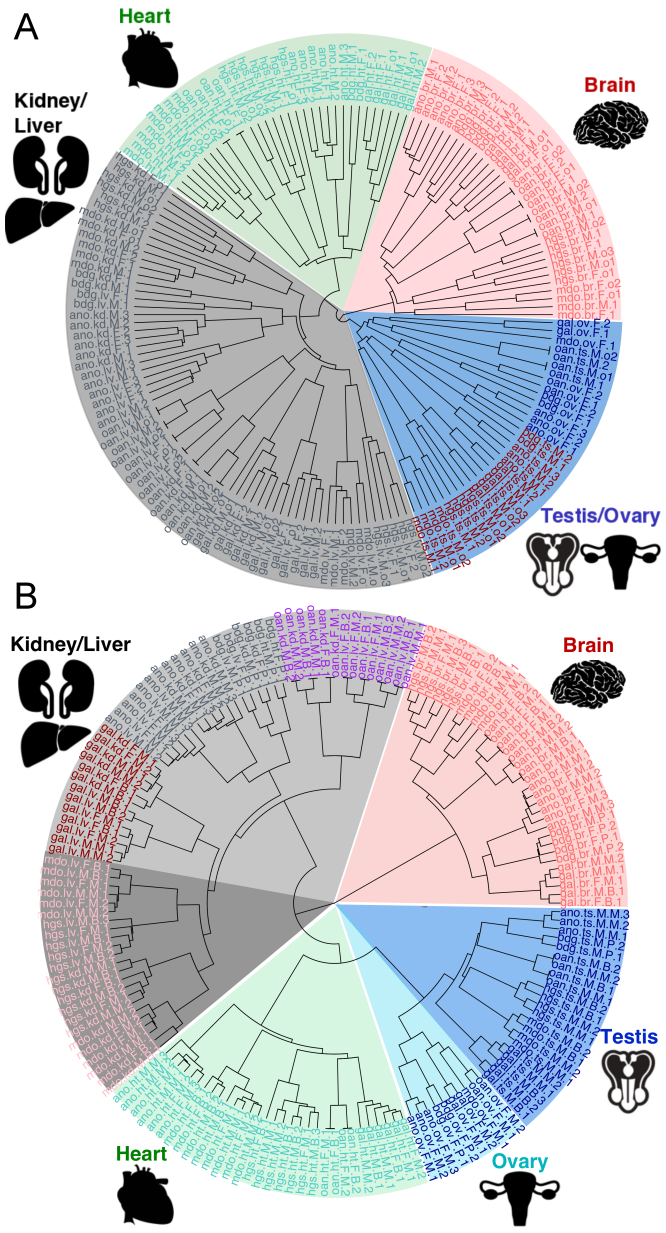


Figure 2: **Tissue specific *vs* species specific clustering of gene expression in amniotes.** a, Clustering of samples based on expression values, calculated as transcripts per million (TPM) of one to one orthologous genes expressed in heart, brain, kidney, liver, testis and ovary (n=6596). UPGMA (Unweighted Pair Group Method with Arithmetic Mean) hierarchical clustering was used with distance between samples calculated using the average of all distances between pairs. b, Clustering of samples based on expression values, calculated as transcripts per million (TPM) of one to one orthologous genes expressed in heart, brain, kidney, liver, testis and ovary (n=6596). Ward's minimum variance hierarchical clustering was used with distance between samples measured by the squared Euclidean distance.

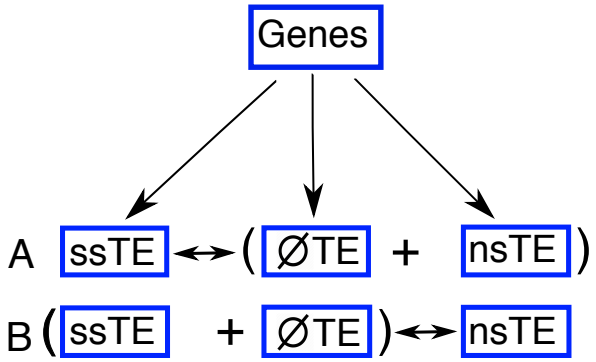


Figure 3: **Gene sets for expression comparison.** Because there were too few genes (orthologs or non-orthologs) with no TE insertions, we designed comparison sets based on the following scheme. We split our gene sets (either ortholog or non-ortholog) into three subsets: those containing recent species specific TE insertions (ssTE), those containing non-recent species specific TE insertions (nsTE) and those containing no TE (∅TE)

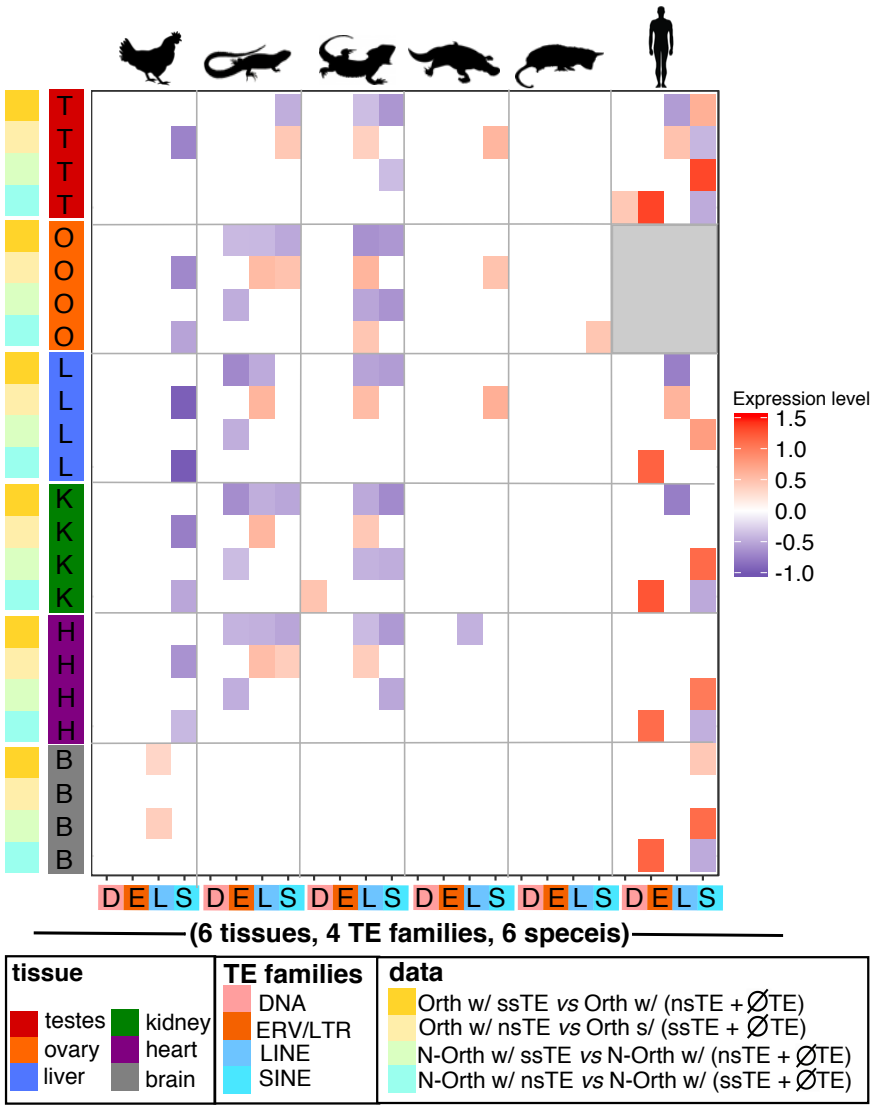


Figure 4: **Changes in the levels of ortholog/non-ortholog gene expression as a function of TE insertion.** This figure shows the association between orthologous/non-orthologous gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon(pogona), platypus, opossum and human) with the presence of recent species-specific TE insertions (ssTE) or non-recent species specific TE insertions (nsTE) (from left to right: DNA, ERV/LTR, LINE or SINE). A weighted bootstrap approach was used to compare the median difference in gene expression levels of orthologous/non-orthologous genes with a ssTE/nsTE insertion compared to orthologous/non-orthologous gene without ssTE/nsTE. Gene expression levels are log2-transformed. Comparisons without statistically significant gene expression changes are shown in white. Statistically significant increased gene expression shown in red and statistically significant decreased gene expression in blue. Grey shading indicates no samples were available for this comparison.

Chapter 5

Retrotransposons: Genomic and Trans-Genomic Agents of Change

"Research is what I'm doing when I don't know what I'm doing." — Lewis Thomas

The general view of transposable elements is that they are "selfish DNA parasites" because they spread through their hosts, such as humans, animals, plants as well as bacteria and provide for their own survival. Thanks to the Barbara McClintock, who first suggested the existence of retrotransposons and their influence in reshaping genomes in evolution, they are no longer seen as "selfish" or "junk". Indeed, the insertions of retrotransposon into human genome can cause the genetic dysfunction and alteration of gene expression contributing to cancer and other human diseases. The following excerpt appears as chapter 4 in *Evolutionary Biology: Biodiversification from Genotype to Phenotype*, discussing the role of retrotransposons as drivers of genome evolution.

Statement of Authorship

Title of Paper	Retrotransposons: Genomic and Trans-Genomic Agents of Change
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	David L. Adelson, Reuben M. Buckley, Atma M. Ivancevic, Zhipeng Qu, Lu Zeng (2015). Retrotransposons: Genomic and Trans-Genomic Agents of Change. Springer International Publishing. Pontarotti (ed.), Evolutionary Biology: Biodiversification from Genotype to Phenotype. DOI: 10.1007/978-3-319-19932-0_4

Principal Author

Name of Principal Author (Candidate)	Lu Zeng		
Contribution to the Paper	Designed one figures, provided suggestions and proof-read the book chapter.		
Overall percentage (%)	10%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	20/02/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Wrote the book chapter		
Signature		Date	26/2/18

Name of Co-Author	Atma M. Ivancevic		
Contribution to the Paper	Designed figures, provided suggestions and proof-read the book chapter		
Signature		Date	26/02/18

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Reuben M. Buckley		
Contribution to the Paper	Designed figures, provided suggestions and proof-read the book chapter		
Signature		Date	23/5/2017

Name of Co-Author	Zhipeng Qu		
Contribution to the Paper	Designed figures, provided suggestions and proof-read the book chapter		
Signature		Date	26/02/2018

Chapter 4
Retrotransposons: Genomic and Trans-Genomic Agents of Change

David L. Adelson, Reuben M. Buckley, Atma M. Ivancevic, Zhipeng Qu and Lu Zeng

Abstract Genome structure in higher eukaryotes is highly dependent on the type and abundance of transposable elements, particularly retrotransposons, in their non-coding DNA. Retrotransposons are generally viewed as genomic parasites that must be suppressed in order to ensure genome integrity. This perception is based on the instances of retrotransposons having caused deleterious structural variation in genomes. Recent data are beginning to provide a more positive view of the impact of retrotransposons, particularly in mammals, where the evolution of the placenta has depended on the exaptation of a type of retrotransposon, endogenous retroviruses. Finally, exosome trafficking of retrotransposons between cells has been shown to induce the innate immune system gene expression, possibly indicative of a role for retrotransposons in the regulation of the innate immune system. It may be time for us to review the status of retrotransposons and reclassify them as symbionts rather than parasites.

4.1 Evolutionary Origin and Structure of Retrotransposons

Genome structure and function are two sides of the same coin, and retrotransposons (AKA retrotransposable elements, retroelements and retroposons), self-replicating DNA sequences that are found in all eukaryotic taxa, have the capacity to make larger changes to genome structure than other sources of variation—such as DNA polymerase errors that lead to single nucleotide variation (SNV). Because retrotransposons can account for the majority of the genome sequence in eukaryotes, their accumulation and clade specificity have been implicated in speciation, regulation of gene expression, exaptation and structural variation. Understanding the

D.L. Adelson (✉) · R.M. Buckley · A.M. Ivancevic · Z. Qu · L. Zeng
School of Biological Sciences, University of Adelaide, North Terrace, Adelaide, SA 5005, Australia
e-mail: david.adelson@adelaide.edu.au

P. Pontarotti (ed.), *Evolutionary Biology: Biodiversification from Genotype to Phenotype*, DOI 10.1007/978-3-319-19932-0_4

mechanisms that govern retrotransposon distribution and replication is thus of fundamental importance.

The evolutionary origin of retrotransposons is a matter of debate, but sequence similarity of their reverse transcriptases with the catalytic subunit of telomerase (Eickbush 1997; Lingner et al. 1997) and phylogenetic studies of reverse transcriptase sequences can be interpreted to indicate that reverse transcriptase may have evolved from telomerase, or telomerase is the result of co-opting reverse transcriptase. However, there are also good arguments for the ancient, prokaryotic origin of reverse transcriptase as a descendant of group II introns, which are mobile, self-splicing introns (Boeke 2003).

Retrotransposons can be divided into four major classes (Eickbush and Jamburuthugoda 2008). This classification is based on the reverse transcriptase enzyme required for replication and encoded by these elements. In vertebrates, retrotransposons can account for half of the genome sequence, and in plants, up to 70 % of the genome. This chapter is focused on the mammalian/vertebrate retrotransposons and these are commonly described as falling into two broad categories: those containing long terminal repeats (LTR) and those not containing LTR (non-LTR) (Jurka et al. 2007).

Non-LTR retrotransposons encode their own internal promoter and one or two open reading frames (ORFs) with reverse transcriptase and endonuclease activities that are used for replication (Fig. 4.1). LTR containing retrotransposons resemble (endogenous) retroviruses (ERVs) in that they can contain additional ORFs similar to those found in retroviruses, and these are referred to as endogenous retrovirus-like elements (ERVL). ERVL LTR retrotransposons are believed to have evolved from DNA transposons (Bao et al. 2010) and then acquired additional genes from viruses such as *env*, allowing them to become retrovirus-like and to produce infectious particles.

4.2 The Retrotransposon Life cycle

Retrotransposons replicate via an RNA intermediate that is reverse transcribed and reinserted into the genome (Fig. 4.1) at short target motifs (Fig. 4.2) (Cost and Boeke 1998). For non-LTR retrotransposons, also called long interspersed elements (LINE), transcription is initiated by an internal Pol II promoter and the resulting transcript is then translated to produce two proteins, one of which, ORF2p has both reverse transcriptase and endonuclease activities (Feng et al. 1996; Moran et al. 1996). ORF2p has the ability to recognise short target sequences and initiate nicks at those locations which subsequently serve to prime the reverse transcription of the retrotransposon RNA directly into the genome (Eickbush and Jamburuthugoda 2008; Morrish et al. 2002).

Some retrotransposons do not contain ORFs (non-autonomous) and are dependent on retrotransposons that do (autonomous) (Jurka et al. 2007). Autonomous retrotransposons are longer (LINEs), whereas the shorter, non-autonomous

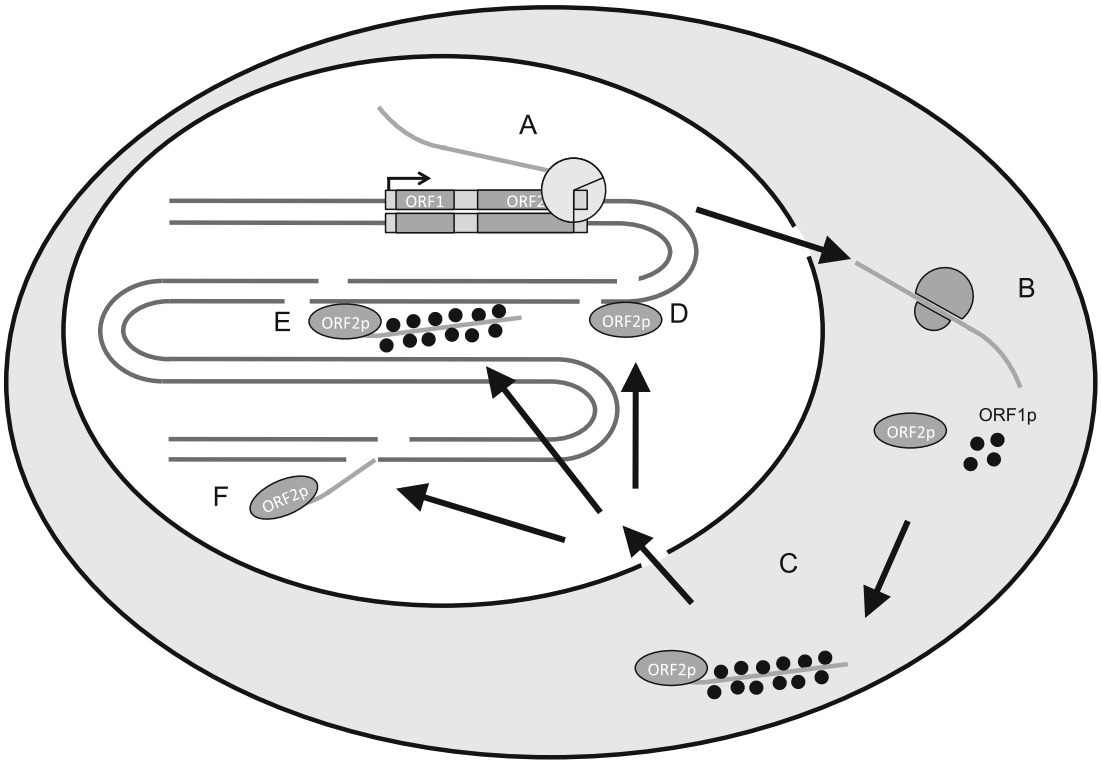


Fig. 4.1 Retrotransposon life cycle: *A* TEs are transcribed by RNA Pol II and exported to the cytoplasm (Swergold 1990). *B* In the cytoplasm, ORF1 and ORF2 are both translated. The ORF1 protein (*ORF1p*) is an RNA-binding protein believed to aid the entry of LINE L1 RNA into the nucleus (Martin 2006). The ORF2 protein (*ORF2p*) has both endonuclease and reverse transcriptase activities (Feng et al. 1996; Moran et al. 1996). *C* To enter the nucleus, ORF1p and ORF2p form a complex with the L1 RNA known as a ribonuclear protein (*RNP*) (Martin 2006). *D* The endonuclease activity of ORF2p creates double-stranded breaks without insertion of TEs (Gasior et al. 2006). *E* The endonuclease activity is essential for the process of target-primed reverse transcription (*TPRT*). TPRT requires that ORF2p creates a nick in each strand at the integration site. The LINE L1 RNA is then used as a template for the reverse transcriptase activity of ORF2p (Cost et al. 2002). *F* L1 RNA is able to insert into and aid in repairing double-stranded breaks independent of the endonuclease activity of ORF2p (Morrish et al. 2002)

elements are called short interspersed elements (SINEs). While LINEs are usually ubiquitously distributed across taxa, SINEs are usually clade specific, as they result from the fusion of an internal promoter containing transcript with the 3' end of a LINE.

The mechanism of SINE creation is still an open question, but most likely is a function of aspects of the LINE life cycle. SINEs have a composite structure: a 5' end similar to 5' tRNA, 7SL RNA or 5S rRNA promoters, a unique region and a 3' end similar to the 3' tail of LINEs (Piskurek and Jackson 2012). The most accepted hypothesis on SINE origins is based on the proposed template-switching mechanism of Buzdin et al. (Buzdin et al. 2002; Gilbert and Labuda 2000; Gogvadze and Buzdin 2009, Kramerov and Vassetzky 2005; Ohshima and Okada 2005). This template-switching mechanism is based on the study of pseudogenes, where the LINE (L1) reverse transcriptase switches from its own L1 mRNA to other nearby

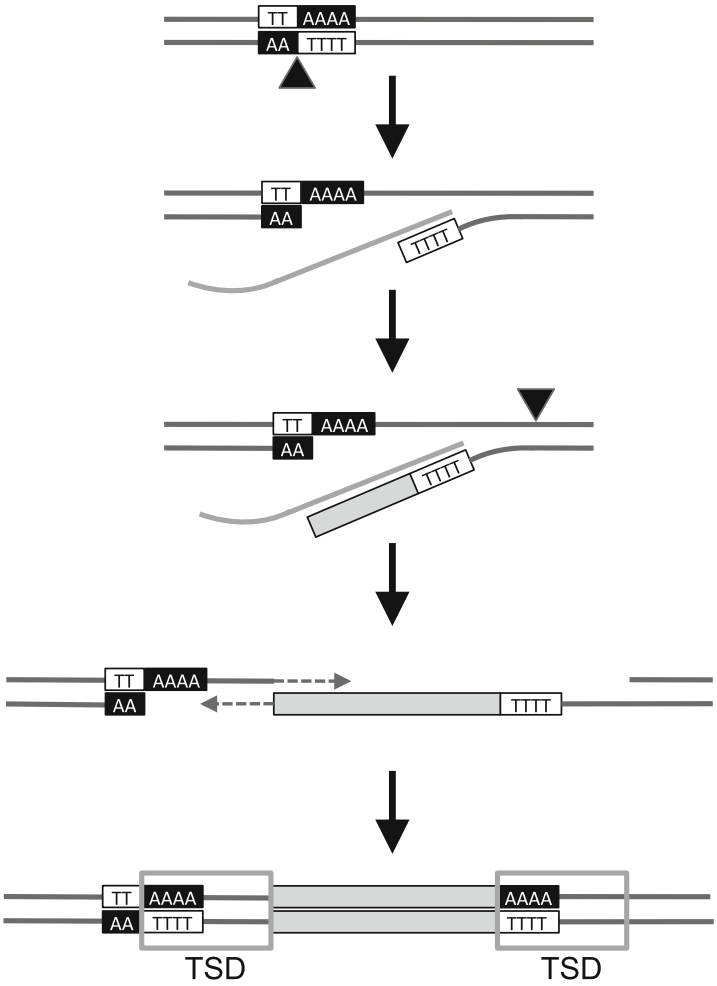


Fig. 4.2 Target-primed Reverse Transcription (*TPRT*) is how retrotransposons are inserted into the genome. ORF2p endonuclease activity creates a nick in the DNA at the AA/TTTT target site (Cost and Boeke, 1998). ORF2p reverse transcriptase activity then uses the cDNA copy as a template for DNA synthesis. Next ORF2p endonuclease activity creates a second nick in the DNA. The second DNA strand is then synthesised via double-strand break (*DSB*) repair and results in the formation of short target site duplications (*TSD*)

mRNA sequences through an RNA–RNA recombination process, thus creating new recombinant pseudogenes (and possibly SINEs) during L1 insertion (Buzdin et al. 2002; Gogvadze et al. 2007; Ichiyanagi et al. 2007; Piskurek and Jackson 2012). However, other investigators have suggested direct transposon into transposon (TnT) insertion as an alternative mechanism for the creation of novel transposable elements (Giordano et al. 2007; Ichiyanagi et al. 2007; Kriegs et al. 2007). The TnT mode of retrotransposon generation is what has led to the formation of SVA (SINE/VNTR/Alu) elements in humans, which are chimeric elements that can be mobilised by L1 elements and contain Alu-like sequence, Variable Number of Tandem Repeats (VNTR) sequence and SINE-R sequence resulting from a series of TnT events (Ostertag et al. 2003). The template-switching and TnT mechanisms are not mutually exclusive, and it is clear that both operate to create new SINEs, but at present we do not know which mechanism dominates.

Because retrotransposons can control their own expression through internal promoters [Pol II for LINEs and Pol III for SINEs and ERVs (Belancio et al. 2010a; Dieci et al. 2013)], expression is inextricably linked to the retrotransposon replication and to the evolution of new SINEs. As a result of this ability to autonomously insert new copies from expressed sequences into the genome, eukaryotes

have evolved mechanisms to keep retrotransposon expression in check in order to avoid large-scale deleterious structural variation.

4.2.1 Retrotransposon Suppression

There appear to be two main mechanisms for retrotransposon suppression: transcriptional repression and post-transcriptional degradation (Fig. 4.3). Transcriptional repression can be caused by methylation of retrotransposon promoters or alteration of chromatin state to make retrotransposons transcriptionally inaccessible. Proof for the importance of methylation is evident from the phenotype of *dnmt3l* (DNA (cytosine-5)-methyltransferase 3-like) knockout mice (Bourc'his and Bestor 2004; Webster et al. 2005), which undergo meiotic catastrophe associated with the rampant expression of retrotransposons in male germ cells. The *dnmt3l* locus encodes a protein that regulates methyl transferase activity required to methylate and suppress the activity of CpG islands in retrotransposon promoters (Vlachogiannis et al. 2015). In addition to CpG island methylation, transcription can be repressed by the alteration of chromatin status (Fadloun et al. 2013), and this may be mediated by piRNA transported to the nucleus (Kuramochi-Miyagawa et al. 2008).

Post-transcriptional degradation of retrotransposon RNA in the male germ line is mediated by piRNAs derived from retrotransposon sequences and amplified by the ping-pong reaction (Aravin et al. 2008). In the female germ line, the situation appears to be different, with siRNAs shown to mediate retrotransposon transcript destruction via the RNA-induced silencing complex (RISC) pathway (Ciaudo et al. 2013; Watanabe et al. 2008).

There may also be additional mechanisms that can suppress retrotransposons at the translational level (Grivna et al. 2006; Tanaka et al. 2011) or even at the post-translational level to interfere with ORF proteins binding to retrotransposon transcripts (Fig. 4.3) (Goodier et al. 2012). In spite of all of these mechanisms to suppress retrotransposons at various steps in their life cycle, they are still transcribed at some developmental stages and in many somatic tissues (Belancio et al. 2010b). Perhaps suppression is a loaded term in this context and perhaps what we are observing is actually the regulation of retrotransposon expression.

4.2.2 Retrotransposon Expression

At certain phases of the mammalian life cycle, retrotransposons are negatively regulated to a lesser degree and are therefore transcribed and able to retrotranspose. Because methylation of cytosine to 5-methyl-cytosine (5mC) is critical to retrotransposon silencing, retrotransposons are potentially most active at times of low genomic 5mC content, which occurs in mouse embryos at around 3.5 days of embryonic development and also in primordial germ cells (Hackett and Surani 2013).

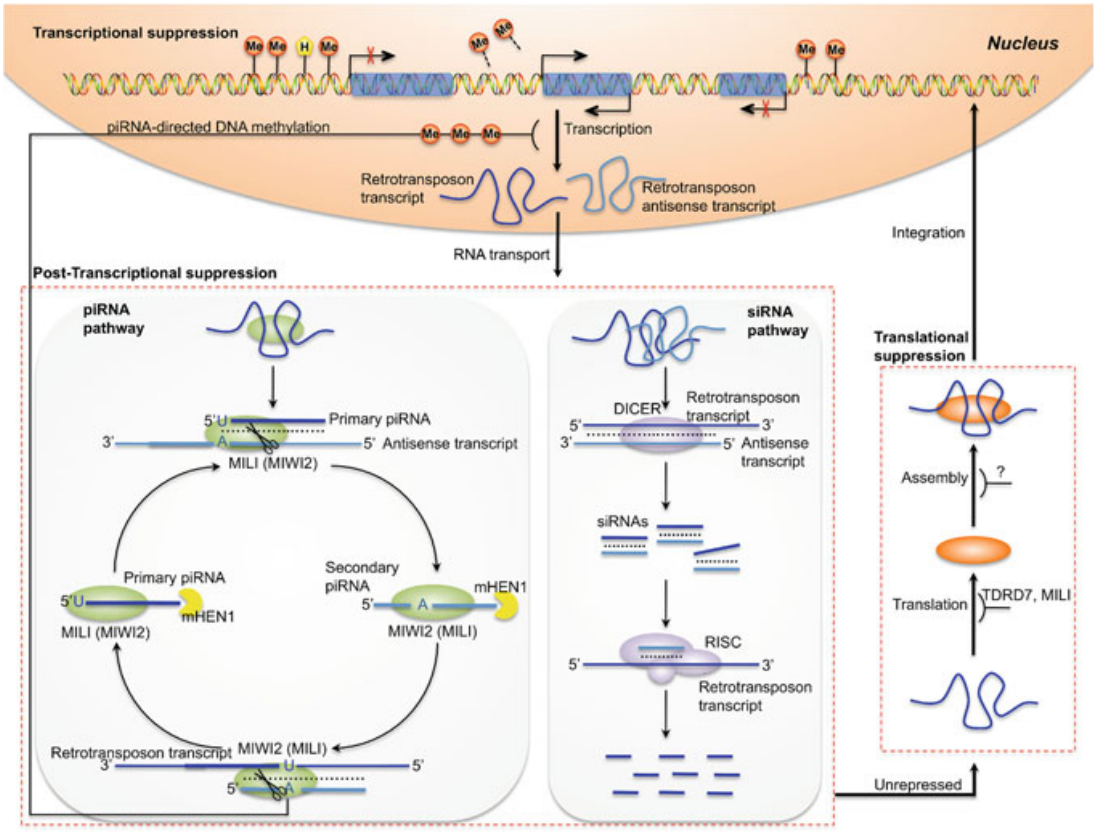


Fig. 4.3 A schematic overview of retrotransposon suppression. Retrotransposons can be suppressed by different mechanisms throughout their life cycle (Crichton et al. 2014). Transcriptional suppression: In most cell types, retrotransposons are in a repressed state due to high levels of DNA methylation or histone modifications (Fadloun et al. 2013; Meissner et al. 2008). In some specific developmental stages and cell types, some retrotransposon RNAs can be transcribed bidirectionally and transported from the nucleus to the cytoplasm (Fadloun et al. 2013). Post-transcriptional suppression: Retrotransposon RNAs can be silenced through the piRNA pathway (mostly in the male germ line) or siRNA pathway (mostly in the female germ line). The ping-pong cycle is a well-characterised model for piRNA synthesis. In the mouse, sense retrotransposon RNAs are processed into primary piRNAs. MILI (or MIWI2) is recruited to cleave antisense retrotransposon RNAs into secondary piRNAs with the guidance of primary piRNAs, and mHEN1 is used to subsequently methylate their 3' termini. Secondary piRNAs then bind with MIWI2 (or MILI) to cleave sense retrotransposon RNAs into primary piRNAs and close the loop of the ping-pong cycle (Aravin et al. 2008). piRNAs can also be transported to the nucleus to repress the transcription of retrotransposon by directing DNA methylation (Kuramochi-Miyagawa et al. 2008). For the siRNA pathway, sense and antisense retrotransposon transcripts can form double-strand RNAs, which are cleaved into double-strand siRNAs by DICER. Then, double-stranded siRNAs are unwound and loaded into the RISC to guide the degradation of retrotransposons (Ciaudo et al. 2013; Watanabe et al. 2008). Translational suppression: The Tudor domain-containing protein TDRD7 and MILI might be involved in the suppression of retrotransposon activity during translation (Grivna et al. 2006; Tanaka et al. 2011). Other repression mechanisms may also exist at later stages, such as the assembly stage of retrotransposon RNA and retrotransposon-encoded proteins (Goodier et al. 2012)

However, it is primarily in early embryos that L1 retrotransposons are transcribed and retrotranspose (Kano et al. 2009). Presumably, other suppression mechanisms keep retrotransposons in check in primary germ cells. In spite of significant levels of global 5mC in the genome at other stages of development, retrotransposons are also activated in specific somatic tissues, indicating that retrotransposon suppression is more complex than just ensuring high levels of 5mC, and it may be less stringent in some tissues/cell types. Faulkner et al. (2009) showed that up to 30 % of mouse or human transcripts from all tissues are of retrotransposon origin and that retrotransposons were transcribed in all tissues surveyed. Retrotransposon expression per se does not always mean that retrotransposition is occurring, as some retrotransposons have inserted into UTRs and are therefore transcribed as part of a mRNA. However, it has been shown in both neural progenitor cells and in the human brain that retrotransposition does occur at a detectable level, altering the genomic landscape of that tissue (Baillie et al. 2011; Coufal et al. 2009).

Retrotransposon expression and subsequent retrotransposition have significant impacts on the genomes of both germ line (via germ line insertions and early embryonic insertions) and soma. Germ line insertions can then be transmitted through vertical inheritance, while somatic insertions are not currently believed to contribute to the vertical inheritance of novel insertions. However, there is another mode of retrotransposon transmission: horizontal transfer, where retrotransposon sequences jump to another cell or species, and this type of transfer may be the result of a more general mechanism of intercellular retrotransposon transfer.

4.3 Horizontal Transfer

Horizontal transfer of transposons has been demonstrated in plants, insects and vertebrates. In the context of retroviruses (including ERVs that have maintained ORFs to support an infectious life cycle), horizontal transfer is a relatively commonplace event. For example, in plants, horizontal transfer of transposable elements is both widespread and frequent (El Baidouri et al. 2014). In animals, horizontal transfer of DNA transposons is also widespread (Ivancevic et al. 2013). A good example is in *Drosophila melanogaster* where P-elements swept through the population starting in the 1950s via horizontal transfer (Daniels et al. 1990). *Mariner* elements are also horizontally transmitted between species, including both insects and mammals (Lampe et al. 2003; Lohe et al. 1995; Maruyama and Hartl 1991). Furthermore, Space Invader (*SPIN*) elements have been horizontally transferred in mammals and other tetrapods, as have OC1 elements (Gilbert et al. 2010; Pace et al. 2008). It was not until the 1990s that the first evidence for horizontal transfer of retrotransposons was published, when the patchy phylogenetic distribution and likely horizontal transfer of BovB retrotransposons was first reported (Kordis and Gubensek 1998, 1999a).

4.3.1 BovB: An Example of Widespread Horizontal Transfer

The BovB retrotransposon (also known as LINE-RTE) is a 3.2 kb LINE with at least one large ORF encoding a reverse transcriptase and a possible small ORF1 overlapping with the large ORF (Malik and Eickbush 1998). In cattle and sheep, over a thousand full length BovB, hundreds of thousands of 5' truncated BovB fragments and derived SINEs (Bov-tA and Bov-tA2 (Lenstra et al. 1993; Okada and Hamada 1997) account for ~25 % of the genome sequence (Adelson et al. 2009; Jiang et al. 2014). The high degree of sequence conservation of BovB with sequences detected from the venom gland of *Vipera ammodytes* gave the first support to the idea of horizontal transfer of this retrotransposon (Kordis and Gubensek 1998, 1999b). BovB is now known to have a widespread, but patchy phylogenetic distribution, coupled to a high degree of sequence conservation, two of the hallmarks of horizontally transferred DNA (Fig. 4.4).

Even though BovB has horizontally transferred across a wide range of species, it has not always colonised the genome to the same extent in different species. Some

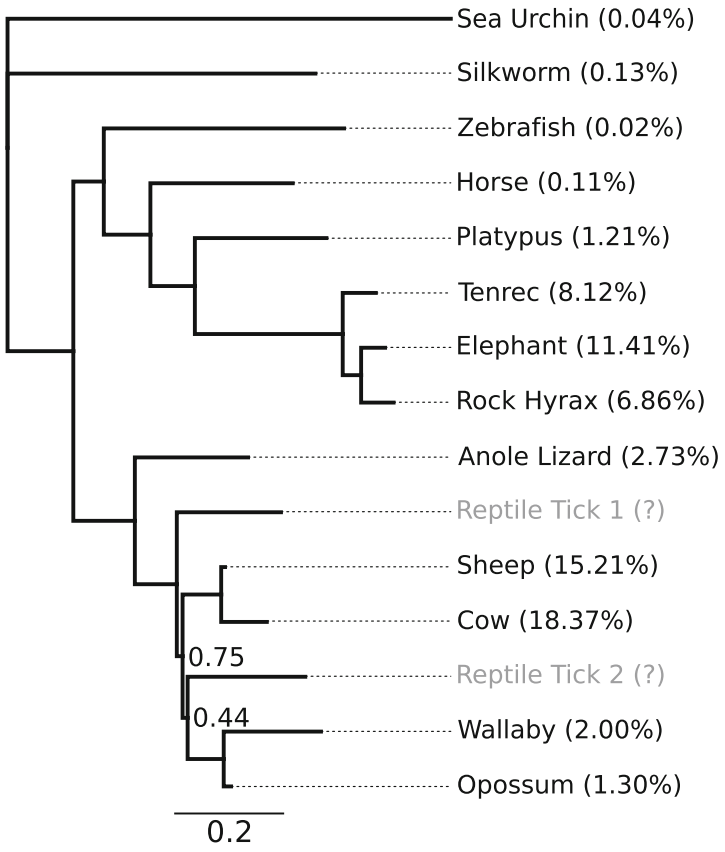


Fig. 4.4 BovB phylogeny Maximum likelihood tree of aligned BovB sequences based on Walsh et al. (2013), showing the sporadic distribution, sequence similarity and abundance of BovB elements across taxa. Local support values are only shown if <0.9. The labels at each branch tip give the species common name and (*in brackets*) the percentage of genome sequence identified as BovB elements for that species. Reptile Tick 1 is *Bothriocroton hydrosauri*, Reptile Tick 2 is *Amblyomma limbatum*; and the BovB genome coverage for these ticks is unknown

lineages such as ruminants and afrotheria have a high percentage of their genomes derived from BovB, whereas in other species BovB has not retrotransposed as prolifically (Fig. 4.4). This difference may be indicative of either variability in how different species suppress retrotransposons or it may simply reflect stochasticity in the population dynamics of retrotransposon expansion in different genomes. Presumably, the initial horizontal transfer event that results in retrotransposition and replication needs only a single germ line incorporation which can either replicate exponentially or "fizzle out" within the "genomic ecosystem" (Brookfield 2005; Le Rouzic et al. 2007). It is clear based on the currently available small and biased (towards mammals) sample of available genome sequences that retrotransposons as exemplified by BovB are capable of widespread and near ubiquitous horizontal transfer, and that this transfer might be enabled by parasites, such as ticks, that feed on blood. However, what is currently lacking is/are the molecular mechanism(s) for these transfers.

4.3.2 Possible Mechanisms/Modes of Transfer

A number of vectors, including arthropods, viruses, snails and DNA transposons, have been proposed for horizontal transfer, and the current state of knowledge was recently summarised by Ivancevic et al. (2013). It is relatively easy to see how a virus or transposon might act as a vector to package or transpose retrotransposons, but at the molecular level, it is not as obvious how eukaryotic vectors might effect the transfer of retrotransposon sequences between species, let alone into the germ line of another species.

4.3.2.1 Viruses as Vectors

For retrotransposons, the only example at present of a molecular virus vector is the taterapox virus (a dsDNA virus) which may have mediated transfer of Sauria SINE between reptiles and West African rodents (Piskurek and Okada 2007). This can be viewed as a highly unusual transfer, as a non-autonomous retrotransposon should not be as likely to colonise a new genome after transfer as an autonomous retrotransposon, such as a LINE. However, if cognate autonomous LINEs are present in both source and recipient species, a non-autonomous SINE could replicate effectively in the recipient species. RNA viruses have also been proposed as vectors of horizontal transfer for retrotransposons as they might package non-LTR retrotransposon transcripts inside infectious virus particles, but a tangible example for this type of transfer has yet to be demonstrated. Interestingly, *Mariner*-like DNA transposons are the plausible vectors for transfer of the CR1 retrotransposon in butterflies and moths (Sormacheva et al. 2012).

4.3.2.2 Endogenous Retroviruses/LTR Retrotransposons

As mentioned in Sect. 4.1, LTR retrotransposons are believed to have arisen from retrotransposons that acquired viral genes allowing them to become infectious, possibly leading to the evolution of retroviruses (Shimotohno and Temin 1981). In addition, waves of retroviral invasions into eukaryotic genomes have resulted in the formation of ERVs. While some ERVs have remained endogenous, occasionally they are able to become infectious and transfer to other genomes, where they can cause disease and eventually become domesticated. This is currently the case for a rodent ERV that has infected Koalas and is causing leukaemia in its new host while colonising the germ line as a new ERV (Tarlinton et al. 2006). Over time, domesticated retroviruses (ERVs) have contributed significantly to the genomic landscape of eukaryotes and have been co-opted into various aspects of eukaryotic biology (Feschotte and Gilbert 2012). In addition to this evolution of the capacity for horizontal transfer via infection, it is possible that retroviruses could package non-infectious non-LTR retrotransposons as a part of their viral payload. While there is no solid evidence for such transfer, exosomes/microvesicles are able to incorporate virus particles and transfer them to adjacent cells. This raises the question of whether exosomes can also transfer retrotransposon sequences directly.

4.3.2.3 Exosomes/Vesicles as Vectors

Exosomes are a class of membrane vesicle that has recently been shown to contain protein and RNA including miRNAs, piRNAs and retrotransposon sequences that they can transport from cell to cell (Batagov and Kurochkin 2013, Li et al. 2013; Skog et al. 2008; Valadi et al. 2007; Villarroya-Beltri et al. 2013; Yuan et al. 2009). Furthermore, exosome transport of Pol III-produced retrotransposon sequences has been specifically shown to regulate cancer therapy resistance pathways, including interferon-stimulated genes by direct activation of retinoid acid-inducible gene 1 (RIG-I) (Boelens et al. 2014). One of the hallmarks of Pol III transcripts is their 5' triphosphate group, which is recognised specifically by RIG-I as a trigger for activation. Pol III is responsible for the transcription of primarily housekeeping-type genes such as tRNAs and rRNAs, but it also transcribes many other loci, including SINEs that have originated from a fusion of Pol III promoter containing transcripts with LINE 3' sequences (Belancio et al. 2010b; Dieci et al. 2013). Because retrotransposons are known to be somatically expressed (see Sect. 4.2.2) in many tissues and cell types, they are likely to be present in exosomes exported by those cell types.

In the context of horizontal transfer, one can envision a number of potential scenarios for intercellular transport of retrotransposon sequences by exosomes (Fig. 4.5). Exosome-mediated transfer could allow transfer of retrotransposon sequences from a mammal or reptile to somatic cells of a parasite such as a tick through blood-borne exosomes. Within the tick, exosome-mediated transfer could then allow transmission to the germ line from the soma and eventual transmission back to other species used as food sources by that species of tick.

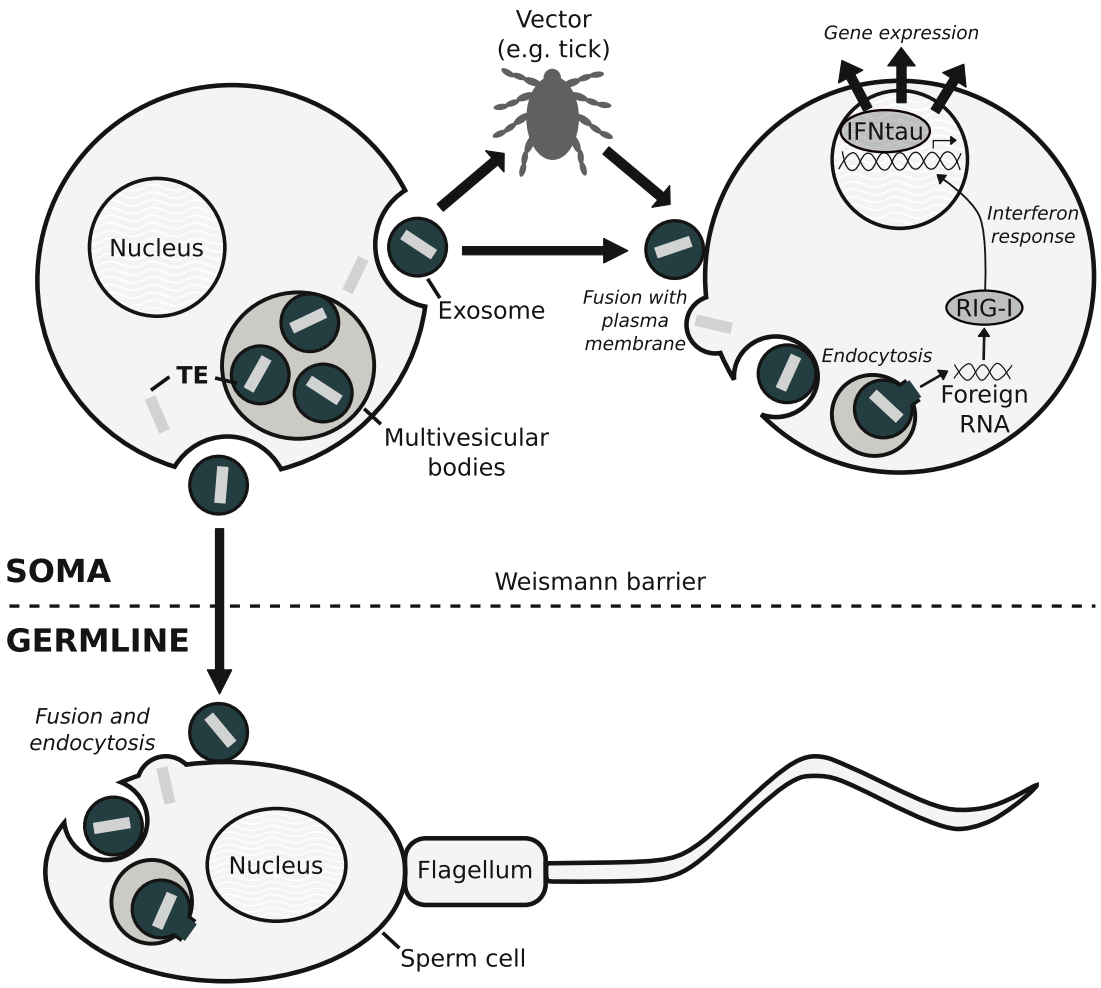


Fig. 4.5 Possible scenarios of intercellular transfer of transposable elements via exosomes. TEs packaged in exosomes can be transferred between both somatic and germline cells. Within an organism, a TE can travel from a somatic, exosome-generating cell directly (e.g. through the blood) into a somatic, exosome-target cell by fusing with the plasma membrane and undergoing endocytosis. Similarly, TEs can be horizontally transferred between the somatic cells of different organisms or species, via some kind of vector (e.g. a parasite). Exosomes can also carry TEs from the soma to the germ line, making them a permanent change in the genome that is eventually passed down to the offspring. Note that for simplicity only entry to the male germ line is shown above. In addition to the transfer of TEs, once inside the target cell, this "foreign RNA" from the TE can trigger an interferon pathway response by inducing the interferon signal transduction pathway via RIG-I. For example, in ruminants, exosomes loaded with ERV/TE RNAs trigger pattern recognition receptors, stimulating the innate immune system and production of interferon-tau, which plays a role in pregnancy recognition and placentation (see Sect. 4.4.4)

While one might envision that the existing piRNA-based suppression system might degrade these retrotransposon sequences rapidly, it also appears that retrotransposon sequences (as exosome cargo) have been co-opted into a signalling role for the innate immune system in vertebrates and used to activate interferon-stimulated genes in the absence of interferon (Dreux et al. 2012; Li et al. 2013). This would not be the first time that retrotransposon sequences have been co-opted for gene regulation (Feschotte 2008; Feschotte and Gilbert 2012), but it introduces a

new dimension of intercellular regulation of gene expression in the context of the evolutionary impact of retrotransposons.

4.4 Evolutionary Impacts

Retrotransposons are known to affect genome structure and hence function. The specific types of structural changes they introduce upon retrotransposition can have a wide-ranging set of subsequent effects in terms of genome structure, gene expression and gene function. More recently, it has become clear that retrotransposons have had a profound impact on the evolution of placentation in mammals.

4.4.1 Genome Structure

Retrotransposon insertion can directly perturb gene structure, but it can also have significant effects on a larger scale (Fig. 4.6). In particular, if retrotransposons form an array of elements with the same orientation on a chromosome, they can serve as

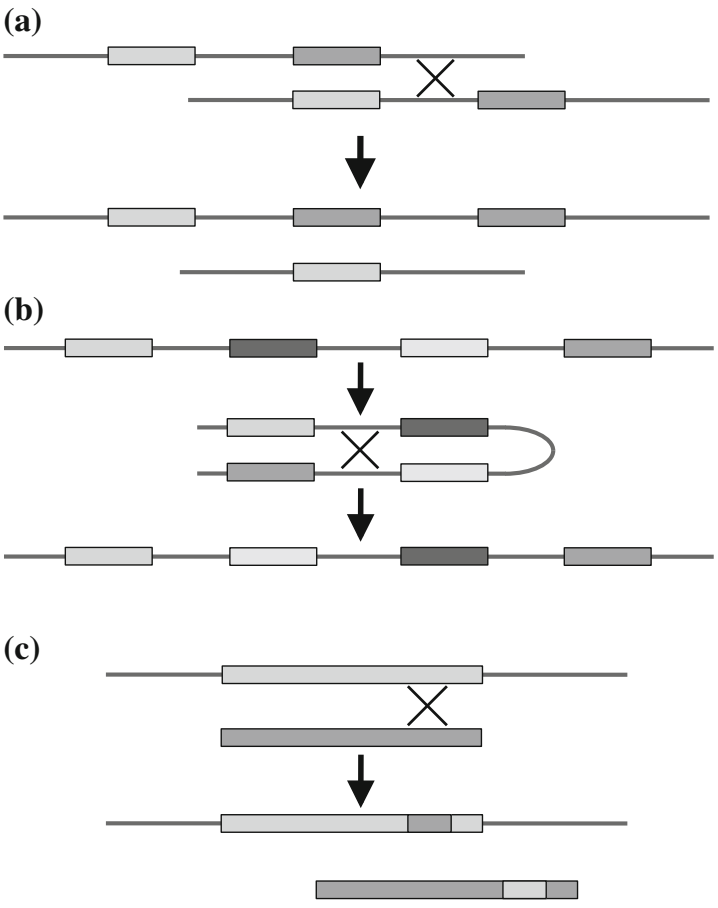


Fig. 4.6 Retrotransposons can lead to changes in genome structure. **a** Changes in CNVs result from non-allelic homologous recombination (*NAHR*) caused by the insertion of many TEs from the same family (Stankiewicz and Lupski 2002; Startek et al. 2015). **b** Chromosomal inversion is also the result of NAHR (Stankiewicz and Lupski 2002). **c** SINE elements have potential to drive change through gene conversion (Roy et al. 2000)

a substrate for non-allelic homologous recombination (NAHR) leading to segmental duplication (Fig. 4.6a) (Stankiewicz and Lupski 2002; Startek et al. 2015). However, statistical analysis of repeats in flanking regions of segmental duplications found that only ~10 % of segmental duplications could be attributed to flanking repetitive elements (Zhou and Mishra 2005). Other types of rearrangements have been shown to result from arrays of repeats such as inversions (Fig. 4.6b) and gene conversion (Fig. 4.6c).

While it is clear that retrotransposons can have indirect effects on genome structure as mentioned above, given the limitations inherent in identifying small segmental duplications and copy number variants the precise magnitude of these effects is unknown.

4.4.2 Gene Expression

As shown in Fig. 4.7, transposable elements can insert into and next to genes, affecting gene expression through multiple mechanisms, including epigenetic silencing of transcription, shortening a transcript via premature poly-Adenylation,

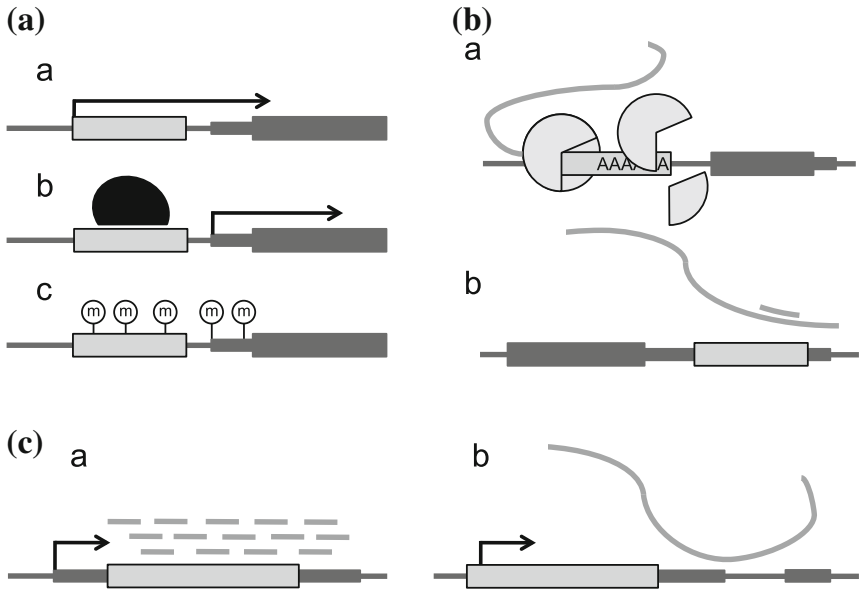


Fig. 4.7 Retrotransposons can alter gene expression. **a** 5' insertion of a retrotransposon with respect to a gene. *a* TEs are able to act as alternative promoters to adjacent genes (Faulkner et al. 2009; Speek 2001). *b* TEs are able to act as transcription factor binding sites (*TFBS*) and are thereby able to modulate gene expression (Bourque et al. 2008). *c* In plants, epigenetic silencing of TEs silences nearby genes; this is also likely to occur in animals (Buckley and Adelson 2014; Hollister and Gaut 2009). **b** 3' insertion of a retrotransposon *a* polyA signal/tail of the retrotransposon can result in shortened transcripts (Lee et al. 2008; Perepelitsa-Belancio and Deininger 2003). *b* Retrotransposon insertion in the 3' UTR of a gene can provide a target site for piRNAs which down-regulate gene expression (Watanabe et al. 2014). **c** Intergenic insertion of TEs. *a* Insertion of TEs into a piRNA cluster results in piRNAs that can target genes carrying TE-derived sequences (Yamamoto et al. 2013). *b* TEs involved in the origin and evolution of lncRNA (Kapusta et al. 2013)

driving piRNA expression or altering 3' UTR structure to affect mRNA stability. Analysis of retrotransposon insertions into or near genes has shown that many genes have been altered in ways that are likely to alter expression (Jjingo et al. 2011; Jordan et al. 2003) and analysis of enhancers has shown that retrotransposons drive the evolution of eukaryotic enhancers (McDonald et al. 1997). All of these effects on gene expression are subject to selection and are therefore part of the evolutionary process. Not all insertions into genes will affect regulation of gene expression, some can directly affect the coding sequence or coding potential of genes through exaptation.

4.4.3 Exaptation

When retrotransposons contribute to non-coding or protein coding exon sequences, they are referred to as exaptations. These exaptations may or may not be subject to immediate purifying selection, depending on the type of change they cause. Some exaptations that prove beneficial are selected for, but these are rare. Many examples of exaptation come from non-coding transcripts, where retrotransposon insertions have led to novel piRNA and miRNA transcripts (Jurka et al. 2007; Yamamoto et al. 2013). In fact, only ~50 instances of coding sequences derived from LTR retrotransposons syntenic between human and mouse have been identified (Jurka et al. 2007). One of these encodes the PEG10 (paternally expressed gene 10) locus, which is required for placentation. Occasionally, insertion of a retrotransposon sequence into an intron can lead to exonisation of part of the retrotransposon sequence as an alternative transcript through the presence of splice donor/acceptor sites in the sequence (Fig. 4.8). When this happens, sometimes the alternative transcripts are deleterious because of impaired function, and the regulation of alternative splicing may then become an additional regulatory mechanism for the affected gene (Lorenz et al. 2007).

4.4.4 Innate Immunity/Pregnancy Recognition

Some exaptations of retrotransposon sequences have been well-characterised, particularly in terms of the evolution of placentation. There is strong evidence for exaptation of ERV genes in both mouse and hominoid primates required for placental function (Chuong 2013; Haig 2012; Mallet et al. 2004). One of the most striking such exaptations is the role of endogenous jaagsiekte retrovirus (enJSRV) in ruminant pregnancy recognition and placentation. The domestic ruminant conceptus expresses interferon-tau (IFNT) from days 10 to 12, which dramatically alters gene expression in the uterine epithelium and stroma (Bazer et al. 2008; Dunlap et al. 2006; Gray et al. 2006; Spencer and Bazer 1995). At the same time, enJSRVs are released into the ruminant reproductive tract and they are known to

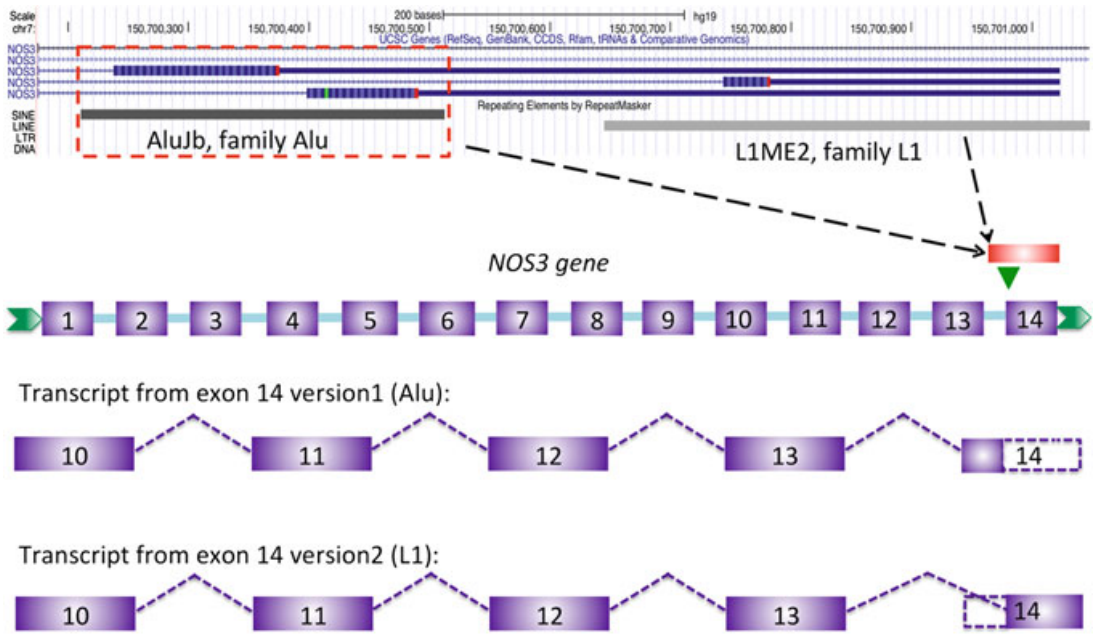


Fig. 4.8 Retrotransposon exaptation influences mRNA processing and can cause multiple splice variants. At the top, the UCSC browser (Kent et al. 2002) track for the human NOS3 gene is shown, including repeat element annotation. *Below*, a schematic of the 3' end of the human NOS3 gene illustrating an Alu element (*black bar*) inserted into intron 13. This retrotransposon provides exon 14 alternative splicing version 1. An adjacent L1 insertion can result in exon 14 alternative splicing version 2 (Lorenz et al. 2007). Dashed lines indicate a splicing event

regulate key peri-implantation development in the embryo and placenta (Dunlap et al. 2005, 2006). enJSRVs therefore have been exapted to regulate key aspects of development associated with implantation and placentation by virtue of their ability to trigger expression of IFNT expression in the conceptus. Recently, exosomes have been shown to be part of the specific mechanism used to trigger IFNT expression in this system, but without specifically testing for retrotransposon RNA content (Ruiz-Gonz ez et al. 2014, 2015). We speculate that exosomes loaded with retrotransposon sequences may also be involved in pregnancy recognition more generally in order to activate the STAT1 pathway in an interferon-free fashion.

SINE/ERV transcripts packaged into exosomes can trigger RIG-I in target cells leading to IFN independent activation of the IFN pathway, leading us to speculate that the role of retrotransposons is broader than previously thought, and that they may be involved in global regulation of the innate immune system.

4.5 Conclusion

Retrotransposons are abundant, found in a broad phylogenetic distribution and yet in spite of clade specific non-autonomous variants, exhibit a significant degree of commonality. Furthermore, their transcription is highly regulated, rather than

suppressed at all times. These facts, along with the evidence of pervasive and widespread horizontal transfer and an exosome-based mechanism for transfer that has likely co-evolved with the innate immune system and placentation, suggest to us that retrotransposons are not genomic parasites but rather genomic symbionts. We hypothesise that mammals and other vertebrates depend on these symbionts for cell-to-cell signalling in innate immunity and reproduction.

Acknowledgments The authors wish to thank R. Daniel Kortschak and Joy M. Raison for helpful discussions and advice.

References

Adelson DL, Raison JM, Edgar RC (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci USA 106:12855

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 31:785

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nat Cell Biol 479:534

Bao W, Kapitonov VV, Jurka J (2010) Ginger dna transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA 1(1):3. doi:10.1186/1759-8753-1-3

Batagov AO, Kurochkin IV (2013) Exosomes secreted by human cells transport largely mrna fragments that are enriched in the 3'-untranslated regions. Biol Direct 8:12. doi:10.1186/1745-6150-8-12

Bazer FW, Burghardt RC, Johnson GA, Spencer TE, Wu G (2008) Interferons and progesterone for establishment and maintenance of pregnancy: interactions among novel cell signaling pathways. Reprod Biol 8(3):179–211

Belancio VP, Roy-Engel AM, Deininger PL (2010a) All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol 20(4):200–210. doi:10.1016/j.semcancer.2010.06.001

Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P (2010 b) Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res 38:3909

Boeke JD (2003) The unusual phylogenetic distribution of retrotransposons: a hypothesis. Genome Res 13(9):1975–1983. doi:10.1101/gr.1392003

Boelens MC, Wu TJ, Nabet BY, Xu B, Qiu Y, Yoon T, Azzam DJ, Twyman-Saint Victor C, Wiemann BZ, Ishwaran H, Ter Brugge PJ, Jonkers J, Slingerland J, Minn AJ (2014) Exosome transfer from stromal to breast cancer cells regulates therapy resistance pathways. Cell 159 (3):499–513. doi:10.1016/j.cell.2014.09.051

Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3 l. Nature 431(7004):96–99. doi:10.1038/nature02886

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18(11):1752–1762. doi:10.1101/gr.080663.108

Brookfield JFY (2005) The ecology of the genome—mobile dna elements and their hosts. Nat Rev Genet 6(2):128–136. doi:10.1038/nrg1524

Buckley RM, Adelson DL (2014) Mammalian genome evolution as a result of epigenetic regulation of transposable elements. Biomol Concepts 5(3):183–194. doi:10.1515/bmc-2014-0013

Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T, Lebedev Y, Sverdlov E (2002) A new family of chimeric retrotranscripts formed by a full copy of u6 small nuclear rna fused to the 3' terminus of l1. Genomics 80(4):402–406

Chuong EB (2013) Retroviruses facilitate the rapid evolution of the mammalian placenta. Bioessays 35:853

Ciaudo C, Jay F, Okamoto I, Chen CJ, Sarazin A, Servant N, Barillot E, Heard E, Voinnet O (2013) Rnai-dependent and independent control of line1 accumulation and mobility in mouse embryonic stem cells. PLoS Genet 9(11):e1003,791. doi:10.1371/journal.pgen.1003791

Cost GJ, Boeke JD (1998) Targeting of human retrotransposon integration is directed by the specificity of the l1 endonuclease for regions of unusual dna structure. Biochemistry 37 (51):18081–18093

Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human l1 element target-primed reverse transcription in vitro. EMBO J 21(21):5899–5910

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH (2009) L1 retrotransposition in human neural progenitor cells. Nat Cell Biol 460:1127

Crichton JH, Dunican DS, Maclennan M, Meehan RR, Adams IR (2014) Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. Cell Mol Life Sci 71(9):1581–1605. doi:10.1007/s00018-013-1468-0

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A (1990) Evidence for horizontal transmission of the p-transposable element between drosophila species. Genetics 124:339

Dieci G, Conti A, Pagano A, Carnevali D (2013) Identification of rna polymerase iii-transcribed genes in eukaryotic genomes. Biochim Biophys Acta 1829(3–4):296–305. doi:10.1016/j.bbagrm.2012.09.010

Dreux M, Garaigorta U, Boyd B, Décembre E, Chung J, Whitten-Bauer C, Wieland S, Chisari FV (2012) Short-range exosomal transfer of viral rna from infected cells to plasmacytoid dendritic cells triggers innate immunity. Cell Host Microbe 12(4):558–570. doi:10.1016/j.chom.2012.08.010

Dunlap KA, Palmarini M, Adelson DL, Spencer TE (2005) Sheep endogenous betaretroviruses (enjsrvs) and the hyaluronidase 2 (hyal2) receptor in the ovine uterus and conceptus. Biol Reprod 73(2):271–279. doi:10.1095/biolreprod.105.039776

Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, Spencer TE (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 103(39):14390–14395. doi:10.1073/pnas.0603836103

Eickbush TH (1997) Telomerase and retrotransposons: which came first? Science (New York, NY) 277(5328):911–912

Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221

El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24(5):831–838. doi:10.1101/gr.164400.113

Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, Carninci P, Torres-Padilla ME (2013) Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of line-1 by rna. Nat Struct Mol Biol 20(3):332–338. doi:10.1038/nsmb.2495

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41:563

Feng Q, Moran J, Kazazian H, Boeke J (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905

Feschotte C (2008) Opinion—transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397

Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13(4):283–296. doi:10.1038/nrg3199

Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. Journal of Molecular Biology

Gilbert C, Schaack S, Pace JK, Brindley PJ, Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347–1352

Gilbert N, Labuda D (2000) Evolutionary inventions and continuity of core-sines in mammals. J Mol Biol 298(3):365–377. doi:10.1006/jmbi.2000.3695

Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, Warburton P (2007) Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol 3: e137

Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci 66(23):3727–3742. doi:10.1007/s00018-009-0107-2

Gogvadze E, Barbisan C, Lebrun MH, Buzdin A (2007) Tripartite chimeric pseudogene from the genome of rice blast fungus magnaporthe grisea suggests double template jumps during long interspersed nuclear element (line) reverse transcription. BMC Genomics 8:360. doi:10.1186/1471-2164-8-360

Goodier JL, Cheung LE, Kazazian HH Jr (2012) Mov10 rna helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet 8(10):e1002941. doi:10.1371/journal.pgen.1002941

Gray CA, Abbey CA, Beremand PD, Choi Y, Farmer JL, Adelson DL, Thomas TL, Bazer FW, Spencer TE (2006) Identification of endometrial genes regulated by early pregnancy, progesterone, and interferon tau in the ovine uterus. Biol Reprod 74(2):383–394. doi:10.1095/biolreprod.105.046656

Grivna ST, Pyhtila B, Lin H (2006) Miwi associates with translational machinery and piwi-interacting rnas (pirnas) in regulating spermatogenesis. Proc Natl Acad Sci USA 103 (36):13415–13420. doi:10.1073/pnas.0605506103

Hackett JA, Surani MA (2013) Dna methylation dynamics during the mammalian life cycle. Philos Trans R Soc Lond B Biol Sci 368(1609):20110328. doi:10.1098/rstb.2011.0328

Haig D (2012) Retroviruses and the placenta. Current biology: CB

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19 (8):1419–1428. doi:10.1101/gr.091678.109

Ichiyanagi K, Nakajima R, Kajikawa M, Okada N (2007) Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. Genome Res 17:33

Ivancevic AM, Walsh AM, Kortschak RD, Adelson DL (2013) Jumping the fine LINE between species: horizontal transfer of transposable elements in animals catalyses genome evolution. Bioessays 35:12

Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W, Stanton JA, Brauning R, Barris WC, Hourlier T, Aken BL, Searle SMJ, Adelson DL, Bian C, Cam GR, Chen Y, Cheng S, DeSilva U, Dixen K, Dong Y, Fan G, Franklin IR, Fu S, Fuentes-Utrilla P, Guan R, Highland MA, Holder ME, Huang G, Ingham AB, Jhangiani SN, Kalra D, Kovar CL, Lee SL, Liu W, Liu X, Lu C, Lv T, Mathew T, McWilliam S, Menzies M, Pan S, Robelin D, Servin B, Townley D, Wang W, Wei B, White SN, Yang X, Ye C, Yue Y, Zeng P, Zhou Q, Hansen JB, Kristiansen K, Gibbs RA, Flicek P, Warkup CC, Jones HE, Oddy VH, Nicholas FW, McEwan JC, Kijas JW, Wang J, Worley KC, Archibald AL, Cockett N, Xu X, Wang W, Dalrymple BP (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344(6188):1168–1173. doi:10.1126/science.1252806

Jjingo D, Huda A, Gundapuneni M, Mariño-Ramrez L, Jordan IK (2011) Effect of the transposable element environment of human genes on gene length and expression. Genome Biol Evol 3:259–271. doi:10.1093/gbe/evr015

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Ann Rev Genomics Hum Genet

Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH (2009) L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev 23:1303

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9:e1003470

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at ucsc. Genome Res 12(6):996–1006. doi:10.1101/gr.229102. Article published online before print in May

Kordis D, Gubensek F (1998) Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. Proc Natl Acad Sci USA 95(18):10704–10709

Kordis D, Gubensek F (1999a) Horizontal transfer of non-LTR retrotransposons in vertebrates. Genetica 107:121

Kordis D, Gubensek F (1999b) Molecular evolution of bov-b lines in vertebrates. Gene 238 (1):171–178

Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol 247:165–221. doi:10.1016/S0074-7696(05)47004-7

Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (aves: Galliformes). BMC Evol Biol 7:190. doi:10.1186/1471-2148-7-190

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, Hata K, Li E, Matsuda Y, Kimura T, Okabe M, Sakaki Y, Sasaki H, Nakano T (2008) Dna methylation of retrotransposon genes is regulated by piwi family members mili and miwi2 in murine fetal testes. Genes Dev 22(7):908–917. doi:10.1101/gad.1640708

Lampe DJ, Witherspoon DJ, Soto-Adames FN, Robertson HM (2003) Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. Mol Biol Evol 20(4):554–562. doi:10.1093/molbev/msg069

Le Rouzic A, Boutin TS, Capy P (2007) Long-term evolution of transposable elements. Proc Natl Acad Sci USA 104(49):19375–19380. doi:10.1073/pnas.0705238104

Lee JY, Ji Z, Tian B (2008) Phylogenetic analysis of mrna polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic Acids Res 36(17):5581–5590. doi:10.1093/nar/gkn540

Lenstra JA, van Boxtel JA, Zwaagstra KA, Schwerin M (1993) Short interspersed nuclear element (sine) sequences of the bovidae. Anim Genet 24(1):33–39

Li CCY, Eaton SA, Young PE, Lee M, Shuttleworth R, Humphreys DT, Grau GE, Combes V, Bebawy M, Gong J, Brammah S, Buckland ME, Suter CM (2013) Glioma microvesicles carry selectively packaged coding and non-coding rnas which alter gene expression in recipient cells. RNA Biol 10(8):1333–1344. doi:10.4161/rna.25281

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. Science (New York, NY) 276:561

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL (1995) Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol Biol Evol 12 (1):62–72

Lorenz M, Hewing B, Hui J, Zepp A, Baumann G, Bindereif A, Stangl V, Stangl K (2007) Alternative splicing in intron 13 of the human enos gene: a potential mechanism for regulating enos activity. FASEB J 21(7):1556–1564. doi:10.1096/fj.06-7434com

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

Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci USA 101:1731

Martin SL (2006) The orf1 protein encoded by line-1: structure and function during l1 retrotransposition. J Biomed Biotechnol 2006(1):45621. doi:10.1155/JBB/2006/45621

Maruyama K, Hartl DL (1991) Evidence for interspecific transfer of the transposable element mariner between Drosophila and Zaprionus. J Mol Evol 33:514

McDonald JF, Matyunina LV, Wilson S, Jordan IK, Bowen NJ, Miller WJ (1997) Ltr retrotransposons and the evolution of eukaryotic enhancers. Genetica 100(1–3):3–13

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale dna methylation maps of pluripotent and differentiated cells. Nature 454(7205):766–770. doi:10.1038/nature07107

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87(5):917–927

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV (2002) Dna repair mediated by endonuclease-independent line-1 retrotransposition. Nat Genet 31(2):159–165. doi:10.1038/ng898

Ohshima K, Okada N (2005) Sines and lines: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res 110(1–4):475–490. doi:10.1159/000084981

Okada N, Hamada M (1997) The 3' ends of trna-derived sines originated from the 3' ends of lines: a new example from the bovine genome. J Mol Evol 44(1):52–56

Ostertag E, Goodier J, Zhang Y, Kazazian H (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73:1444

Pace JK, Gilbert C, Clark MS, Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci USA 105:17023

Perepelitsa-Belancio V, Deininger P (2003) Rna truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet 35(4):363–366. doi:10.1038/ng1269

Piskurek O, Jackson DJ (2012) Transposable elements: from dna parasites to architects of metazoan evolution. Genes (Basel) 3(3):409–422. doi:10.3390/genes3030409

Piskurek O, Okada N (2007) Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proc Natl Acad Sci USA 104(29):12046–12051. doi:10.1073/pnas.0700531104

Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL (2000) Potential gene conversion and source genes for recently integrated alu elements. Genome Res 10(10):1485–1495

Ruiz-González I, Xu J, Wang X, Burghardt RC, Dunlap K, Bazer FW (2014) Exosomes, endogenous retroviruses and toll-like receptors: pregnancy recognition in ewes. Reproduction

Ruiz-González I, Minten M, Wang X, Dunlap K, Bazer FW (2015) Involvement of TLR7 and TLR8 in conceptus development and establishment of pregnancy in Sheep. Reproduction

Shimotohno K, Temin HM (1981) Evolution of retroviruses from cellular movable genetic elements. Cold Spring Harb Symp Quant Biol 45(Pt 2):719–730

Skog J, Würdinger T, van Rijn S, Meijer DH, Gainche L, Sena-Esteves M, Curry WT Jr, Carter BS, Krichevsky AM, Breakefield XO (2008) Glioblastoma microvesicles transport rna and proteins that promote tumour growth and provide diagnostic biomarkers. Nat Cell Biol 10(12):1470–1476. doi:10.1038/ncb1800

Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A, Novikov A, Novikova O (2012) Vertical evolution and horizontal transfer of cr1 non-ltr retrotransposons and tc1/mariner dna transposons in lepidoptera species. Mol Biol Evol 29(12):3685–3702. doi:10.1093/molbev/mss181

Speek M (2001) Antisense promoter of human l1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21(6):1973–1985. doi:10.1128/MCB.21.6.1973-1985.2001

Spencer TE, Bazer FW (1995) Temporal and spatial alterations in uterine estrogen receptor and progesterone receptor gene expression during the estrous cycle and early pregnancy in the ewe. Biol Reprod 53(6):1527–1543

Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends Genet 18(2):74–82

Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P, Gambin A (2015) Genome-wide analyses of line-line-mediated nonallelic homologous recombination. Nucleic Acids Res. doi:10.1093/nar/gku1394

Swergold GD (1990) Identification, characterization, and cell specificity of a human line-1 promoter. Mol Cell Biol 10(12):6718–6729

Tanaka T, Hosokawa M, Vagin VV, Reuter M, Hayashi E, Mochizuki AL, Kitamura K, Yamanaka H, Kondoh G, Okawa K, Kuramochi-Miyagawa S, Nakano T, Sachidanandam R, Hannon GJ, Pillai RS, Nakatsuji N, Chuma S (2011) Tudor domain containing 7 (tdrd7) is essential for dynamic ribonucleoprotein (rnp) remodeling of chromatoid bodies during spermatogenesis. Proc Natl Acad Sci USA 108(26):10579–10584. doi:10.1073/pnas.1015447108

Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442 (7098):79–81. doi:10.1038/nature04841

Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO (2007) Exosome-mediated transfer of mrnas and micrornas is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9(6):654–659. doi:10.1038/ncb1596

Villarroya-Beltri C, Gutiérrez-Vázquez C, Sánchez-Cabo F, Pérez-Hernández D, Vázquez J, Martin-Cofreces N, Martinez-Herrera DJ, Pascual-Montano A, Mittelbrunn M, Sánchez-Madrid F (2013) Sumoylated hnrnpa2b1 controls the sorting of mirnas into exosomes through binding to specific motifs. Nat Commun 4:2980. doi:10.1038/ncomms3980

Vlachogiannis G, Niederhuth CE, Tuna S, Stathopoulou A, Viiri K, de Rooij DG, Jenner RG, Schmitz RJ, Ooi SKT (2015) The dnmt3 l add domain controls cytosine methylation establishment during spermatogenesis. Cell Rep. doi:10.1016/j.celrep.2015.01.021

Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL (2013) Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci USA 110:1012

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y, Sasaki H (2008) Endogenous sirnas from naturally formed dsrnas regulate transcripts in mouse oocytes. Nature 453(7194):539–543. doi:10.1038/nature06908

Watanabe T, Cheng EC, Zhong M, Lin H (2014) Retrotransposons and pseudogenes regulate mrnas and lncrnas via the pirna pathway in the germline. Genome Res. doi:10.1101/gr.180802.114

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, Meachem SJ, Antonarakis SE, de Kretser DM, Hedger MP, Peterson P, Carroll BJ, Scott HS (2005) Meiotic and epigenetic defects in dnmt3l-knockout mouse spermatogenesis. Proc Natl Acad Sci USA 102(11):4068–4073. doi:10.1073/pnas.0500702102

Yamamoto Y, Watanabe T, Hoki Y, Shirane K, Li Y, Ichiiyanagi K, Kuramochi-Miyagawa S, Toyoda A, Fujiyama A, Oginuma M, Suzuki H, Sado T, Nakano T, Sasaki H (2013) Targeted gene silencing in mouse germ cells by insertion of a homologous dna into a pirna generating locus. Genome Res 23(2):292–299. doi:10.1101/gr.137224.112

Yuan A, Farber EL, Rapoport AL, Tejada D, Deniskin R, Akhmedov NB, Farber DB (2009) Transfer of micrornas by embryonic stem cell microvesicles. PLoS ONE 4(3):e4722. doi:10.1371/journal.pone.0004722

Zhou Y, Mishra B (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci USA 102(11):4051–4056. doi:10.1073/pnas.0407957102

Chapter 6

Conclusions and Future Directions

> *"Be patient with yourself. Self-growth is tender; it's holy ground. There's no greater investment."*
>
> — Stephen Covey

Despite being considered "junk" DNA, it has become clear that transposable elements (TEs) are a powerful force in genome evolution. However, methods to identify and annotate TEs have many limitations, and the association of TEs with gene expression evolution is poorly characterized.

In my thesis, I developed an *ab initio* pipeline to identify and annotate TEs across multiple genomes. This method is especially useful for finding recently inserted and active TEs; and it generates not only consensus sequences, but keeps the genomic intervals for the original aligned sequences. This allows straightforward analysis of evolutionary dynamics; segmental duplications were also generated as a useful by-product. My pipeline consistently performs well on both draft and well-annotated genomes. In the analysis of the tuatara genome, I implemented a best reciprocal hits method test for horizontal transfer, this provided evidence for TE transfer between multiple genomes, and can be used for further horizontal transfer analysis. In the analysis of gene expression, multiple statistical approaches were used to minimize the conflation of other co-occurring TEs. This enable us to investigate the association between a specific TE element and gene expression.

One of the primary limitations of this thesis was acquiring enough datasets. It is difficult to acquire gene expression data for all the desired species from one study, especially from lizards and birds. Gene expression data from multiple studies may contain batch effects that can cause artefactual associations for some species. Furthermore, since a large

number of genes contain repeats, it is difficult to find genes that contain only one type of TE, or genes without any TE insertions. Therefore, it is not possible in practice to directly compare genes that contain a specific TE class against genes with no TEs.

Another limitation is the poor nomenclature from the RepBase libraries. I discovered that some L2 sequences were annotated as CR1s, and some CR1 sequences as L2s. This may cause incorrect annotation of consensus sequences generated from my method. My TE annotation and identification is mainly focused on finding recently inserted repeats, meaning it is difficult to find ancestral TEs. Furthermore, all of our genome sequences were generated from Illumina technology, therefore if the longest fragment length library used for mate pairs had many shorter fragments, LINES we found may be chimeric assemblies, rather than a full-length LINE.

There are many areas of this thesis that could be explored in more depth. For example, in the study of the tuatara genome, I found L2 elements are the dominant TE class in its genome, and also their potential for horizontal transfer. However, the mechanisms of how L2s could be transferred between monotremes and tuatara is still unknown; further analysis of L2s can be expanded to more species that are located between the evolution of monotremes and tuatara. This may answer questions about L2 content in other species, such as whether most species with L2s still have active elements or are mainly truncated. The only paper describing full-length L2 structure was published 20 years ago and was from the *fugu* genome, and was subsequently found to be a CR1 element. The uncertainty of L2 structure makes it difficult to understand the L2 transposition machinery. Further research could focus on resolving this.

With regards to gene expression analysis, further work should expand the analysis of gene expression to a broader selection of species, for example ray-finned fish, additional birds, reptiles, and mammals. This would help provide significant insight into genome evolutionary dynamics of complex organisms. Development of paralog analysis could also help to improve the understanding of the association between gene expression and TEs. Because paralogous genes were generated from gene duplication events, they tend to be more species-specific. The combined analysis of both orthologs and paralogs will help to better understand the role of TEs in gene expression and genome evolution.

Collectively, my findings demonstrate that transposable elements play an important role in gene expression during amniote evolution.

Appendix A

Supplementary for Chapter 2

S1 Appendix

bíogo written by: Dan Kortschak
Document organised by: Lu Zeng
February 26, 2018

krishna and igor are *ab initio* repeat family identification and annotation programs, that identify repeat element boundaries and family relationships from whole-genome sequence data. These programs build, refine and classify consensus models of putative interspersed repeats. krishna and igor are built using bíogo (https://github.com/biogo/biogo/), a bioinformatics library for the Go language.
Disclaimer This document is provided to assist researchers with linux command line experience. We have done our best to provide usable instructions, examples and advice, but users assume full responsibility for the output they generate and the authors accept no responsibility for user generated output from any programs or methods listed herein.

For up to date versions of the carp documentation and code, please go to
https://github.com/carp-te.

Prerequisites

Download Go

Available at (https://golang.org/dl).
For installation details, follow the instructions on the Go installation page.

Git

To perform the next step you will need Git to be installed. (Check that you have a git command before proceeding.)

If you do not have a working Git installation, follow the instructions on the Git download page.

Download bíogo Packages

Note: For convenience, add the workspace's bin subdirectory to your PATH, or add in $HOME/.profile.

```
export PATH=$PATH:$(go env GOPATH)/bin
```

Download and install krishna and igor packages from github.

```
go get -u github.com/biogo/examples/krishna
go get github.com/biogo/examples/krishna/matrix
go get github.com/biogo/examples/igor
go get github.com/biogo/examples/igor/seqer
go get github.com/biogo/examples/igor/gffer
```

Install CENSOR

Install censor to screen target genomes against a reference collection of repeats with masking symbols, as well as generating a report classifying all repeats found. CENSOR needs WU-BLAST/NCBI-BLAST and BioPerl installed.

CENSOR, along with instructions for installation, is available at CENSOR download page.

WU-BLAST can be downloaded at WU-BLAST download, NCBI-BLAST can be downloaded at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/legacy.NOTSUPPORTED/2.2.26/. Note that we used NCBI-BLAST legacy code instead of the BLAST+ code because of the CENSOR dependency for that code. BioPerl can be downloaded at BioPerl download.

Install MUSCLE

Install MUSCLE to generate consensus sequences.
MUSCLE is available at MUSCLE download page. The installation information can be seen at MUSCLE Install page.

Example run

In this example, the human genome was downloaded as chromosomes (24 chromosomes) from UCSC into files called chr*.fa.

Use krishna to do pairwise alignment between human genome sequences

krishna-matrix helps you to align sequences by using a matrix table.

The default minimum hit length for krishna (-dplen) is 400bp, and minimum hit identity (-dpid) is 94%. The smaller the length and the lower the hit identity parameters you use, the more time and memory you will need.

If you want to change running parameters, for example, minimum hit length of 200bp, and minimum hit identity of 90%, just specify the parameters when running matrix -krishnaflags="-tmp=/scratch -threads=8 -log -filtid=0.9 -filtlen=200".

Now run the job (the krishna output files end with .gff):

```
cd /your/path/here/human
matrix -threads=8 -krishnaflags="-tmp=./ -threads=2 -log -filtid=0.94
-filtlen=400" chr*.fa
```

-tmp: store the temporary files generated from running krishna, you can specify your own directory.
-threads (matrix): number of concurrent krishna instances to run. (default 6)
-threads (krishnaflags): number of threads to use for alignment. (default 1)
-filtid: minimum hit identity.
-filtlen: minimum hit length.

If genome sequence files are very big and consist of multiple contigs or scaffolds (>200MB), you can use bundle to split them into smaller files. For example,

```
go get github.com/biogo/examples/bundle
bundle -bundle 80000000 -in seq.fa
```

-bundle: specifies the total sequence length in a bundle. (default 20000000, 20MB).
-in: the genomes sequences you need to split.
Then run krishna job.

Use igor to report repeat feature family groupings in JSON format.

After running krishna, igor will take the pairwise alignment data to cluster repeat families.

```
find ./ -maxdepth 1 -name '[!.]*.gff' -print0 | xargs -r0 cat > hg_krishna.gff
igor -in hg_krishna.gff -out hg94_krishna.json
```

Use seqer to generate consensus sequences from genome intervals .

seqer returns multiple fasta sequences corresponding to feature intervals described in the JSON output from igor.

gffer converts the JSON output of igor to gff. seqer will produce fastq consensus sequence output from either MUSCLE or MAFFT.

```
gffer < hg94_krishna.json > hg94_krishna.igor.gff
cat chr*.fa > hg19v37.mfa
seqer -aligner=muscle -dir=consensus -fasta=true -maxFam=100 -subsample=true -minLen=0.95 -threads=12 -ref=hg19v37.mfa hg94_krishna.igor.gff
```

-fasta: Output consensus as fasta with quality case filtering
-maxFam: maxFam indicates maximum family size permitted (0 == no limit).
-minLen: Minimum proportion of longest family member.
-threads: Number of concurrent aligner instances to run.

Benchmarks

Genome	Krishna Threads	Genome DB Size	Krishna run time (hh:mm)	Igor run time (hh:mm)	Seqer run time (hh:mm)
Human	8	3.0G	~200	128:30	2:23
Bearded Dragon	8	1.8G	~23	73:11	<4
Anolis	6	1.8G	76:52	97:32	2:40
Chicken	4	1017M	5	<4	<1
Opossum	8	3.5G	~83	61:48	4:52
Platypus	8	2.0G	~99	191:34	10:16

Analysis runs on a machine with 512GB RAM, running Red Hat Linux.

Repeat Library Annotation

Previous steps have generated repeat consensus sequences from the human genome, now we are going to annotate these repeat consensus sequences.
All the files used below can be found at
(https://data.mendeley.com/datasets/k88h5xnhcb/draft?a=d401233a-5af8-4879-81e8-c049b7133c8c).
All the code used below can be found at
(https://github.com/carp-te/carp-documentation/tree/master/code)**.**

Annotate consensus sequences

Notes: For Java code used here you may need to specify the directories where your input data is and where you want your output written.

Annotate consensus sequences with repeat families.

Use censor to annotate consensus sequences with the Repbase library. The Vertebrates.fa we use here is the Repbase vertebrates repeat libraries downloaded on 1st March, 2016. You can download it from http://www.girinst.org/repbase/update/browse.php?type=All&format=FASTA&autonomous=on&nonautonomous=on&simple=on&division=Vertebrata&letter=A.

```
find ./consensus -maxdepth 1 -name '[!.]*.fq' -print0 | xargs -r0 cat >
 ConsensusSequences.fa

censor -bprm cpus=8 -lib ~/Vertebrates.fa -lib ~/our_known_reps_20130520.fasta
 ConsensusSequences.fa
```

For people that are not able to access WU-BLAST or prefer to use another aligner, CENSOR can also use NCBI BLAST instead (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), please find details for CENSOR installation at CENSOR download page.

```
censor.ncbi -lib Vertebrates.fa -lib our_known_reps_20130520.fasta
 ConsensusSequences.fa
```

RepeatMasker can also be used to replace CENSOR in this step (RM-Blast was used as search engine).

```
cat Vertebrates.fa our_known_reps_20130520 > combined_library.fa

RepeatMasker -pa 16 -a -nolow -norna -dir ./ -lib combine_library.fa
 ConsensusSequences.fa

perl format_RMSK.pl ConsensusSequences.fa > tmp

mv tmp ConsensusSequences.fa
```

The censor output usually contains 5 files: ConsensusSequences.fa.map, ConsensusSequences.fa.aln, ConsensusSequences.fa.found, ConsensusSequences.fa.idx, ConsensusSequences.fa.masked.

Classify consensus sequences.

ConsensusSequences.fa and ConsensusSequences.fa.map are required in this step. You will also need to specify the directories for your input data and where you want your output written in the java code. Edit the source, compile and run.

```
javac ClassifyConsensusSequences.java
java ClassifyConsensusSequences
```

This should generate 5 output files: known.txt, partial.txt, check.txt, notKnown.fa, notknown.fa.gff. Then we need to further annotate these notKnown.fa consensus sequences.

Filter sequences.

This step contains three parts: 1) Identify potential protein sequences; 2) Identify GB_TE sequences; 3) Identify retrovirus sequences. You can run each part separately in parallel to save time. From these three steps, you will get three output files for following steps: 1) notKnown.fa.spwb.gff, 2) notKnown.fa.tewb.gff, 3) notKnown.fa.ervwb.gff.

First download uniprot protein dataset (uniprot_sprot.fasta.gz) from
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz.

Then download GB_TE dataset. First install EDirect from
EDirect download page. Then download the data:

```
esearch -db protein -query "reverse transcriptase or transposon or repetitive element
    or RNA-directed DNA polymerase or pol protein or non-LTR retrotransposon or mobile
     element or retroelement or polyprotein or retrovirus or (group-specific antigen
    gag) or polymerase (pol)" | efetch -format fasta > 260118_GB_TE.fa
```

Retrovirus datasets can be downloaded from https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=11632.

Here are examples using WU-BLAST:
1) Identify potential protein sequences

```
gzip -d uniprot_sprot.fasta.gz

xdformat -p -k uniprot_sprot.fasta

blastx ./report_run/sprot notKnown.fa  -gspmax=1 -E 0.00001 -B 1 -V 1 -cpus=32 >
    notKnown.fa.spwb

python ./report_run/wublastx2gff.py notKnown.fa.spwb > notKnown.fa.spwb.gff
```

2) Identify GB_TE sequences

```
xdformat -p -k GB_TE.21032016.fa -o GB_TE.new

blastx ./BlastDB/GB_TE.new notKnown.fa -gspmax=1 -E 0.00001 -B 1 -V 1 -cpus=32 >
    notKnown.fa.tewb

python ./report_run/wublastx2gff.py notKnown.fa.tewb > notKnown.fa.tewb.gff
```

3) Identify potential retrovirus sequences

```
xdformat -n -k all_retrovirus.fasta

tblastx ./BlastDB/all_retrovirus.fasta notKnown.fa -gspmax=1 -E 0.00001 -B 1 -V 1 \
-cpus=32 > notKnown.fa.ervwb

python ./report_run/wublastx2gff.py notKnown.fa.ervwb > notKnown.fa.ervwb.gff
```

-gspmax: max. number of gapped HSPs (GSPs) saved per subject sequence (default 0; 0 => unlimited).
-B -V: the B and V options limit the number of subject sequences for which any results whatsoever are reported, regardless of the number of HSPs or GSPs found.
-E: Expectation value (E) threshold for saving hits.
-cpus: no. of processors to utilize on multi-processor systems.

If you wish to use NCBI-BLAST, we tested different NCBI-BLAST parameters, to make the results as consistent as possible to the WU-BLAST results. See below for examples using NCBI-BLAST:

1) Identify potential protein sequences

```
makeblastdb -in uniprot_sprot.fasta -dbtype prot

blastx -db uniprot_sprot.fasta -query notKnown.fa -max_hsps 1 -seg no -evalue 0.00001
    -num_threads 32 -max_target_seqs 1 -word_size 2 -outfmt 6 -out notKnown.fa.spwb.
    ncbi

awk '{print $1"\t""blast""\t""hit""\t"$7"\t"$8"\t"$11"\t"."\t"."\t""Target sp|"
$2" "$9" "$10}' notKnown.fa.spwb.ncbi > tmp

awk '{if($4>$5) print $1"\t"$2"\t"$3"\t"$5"\t"$4"\t"$6"\t"$7"\t"$8"\t"$9" "$10" "$11
" "$12; else print $0}' tmp > notKnown.fa.spwb.gff
```

2) Identify GB_TE sequences

```
makeblastdb -in GB_TE.21032016.fa -dbtype prot -out GB_TE.new

blastx -db GB_TE.new -query notKnown.fa -max_hsps 1 -seg no -evalue 0.00001 \
-num_threads 32 -max_target_seqs 1 -word_size 2 -outfmt 6 -out notKnown.fa.tewb.ncbi

awk '{print $1"\t""blast""\t""hit""\t"$7"\t"$8"\t"$11"\t"."\t"."\t""Target sp|"
$2" "$9" "$10}' notKnown.fa.ervwb.ncbi > tmp

awk '{if($4>$5) print $1"\t"$2"\t"$3"\t"$5"\t"$4"\t"$6"\t"$7"\t"$8"\t"$9" "$10" "$11
" "$12; else print $0}' tmp > notKnown.fa.tewb.gff
```

3) Identify potential retrovirus sequences

```
makeblastdb -in all_retrovirus.fasta -dbtype nucl

tblastx -db all_retrovirus.fasta -query notKnown.fa -max_hsps 1 -seg no -evalue
    0.00001 -num_threads 32 -max_target_seqs 1 -word_size 2 -outfmt 6 -out notKnown.fa
    .ervwb.ncbi

awk '{print $1"\t""blast""\t""hit""\t"$7"\t"$8"\t"$11"\t"."\t"."\t""Target sp|"
$2" "$9" "$10}' notKnown.fa.ervwb.ncbi > tmp

awk '{if($4>$5) print $1"\t"$2"\t"$3"\t"$5"\t"$4"\t"$6"\t"$7"\t"$8"\t"$9" "$10" "$11
" "$12; else print $0}' tmp > notKnown.fa.ervwb.gff
```

-db: BLAST database name
-query: Input file name
-max_hsps: Set maximum number of HSPs per subject sequence to save for each query, NCBI-BLAST doesn't have gsps option, but we tested with hsps in WU-BLAST, the result remain almost the same
-seg: Filter query sequence with SEG (Format: 'yes', 'window locut hicut', or 'no' to disable), default WU-BLAST is off
-evalue: Expectation value (E) threshold for saving hits
-num_threads: Number of threads (CPUs) to use in the BLAST search
-max_target_seqs: Maximum number of aligned sequences to keep
-word_size: Word size for wordfinder algorithm.

Get protein information from consensus sequences.

Another java program GetProteins.java will be used. It needs two input files: notKnown.fa, notKnown.fa.spwb.gff (Generated from previous step).

```
javac GetProteins.java
java GetProteins
```

You will get 2 output files: proteins.txt (a list of families that have been identified as proteins and the proteins they match); notKnownNotProtein.fa (a fasta file of the families that were not classified).

Check for simple sequence repeats (SSR).

Check for existence of SSR in the unknown sequences, using phobos. Phobos can be downloaded at (http://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm). We used executable: phobos-linux-gcc4.1.2.

```
phobos-linux-gcc4.1.2 -r 7 --outputFormat 0 --printRepeatSeqMode 0
 notKnownNotProtein.fa > notKnownNotProtein.phobos
```

Identify the sequences that are SSRs from the phobos output.

phobos output will be used to identify SSRs: notKnownNotProtein.phobos.

```
javac IdentifySSRs.java
java IdentifySSRs
```

Your output will be a file called: SSR.txt

Generate annotated repeat library.

There are ten input files that are required to generate a repeat library:
1. ConsensusSequences.fa
2. ConsensusSequences.fa.map
3. notKnown.fa.tewb.gff
4. notKnown.fa.ervwb.gff
5. protein.txt
6. known.txt
7. GB_TE.21032016.fa
8. all_retrovirus.fasta
9. SSR.txt (if you do not have this, leave the definition in, it will generate error messages, but will not stop the program or affect the results.)
10. LA4v2-satellite.fa (you do not have this, or equivalent, as you didn't have any satellites, but leave the definition in-it will cause error messages, but will not stop the program or affect the results.)

```
javac GenerateAnnotatedLibrary.java
java GenerateAnnotatedLibrary
```

This will generate a library called "Human_Repeat_Library.fasta", you can rename this file to whatever you want.

Benchmarks2

Genome	Consensus sequences size	Censor first run time (hh:mm)	reportJ.sh (hh:mm)	phobos run time (hh:mm)
Human	38M	5:14	19:30	<00:10
Bearded Dragon	88M	22:21	<178	<00:10
* New Bearded Dragon	88M	7:13	18:28	<00:10
Anolis	63M	9:42	78	<00:10
Chicken	18M	3:01	<24	<00:10
Opossum	60M	13:17	80	<01:00
Platypus	162M	17:34	115	<01:00

Analysis run on a slurm machine with 4∼16 cpus, and 8GB RAM, running Red Hat Linux.
* New Bearded dragon analysis used same bearded dragon genome, except it was run on a High Performance Computing machine with 32 cpus, running Red Hat Linux.

Supplementary Figures

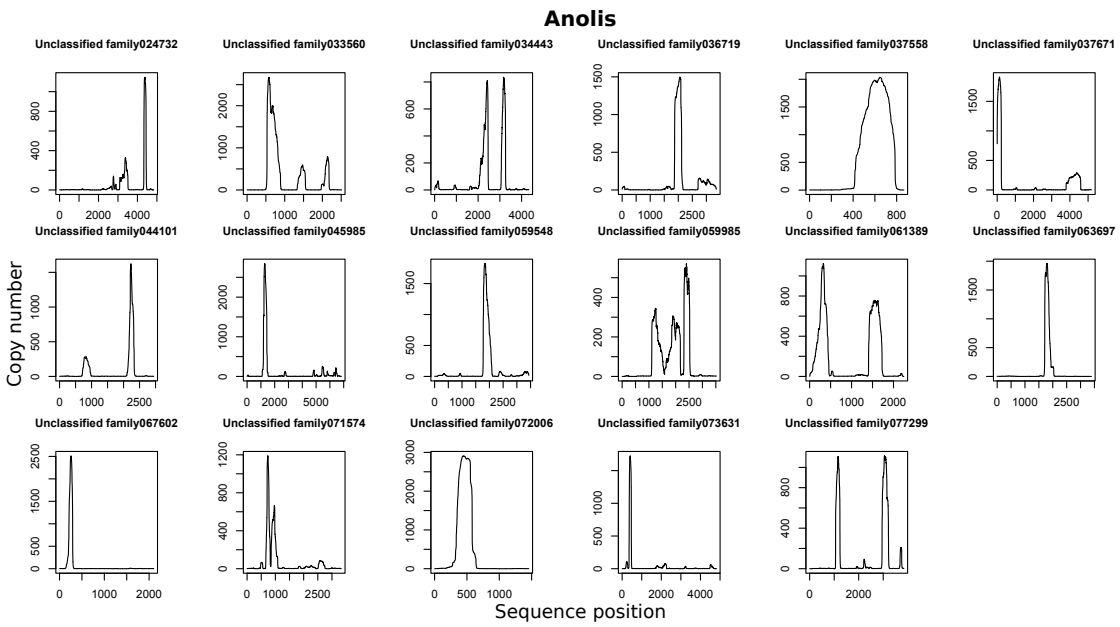


Figure S1: **Coverage plot of high copy number unclassified repeats in the anolis genome.** Shows the coverage plot for the top 12 highest copy number (>2,000 copies) unclassified consensus sequences in the anole genome.

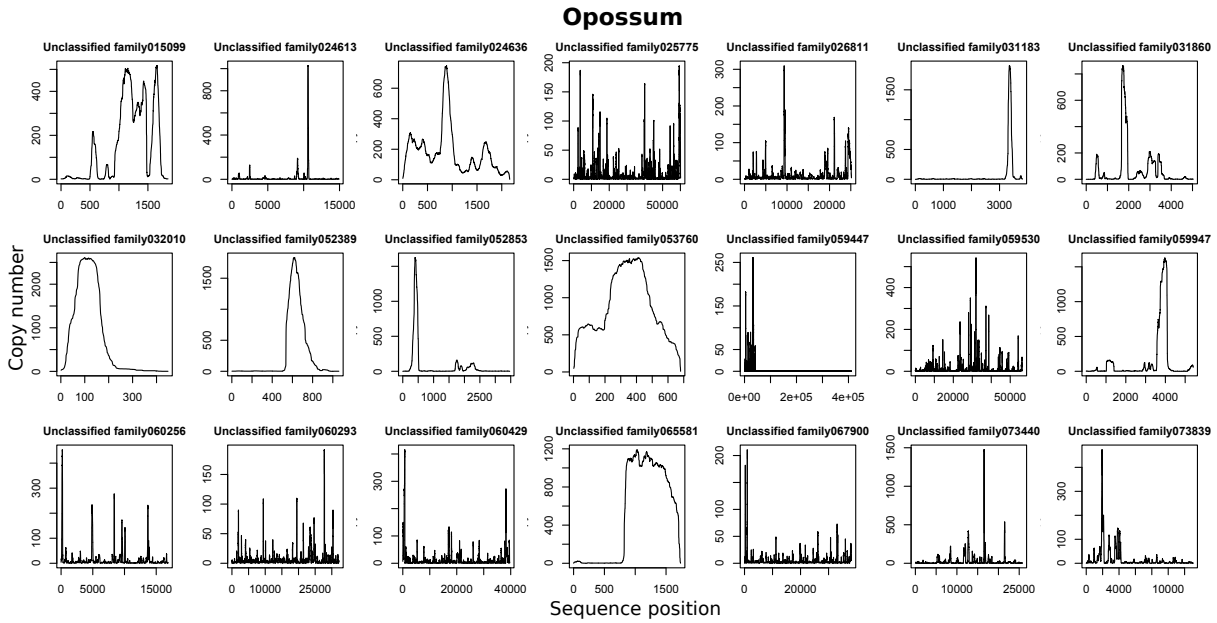


Figure S2: **Coverage plot of high copy number unclassified repeats in the opossum genome.** Shows the coverage plot for the top 21 highest copy number (>2,000 copies) unclassified consensus sequences in the opossum genome.

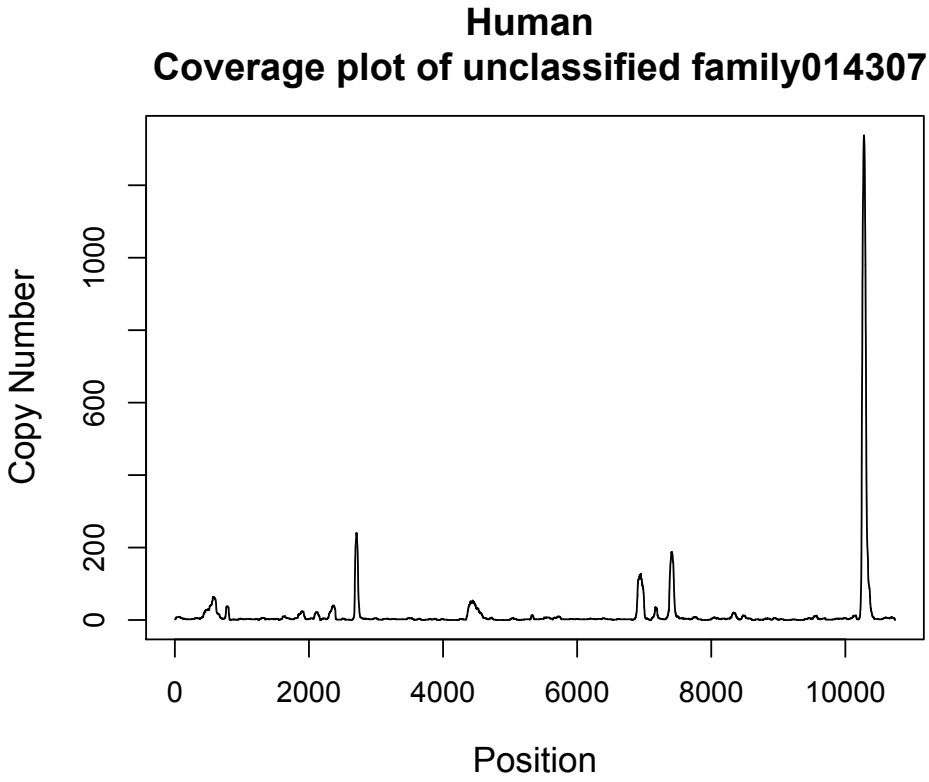


Figure S3: **Coverage plot of high copy number unclassified repeats in the human genome.** Shows the coverage plot for the top 1 highest copy number (>2,000 copies) unclassified consensus sequences in the human genome.

Supplementary Tables

Table S1: **Genome dataset**. Shows the systematic name, common name, genome version, source and submitter for all the genomes tested with our *ab initio* method.
The Following abbreviations are used for submitters:
Genome Sequencing Platform, The Genome Assembly Team = GAT;
Genome Reference Consortium = GRC;
International Chicken Genome Consortium = ICGS;
Washington University = WashU.

No	Systematic Name	Common Name	Genome Version	Source	Submitter
1	*Homo sapiens*	Human	GRCh37(hg19)	NCBI	GRC
2	*Central Pogona Vitticeps*	Bearded Dragon	Pogona_vitticeps.male	NCBI	BRAEMBL
3	*Anolis Carollinensis*	Anolis lizard	PanoCar2	NCBI	Broad
4	*Gallus gallus*	Chicken	galGal4	NCBI	ICGS
5	*Monodelphis domestica*	Opossum	monDom5	NCBI	GAT
6	*Ornithorhynchus anatinus*	Platypus	ornAna1	NCBI	WashU

Table S2: **Assembly statistics**. Shows the systematic name, total sequence length (i.e. genome size, including bases and gaps), scaffold N50 (i.e. scaffold length at which 50% of the total bases in the assembly are in scaffolds of that length or greater), contig N50 and assembly level (complete genome, chromosome, scaffold or contig). Species are listed in the same order as Table 1. However, the information is incomplete for some genomes because the NCBI Assembly database only contains the information provided by the submitters.

No	Species	Total Sequence Length	Scaffold N50	Contig N50	Assemble Level
1	*Homo sapiens*	3,095,677,412	44,983,201	38,440,852	Chromosome
2	*Pogona Vitticeps*	1,716,675,060	2,477,614	-	Scaffold
3	*Anolis Carollinensis*	1,799,143,587	4,033,265	79,867	Chromosome
4	*Gallus gallus*	1,046,932,099	12,877,381	279,750	Chromosome
5	*Monodelphis domestica*	3,605,631,728	59,809,810	108,014	Chromosome
6	*Ornithorhynchus anatinus*	2,073,148,626	958,970	11,554	Chromosome

Table S3: **Assembly method and coverage.**Shows the systematic name, assembly method, sequencing technology and estimated genome coverage for the six genomes in this study. Species are listed in the same order as Table 1.

No	Species	Assembly Method	Sequencing Technology	Genome Coverage
1	*Homo sapiens*	Celera	Sanger	20x
2	*Pogona Vitticeps*	SOAP deNovo	Illumina HiSeq 2000	85.5x
3	*Anolis Carollinensis*	Arachne v.3.0.0	ABI	7.10x
4	*Gallus gallus*	Celera Assemblerv.5.4	Sanger; 454	12x
5	*Monodelphis domestica*	ARACHNE2+	Sanger	6.8x
6	*Ornithorhynchus anatinus*	PCAP	WGS plasmid, fosmid end and BAC end sequences	6x

Table S4: **Benchmarks for each method.** Here we show the compute time used for the seven tested species with CARP and RMD.

	Chicken		**Bearded dragon**		**Anolis**		**Platypus**		**Opossum**		**Human**	
	De novo	RMD	*De novo*	RMD	*De novo*	RMD	*De novo*	RMD	*De novo*	RMD	*De novo*	RMD
Time consumed (h)	37	8	276	39	266	19	434	49	244	21	495	18

Table S5: **Summary of library lengths generated from two methods.** Total length (bp) of consensus sequence libraries generated by CARP and RMD.

	Chicken		**Bearded dragon**		**Anolis**		**Platypus**		**Opossum**		**Human**	
	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD
Well -annotated	4,100,530	154,743	6,758,949	396,987	21,498,947	564,031	4,022,323	204,475	28,379,922	651,545	10,914,900	462,921
Unclassified	10,871,250	20,553	42,128,582	220,612	34,790,047	166,735	142,361,947	40,446	25,672,894	66,497	20,779,056	2,018
Total	14,971,780	175,296	48,887,531	617,599	56,288,994	730,766	146,384,270	244,921	54,052,816	718,042	31,693,956	464,939

Table S6: **Summary of specific TE length generated from two methods.** Comparison of the total consensus sequence lengths (bp) of specific TE types generated by CARP and RMD.

	Chicken		**Bearded dragon**		**Anolis**		**Platypus**		**Opossum**		**Human**	
	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD	CARP	RMD
SINE	0	542	438,341	13,932	705,660	14,210	620,446	42,051	217,725	43,274	3,516,601	21,660
LINE	2,504,529	71,746	5,710,722	217,145	13,287,702	288,130	2,803,327	101,511	17,085,608	364,368	5,115,349	249,096
LTR	1,364,464	56,279	125,606	78,822	4,034,353	121,054	170,087	24,980	10,832,916	212,088	2,028,463	134,103
DNA	31,724	9,635	480,443	79,404	3,461,618	120,828	20,525	14,770	230,518	25,880	76,177	42,308
Others	199,813	16,541	3,837	7,684	9,614	19,809	407,938	21,163	13,155	5,935	178,310	15,754
Total	4,100,530	154,743	6,758,949	396,987	21,498,947	564,031	4,022,323	163,123	28,379,922	651,545	10,914,900	462,921

Table S7: **Repeat content of the chicken genome.** Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the chicken genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
CR1	263,953	75,352,619	7.197
Tx1	19,284	1,282,829	0.123
LINE L2	10,268	991,208	0.094
SINE	9,766	957,771	0.092
Others	46,883	3,327,936	0.318
	350,154	81,912,363	7.824
DNA transposons			
hAT	64,915	5,762,565	0.550
Mariner	35,785	5,210,866	0.498
Charlia	9,267	3,941,040	0.376
DNA	37,393	3,612,850	0.345
Others	120,170	8,667,690	0.828
	267,530	27,195,011	2.597
LTR			
Copia	139,422	9,641,970	0.921
Gypsy	32,502	2,180,147	0.208
BEL	23,661	1,450,356	0.138
Others	26,536	3,756,055	0.359
	222,121	17,028,528	1.626
ERVs			
GGLTR	18,844	7,093,568	0.677
Others	72,068	11,146,021	1.065
	90,912	18,239,589	1.742
SSR	38,969	4,172,125	0.399
Others	112,310	11,734,743	1.121
Well-annotated	1,081,996	160,282,359	15.309
Unknown	84,926	20,806,411	1.987
Total	1,166,922	181,088,770	17.296

Table S8: **Repeat content of the anolis genome.** Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the anolis genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
CR1	206,414	58,082,752	3.228
LINE L2	282,443	75,470,209	4.195
LINE L1	159,992	35,267,057	1.960
RTE	97,078	32,188,630	1.789
R4	65,406	24,750,832	1.376
Others	184,317	37,754,227	2.098
	995,650	263,513,707	14.646
SINEs			
SINE-2	344,568	71,189,836	3.957
Others	9,136	1,110,966	0.062
	353,704	72,300,802	4.019
DNA transposons			
hAT	526,130	77,476,809	4.307
Mariner	409,901	67,917,600	3.775
Helitron	179,050	35,682,369	1.983
DNA	300,556	39,263,225	2.182
Others	112,203	10,576,832	0.588
	1,527,840	230,916,835	12.835
LTR			
Gypsy	241,711	67,543,777	3.754
Copia	53,434	8,461,466	0.470
DIRS	24,293	9,446,400	0.525
Others	43,490	9,248,301	0.514
	362,928	94,699,944	5.263
ERVs			
ERV1/2/3	79,860	12,977,630	0.720
SSR	72,160	8,040,704	0.447
Others	250,255	19,205,549	1.067
Well-annotated	3,642,397	701,655,171	38.997
Unknown	1,678,936	216,304,228	12.022
Total	5,321,333	917,959,399	51.019

Table S9: **Repeat content of the bearded dragon genome**. Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the bearded dragon genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
LINE-2	229,080	40,358,815	2.351
RTE(BovB)	222,363	57,590,102	3.355
RTE	124,644	26,011,403	1.515
LINE(CR1)	194,333	35,912,124	2.092
LINE-1	126,975	15,181,332	0.884
Penelope	95,133	11,527,155	0.671
Other	96,331	12,761,752	0.744
	1,088,859	199,342,683	11.612
SINEs			
SINE-2	206,966	28,277,402	1.647
Other	17,391	1,252,811	0.073
	224,357	29,530,213	1.720
DNA transposons			
Mariner	201,988	22,220,773	1.294
hAT	260,256	23,832,751	1.388
others	215,581	14,924,598	0.870
	677,825	60,978,122	3.552
LTR			
DIRS	62,273	11,501,289	0.670
Gypsy	207,875	15,163,491	0.883
Copia	74,482	5,108,316	0.298
other	45,829	2,782,279	0.162
	390,459	34,555,375	2.013
ERVs			
ERV1/2/3	99,037	6,371,060	0.371
SSR	83,878	8,278,572	0.482
Others	217,184	15,519,668	0.905
Well-annotated	2,781,599	354,575,693	20.655
Unknown	2,882,556	381,401,912	22.217
Total	5,664,155	735,977,605	42.872

Table S10: **Repeat content of the platypus genome.** Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the platypus genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
LINE L2	2,141,746	380,315,682	18.345
CR1	105,713	16,510,780	0.796
BovB	14,412	3,793,080	0.183
Others	156,741	22,358,028	1.078
	2,418,612	422,977,570	20.402
SINEs			
Mon1	2,039,591	376,397,912	18.156
PlatSINE	49,237	11,920,732	0.575
Others	152,447	16,041,435	0.774
	2,241,275	404,360,079	19.505
DNA transposons			
hAT	105,794	10,839,474	0.523
Mariner	90,510	12,440,134	0.600
Others	163,790	12,469,529	0.601
	360,094	35,749,137	1.724
LTR			
Copia	35,510	2,515,002	0.121
Gypsy	150,673	10,227,632	0.493
Other	58,994	6,189,423	0.299
	245,177	18,932,057	0.913
ERVs			
ERV	85,721	8,813,118	0.425
SSR	107,515	13,599,686	0.656
Others	364,116	46,404,979	2.238
Well-annotated	5,822,510	950,836,626	45.863
Unknown	2,321,717	261,289,414	12.603
Total	8,144,227	1,212,126,040	58.466

Table S11: **Repeat content of the opossum genome.** Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the opossum genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
LINE L1	1,252,041	703,738,862	19.518
CR1	1,096,864	225,839,324	6.264
RTE	220,608	73,141,920	2.028
LINE2	50865	5042973	0.140
Others	127,564	9,278,127	0.257
	2,747,942	1,017,041,206	28.206
SINEs			
SINE-1	506,241	88,145,053	2.444
SINE MIR	547,244	67,502,440	1.872
THER	597,652	89,262,998	2.476
Others	746,517	131,606,069	3.650
	2,397,654	376,516,560	10.442
DNA transposons			
hAT	350,343	44,822,893	1.243
Mariner	157,450	23,453,500	0.651
Charlia	17,487	4,683,795	0.130
Others	439,658	37,393,312	1.037
	964,938	110,353,500	3.061
LTR			
Copia	60,653	4,045,526	0.112
Gypsy	289,603	20,968,390	0.582
Others	100,662	9,982,015	0.277
	450,918	34,995,931	0.971
ERVs			
ERV	772,823	347,807,695	9.646
SSR	245,921	34,883,702	0.967
Others	302,420	27,218,434	0.755
Well-annotated	7,882,616	1,948,817,028	54.049
Unknown	1,164,500	141,831,766	3.934
Total	9,047,116	2,090,648,794	57.983

Table S 12: **Repeat content of the human genome.** Shows the copy number, total base pairs (bp) and the percentage of specific repeat class in the human genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
LINE L1	1,061,429	504,783,605	16.306
CR1	348,909	65,998,881	2.132
LINE L2	31,074	2,741,726	0.089
Others	135,468	11,057,731	0.357
	1,576,880	584,581,943	18.884
SINEs			
Alu	1,148,493	285,614,276	9.226
MIR	367,336	49,924,630	1.613
Others	91,907	12,580,275	0.406
	1,607,736	348,119,181	11.245
DNA transposons			
hAT	408,174	60,848,627	1.966
Mariner	191,198	42,157,411	1.362
Others	386,094	36,070,514	1.165
	985,466	13,907,6552	4.493
LTR			
Gypsy	279,091	20,266,383	0.655
THE1	10,979	11,686,418	0.377
Copia	55,336	3,810,198	0.123
Others	90,172	11,450,442	0.370
	435,578	47,213,441	1.525
ERVs			
ERV	802,941	246,168,345	7.952
SSR	147,561	18,249,839	0.590
Others	294,098	34,097,043	1.101
Well-annotated	5,850,260	1,417,506,344	45.790
Unknown	762,391	91,054,381	2.941
Total	6,612,651	1,508,560,725	48.731

Table S13: **Linear regression for unknown sequence copy number against length.** Shows the estimate value, standard error, t-value, p-value and significance codes from linear regression analysis. Significance asterisks follow the conventions of R, i.e. ***, **, *, ., for p-values below 0.001, 0.01, 0.05 and 0.1 respectively.

		Coefficient	Estimate	Std. Error	T	P	
Chicken	*de novo*	Intercept	-1.055	0.037	-28.49	<2e-16	***
		Length	0.689	0.013	53.44	<2e-16	***
	RMD	Intercept	-1.690	0.424	-3.911	0.000148	***
		Length	1.226	0.160	7.653	4.18e-12	***
Bearded Dragon	*de novo*	Intercept	-1.690	0.016	-105.2	<2e-16	***
		Length	1.035	0.006	164.5	<2e-16	***
	RMD	Intercept	1.264	0.123	10.27	<2e-16	***
		Length	0.713	0.053	13.54	<2e-16	***
Anolis	*de novo*	Intercept	-1.822	0.030	-60.40	<2e-16	***
		Length	1.005	0.011	92.78	<2e-16	***
	RMD	Intercept	1.147	0.178	6.435	1.75e-10	***
		Length	0.501	0.074	6.753	2.20e-11	***
Platypus	*de novo*	Intercept	-2.156	0.009	-245.0	<2e-16	***
		Length	1.098	0.003	360.7	<2e-16	***
	RMD	Intercept	-0.329	0.218	-1.51	0.132	
		Length	1.026	0.092	11.10	<2e-16	***
Echidna	*de novo*	Intercept	-1.947	0.015	-126.8	<2e-16	***
		Length	1.104	0.006	194.0	<2e-16	***
	RMD	Intercept	-0.518	0.214	-2.422	0.0159	*
		Length	1.191	0.092	12.931	<2e-16	***
Opossum	*de novo*	Intercept	-0.867	0.032	-26.75	<2e-16	***
		Length	0.700	0.011	61.49	<2e-16	***
	RMD	Intercept	-0.709	0.286	-2.477	0.0135	*
		Length	0.974	0.114	8.526	<2e-16	***
Human	*de novo*	Intercept	-1.733	0.021	-80.7	<2e-16	***
		Length	1.047	0.008	135.1	<2e-16	***
	RMD	Intercept	-0.290	0.248	-1.168	0.244	
		Length	0.617	0.101	6.122	2.05e-09	***

Appendix B

Additional analysis of the echidna genome

Analysis of the echidna genome

Additional analysis have been done on the echidna genome by using CARP, as the echidna genome is still unpublished yet, we could not submit this part to journal.

Supplementary Figures

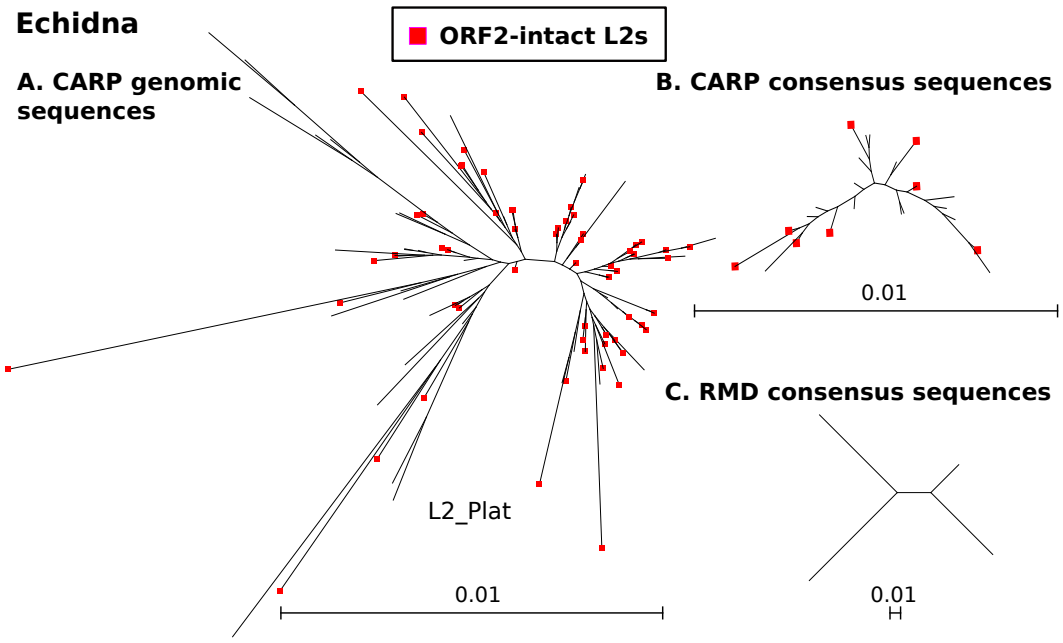


Figure S1: **Phylogenetic analysis of L2 elements in the echidna genome.**
Figure shows the dendrograms of full-length L2 elements in the echidna genome. Panel A) long L2 sequences from the echidna genome. Panel B) Long L2 CARP consensus sequences from echidna. Panel C) Long L2 RMD consensus sequences from echidna. Sequences were aligned with MUSCLE, trees inferred with FastTree and visualized with Archaeopteryx. ORF2-instact L2s are shown with a red dot at the tip of the branch.

Supplementary Tables

Table S1: **Echidna assembly dataset**. Shows the systematic name, common name, genome version, source and submitter for the echidna genome tested with our *ab initio* method. Genome that were acquired through private collaboration (not publicly available) are marked as 'Private' in the Submitter column.

Systematic Name	Common Name	Genome Version	Source	Submitter
Tachyglossus aculeatus	Echidna	Tachyglossus	University of Copenhagen Prof Guojie Zhang	private

Table S2: **Echidna assembly statistics**. Shows the systematic name, total sequence length (i.e. genome size, including bases and gaps), scaffold N50 (i.e. scaffold length at which 50% of the total bases in the assembly are in scaffolds of that length or greater), contig N50 and assembly level.

Species	Total Sequence Length	Scaffold N50	Contig N50	Assemble Level
Tachyglossus aculeatus	1,936,662,216	775,344	-	Scaffold

Table S3: Benchmarks for total time consumption by the echidna genome.

	Echidna	
	De novo	RMD
Time consumed (h)	603	74

Table S4: **Comparison of the total number of specific TE types in each method.** Total number of specific TE types generated by CARP and RMD in the echidna genome. Total length (bp) of specific TE types were shown in parentheses.

	Echidna	
	CARP	RMD
SINE	2,954 (793,326)	26 (23,377)
LINE	10,275 (3,944,588)	108 (97,989)
LTR	67 (27,099)	15 (14,270)
DNA	31 (11,117)	36 (21,725)
Others	106 (46,971)	7 (5,762)
Total	13,433 (4,823,101)	192 (163,123)

Table S5: **Repeat content of the echidna genome.** Shows the copy number, total base pairs (bp) and the percentage of each repeat class in the echidna genome.

Group	Copy number	Total bp	Percentage coverage of genome
Non-LTR retrotransposons			
LINEs			
LINE L2	2,519,577	443,411,531	22.896
CR1	92,892	14,209,630	0.734
BovB	11,685	2,728,940	0.141
Others	123,382	19,105,195	0.986
	2,747,536	479,455,296	24.757
SINEs			
Mon1	1,959,322	349,937,870	18.069
Others	210,469	25,579,379	1.321
	2,169,791	375,517,249	19.390
DNA transposons			
hAT	82062	7937802	0.410
Mariner	78,168	10,507,276	0.542
DNA	34,852	2,548,860	0.132
Others	79,409	5,651,272	0.292
	274,491	26,645,210	1.376
LTR			
Copia	27,716	1,712,219	0.088
Gypsy	123,377	8,275,379	0.427
Others	42,595	3,166,608	0.164
	193,688	13,154,206	0.679
ERVs			
ERV	81,994	6,721,751	0.347
SSR	122,915	17,085,455	0.882
Others	293,565	27,183,445	1.404
Well-annotated	5,883,980	945,762,612	48.835
Unknown	1,049,511	93,141,550	4.809
Total	6,933,491	1,038,904,162	53.644

Appendix C

Supplementary for Chapter 4

Supplementary Figures

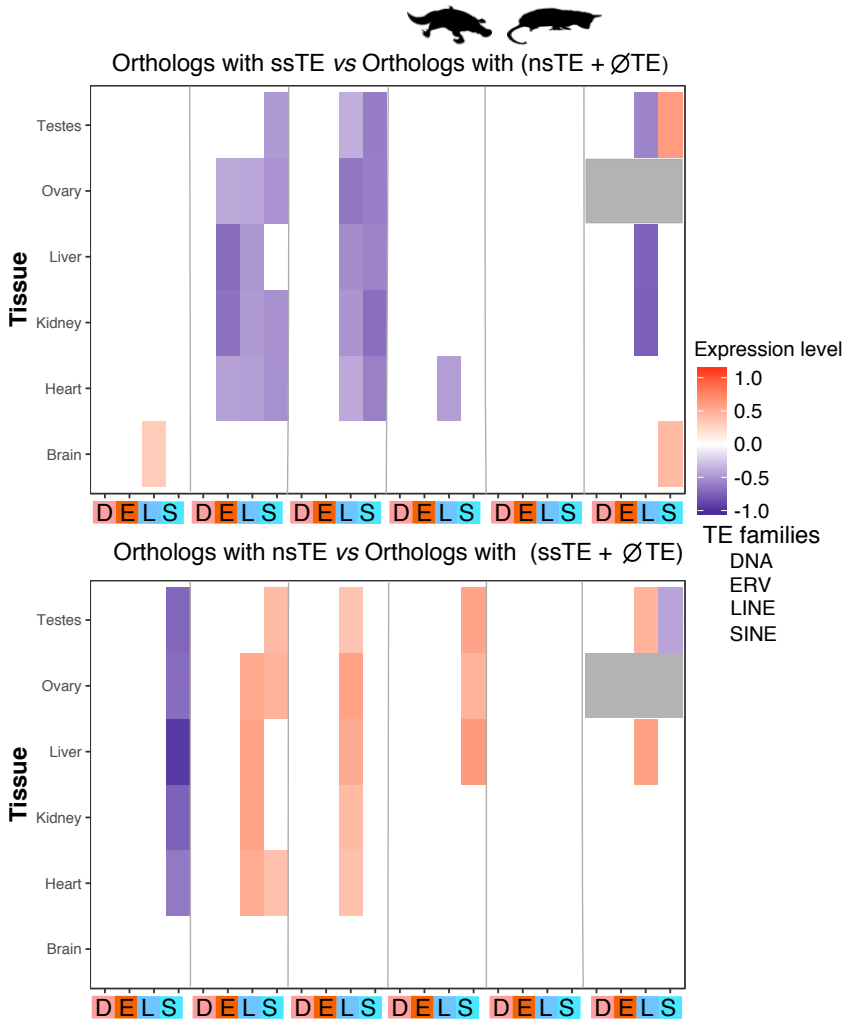


Figure S1: **Change in the levels of ortholog gene expression as a function of TE insertion.** This figure shows the association between ortholog gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon (pogona), platypus, opossum and human) with recent species-specific TE insertions (ssTE) or non-recent species specific TE insertions (nsTE) (from left to right: DNA, ERV/LTR, LINE or SINE). A weighted bootstrap approach was used to compare the median gene expression levels of orthologs with a ssTE/nsTE insertion compared to orthologs without ssTE/nsTE. Gene expression levels are log2-transformed. Comparisons without statistically significant gene expression changes are shown in white. Statistically significant increased gene expression shown in red and statistically significant decreased gene expression in blue. Grey shading indicates no samples were available for this comparison.

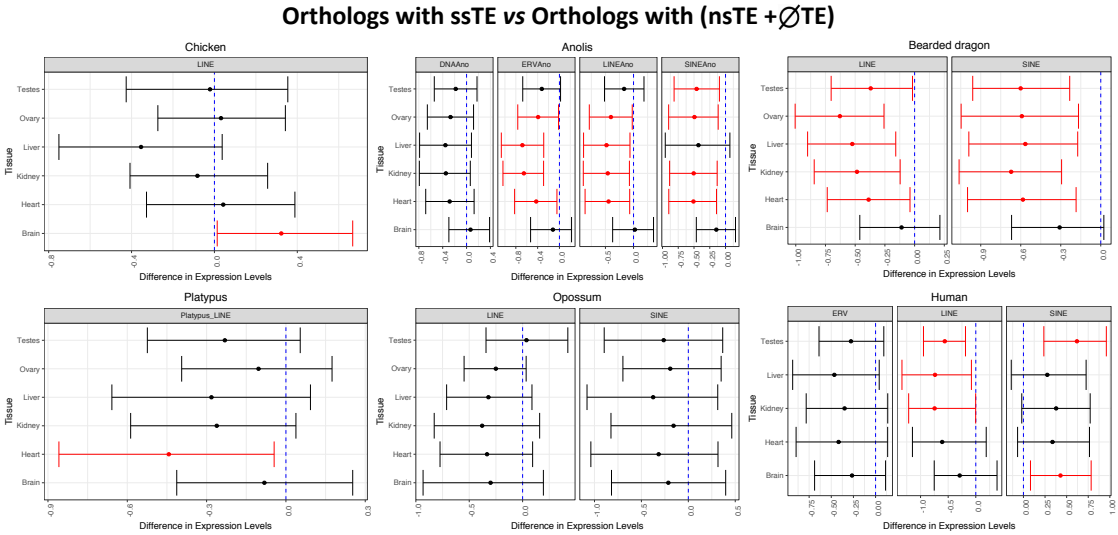


Figure S2: **Change in the levels of ortholog gene expression as a function of species-specific TE insertion.**
This figure shows the association between ortholog median gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon (pogona), platypus, opossum and human) with recent species-specific TE insertions (ssTE) (from left to right: DNA, ERV/LTR, LINE or SINE). Confidence Intervals for the difference in median log2(TPM) counts. Confidence Intervals were obtained using the weighted bootstrap and are 1-α/m intervals, where α=0.05 and m=nTissues*nElements as the total number of intervals presented. Red dots represent the median value from the bootstrap procedure, whilst the vertical line indicates zero. Intervals which do not contain zero are coloured red, and indicate a rejection of the null hypothesis, H_0: $\Delta\theta$=0, where θ represents the parameter of interest.

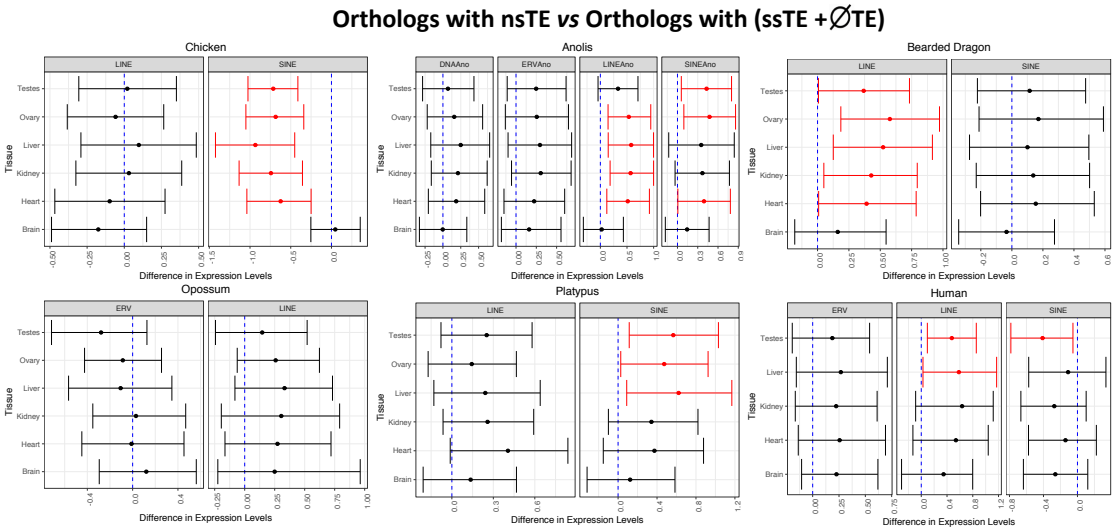


Figure S3: **Change in the levels of ortholog gene expression as a function of non-species specific TE insertion.**
This figure shows the association between ortholog gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon (pogona), platypus, opossum and human) with non-recent species-specific TE insertions (nsTE) (from left to right: DNA, ERV/LTR, LINE or SINE). Confidence Intervals for the difference in median log2(TPM) counts. Confidence Intervals were obtained using the weighted bootstrap and are 1-α/m intervals, where α=0.05 and m=nTissues*nElements as the total number of intervals presented. Red dots represent the median value from the bootstrap procedure, whilst the vertical line indicates zero. Intervals which do not contain zero are coloured red, and indicate a rejection of the null hypothesis, H_0: $\Delta\theta$=0, where θ represents the parameter of interest.

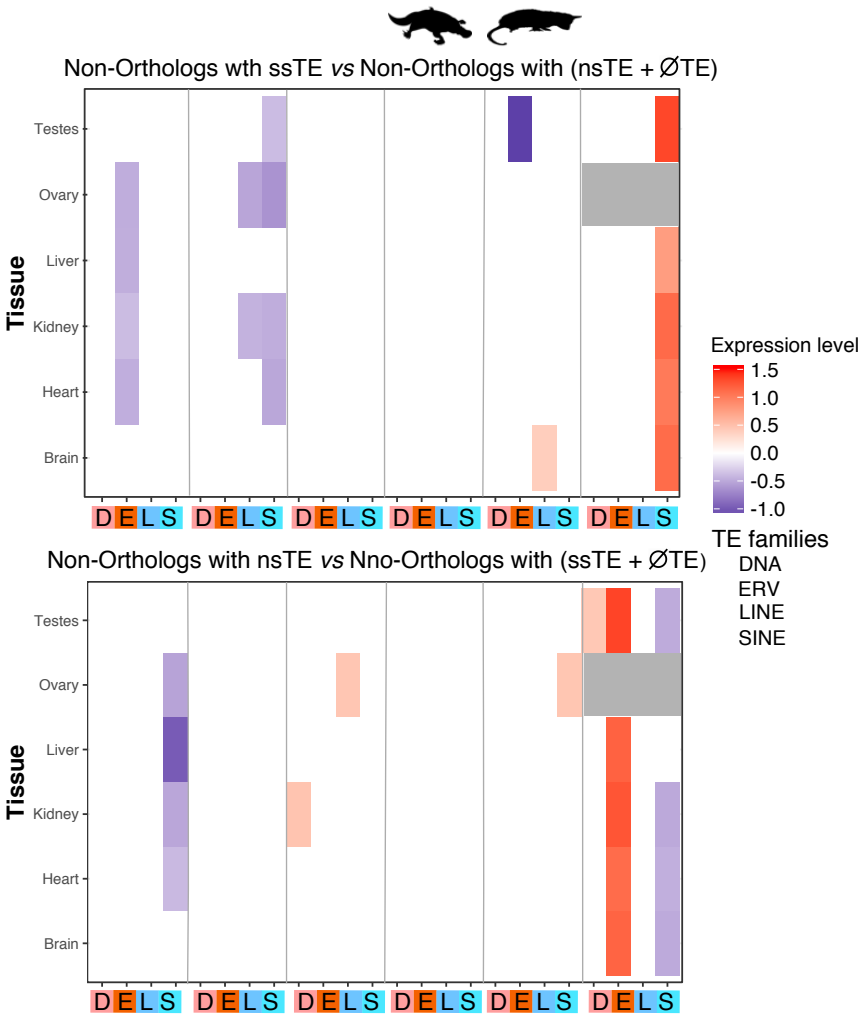


Figure S4: **Change in the level of non-ortholog gene expression as a function of TE insertion.** This figure shows the association between non-ortholog gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon (pogona), platypus, opossum and human) with recent species-specific TE insertions (ssTE) or non-recent species specific TE insertions (nsTE) (from left to right: DNA, ERV/LTR, LINE or SINE). A weighted bootstrap approach was used to compare the median gene expression levels of non-orthologous genes with a ssTE/nsTE insertion compared to non-orthologous gene without ssTE/nsTE. Gene expression levels are log2-transformed. Comparisons without statistically significant gene expression changes are shown in white. Statistically significant increased gene expression shown in red and statistically significant decreased gene expression in blue. Grey shading indicates no samples were available for this comparison.

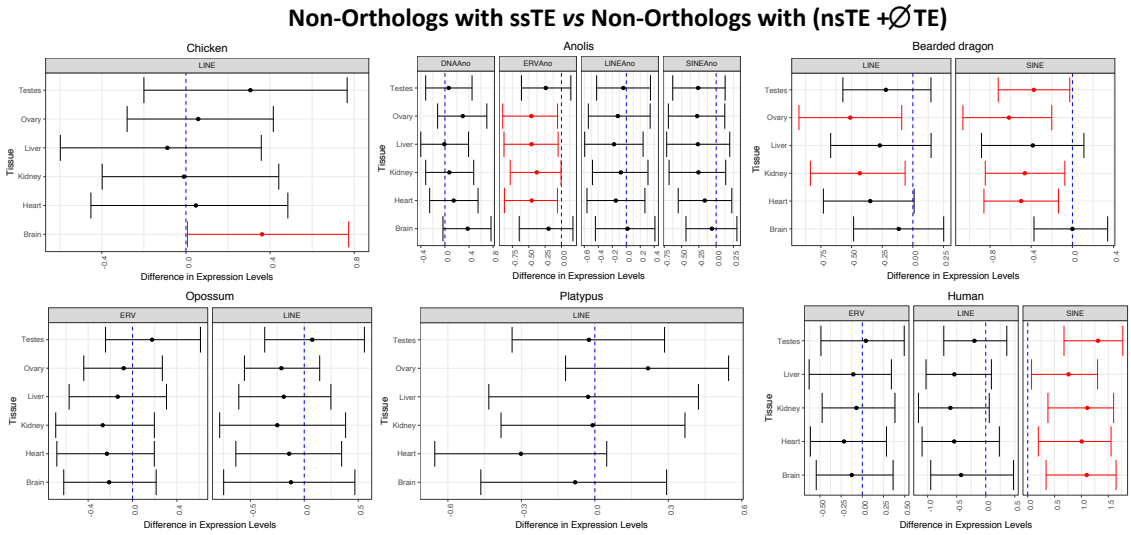


Figure S5: **Change in the level of non-ortholog gene expression as a function of species-specific TE insertion.**
This figure shows the association between non-ortholog gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon (pogona), platypus, opossum and human) with recent species-specific TE insertions (ssTE) (from left to right: DNA, ERV/LTR, LINE or SINE). Confidence Intervals for the difference in median log2(TPM) counts. Confidence Intervals were obtained using the weighted bootstrap and are 1-α/m intervals, where α=0.05 and m=nTissues*nElements as the total number of intervals presented. Red dots represent the median value from the bootstrap procedure, whilst the vertical line indicates zero. Intervals which do not contain zero are coloured red, and indicate a rejection of the null hypothesis, H_0: $\Delta\theta$=0, where θ represents the parameter of interest.

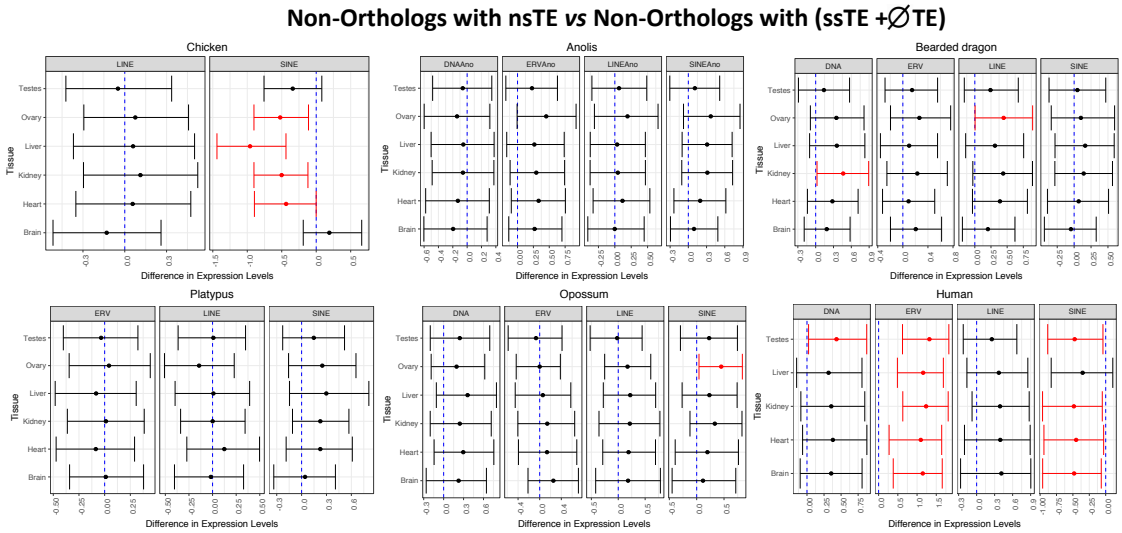


Figure S6: **Change in the level of non-ortholog gene expression as a function of non-species specific TE insertion.**
This figure shows the association between non-ortholog gene expression levels in six species (from left to right: chicken, anole lizard, bearded dragon (pogona), platypus, opossum and human) with non-recent species-specific TE insertions (nsTE) (from left to right: DNA, ERV/LTR, LINE or SINE). Confidence Intervals for the difference in median log2(TPM) counts. Confidence Intervals were obtained using the weighted bootstrap and are 1-α/m intervals, where α=0.05 and m=nTissues*nElements as the total number of intervals presented. Red dots represent the median value from the bootstrap procedure, whilst the vertical line indicates zero. Intervals which do not contain zero are coloured red, and indicate a rejection of the null hypothesis, H_0: $\Delta\theta$=0, where θ represents the parameter of interest.

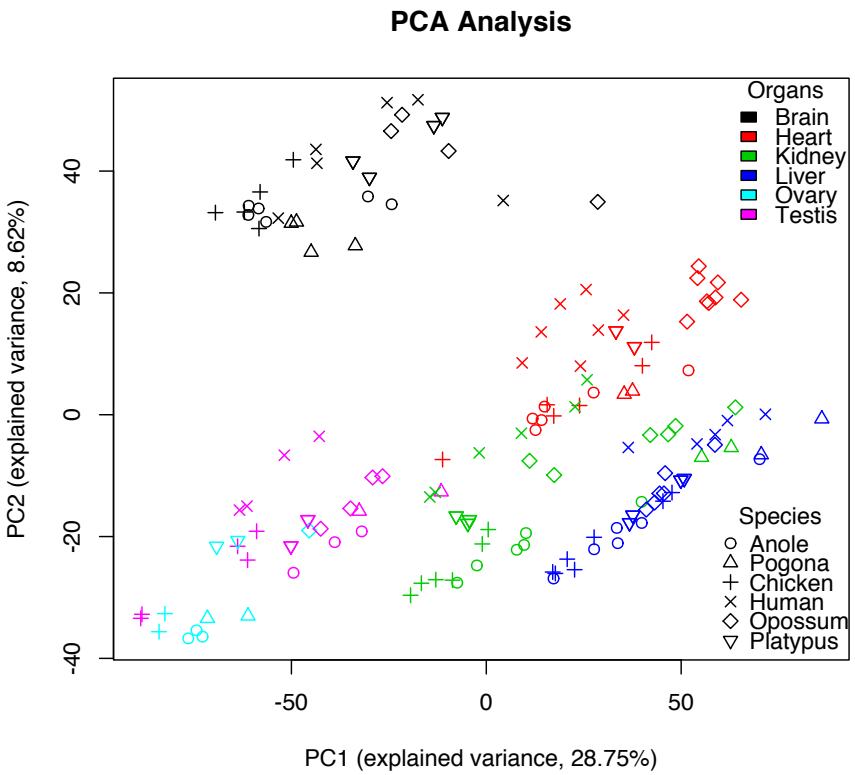


Figure S7: **Factorial map of the principal-component analysis of messenger RNA expression levels.**
This figure shows the PCA analysis of gene expression from six species (anole, bearded dragon (pogona), chicken, human, opossum and platypus) within six organs (brain, heart, kidney, liver, ovary and testis). Human samples did not include ovary. The proportion of the variance explained by the principal components is indicated in parentheses.

Supplementary Tables

Table 1: **Gene expression dataset.** Show the systematic name, common name, Gender, Tissue, layout, source, study and instrument. Gene expression that were acquired through private collaboration (not publicly available) are marked as 'Private' in the Submitter column. The Following abbreviations are used for submitters:
IH2500 = Illumina HiSeq 2500
IH2000 = Illumina HiSeq 2000
IGA IIX = Illumina Genome Analyzer IIX

Read accession(s)	Systematic Name	Common Name	Gender	Tissue	Layout	Source	Study	Instrument
SRR5412144	*Anolis carolinensis*	Anole	Female	Brain	Single	NCBI	Marin	IH2500
SRR5412145	*Anolis carolinensis*	Anole	Female	Brain	Single	NCBI	Marin	IH2500
SRR5412146	*Anolis carolinensis*	Anole	Female	Brain	Single	NCBI	Marin	IH2500
SRR5412147	*Anolis carolinensis*	Anole	Male	Brain	Single	NCBI	Marin	IH2500
SRR5412148	*Anolis carolinensis*	Anole	Male	Brain	Single	NCBI	Marin	IH2500
SRR5412149	*Anolis carolinensis*	Anole	Male	Brain	Single	NCBI	Marin	IH2500
SRR5412150	*Anolis carolinensis*	Anole	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412151	*Anolis carolinensis*	Anole	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412152	*Anolis carolinensis*	Anole	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412153	*Anolis carolinensis*	Anole	Male	Heart	Single	NCBI	Marin	IH2500
SRR5412154	*Anolis carolinensis*	Anole	Male	Heart	Single	NCBI	Marin	IH2500
SRR5412155	*Anolis carolinensis*	Anole	Male	Heart	Single	NCBI	Marin	IH2500
SRR5412156	*Anolis carolinensis*	Anole	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412157	*Anolis carolinensis*	Anole	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412158	*Anolis carolinensis*	Anole	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412159	*Anolis carolinensis*	Anole	Male	Kidney	Single	NCBI	Marin	IH2500
SRR5412160	*Anolis carolinensis*	Anole	Male	Kidney	Single	NCBI	Marin	IH2500
SRR5412161	*Anolis carolinensis*	Anole	Male	Kidney	Single	NCBI	Marin	IH2500
SRR5412162	*Anolis carolinensis*	Anole	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412163	*Anolis carolinensis*	Anole	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412164	*Anolis carolinensis*	Anole	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412165	*Anolis carolinensis*	Anole	Male	Liver	Single	NCBI	Marin	IH2500
SRR5412166	*Anolis carolinensis*	Anole	Male	Liver	Single	NCBI	Marin	IH2500
SRR5412167	*Anolis carolinensis*	Anole	Male	Liver	Single	NCBI	Marin	IH2500
SRR5412168	*Anolis carolinensis*	Anole	Female	Ovary	Single	NCBI	Marin	IH2500
SRR5412169	*Anolis carolinensis*	Anole	Female	Ovary	Single	NCBI	Marin	IH2500
SRR5412170	*Anolis carolinensis*	Anole	Female	Ovary	Single	NCBI	Marin	IH2500
SRR5412171	*Anolis carolinensis*	Anole	Male	Testes	Single	NCBI	Marin	IH2500
SRR5412172	*Anolis carolinensis*	Anole	Male	Testes	Single	NCBI	Marin	IH2500
SRR5412173	*Anolis carolinensis*	Anole	Male	Testes	Single	NCBI	Marin	IH2500
SRR5412242	*Gallus gallus*	Chicken	Female	Brain	Single	NCBI	Marin	IH2500
SRR5412243	*Gallus gallus*	Chicken	Male	Brain	Single	NCBI	Marin	IH2500
SRR5412244	*Gallus gallus*	Chicken	Male	Brain	Single	NCBI	Marin	IH2500
SRR306710	*Gallus gallus*	Chicken	Female	Brain	Single	NCBI	Brawand	IGA IIX
SRR306711	*Gallus gallus*	Chicken	Male	Brain	Single	NCBI	Brawand	IGA IIX
SRR5412245	*Gallus gallus*	Chicken	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412246	*Gallus gallus*	Chicken	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412247	*Gallus gallus*	Chicken	Male	Heart	Single	NCBI	Marin	IH2500

SRR5412248	*Gallus gallus*	Chicken	Male	Heart	Single	NCBI	Marin	IH2500	
SRR306714	*Gallus gallus*	Chicken	Female	Heart	Single	NCBI	Brawand	IGA IIX	
SRR306715	*Gallus gallus*	Chicken	Male	Heart	Single	NCBI	Brawand	IGA IIX	
SRR5412249	*Gallus gallus*	Chicken	Female	Kidney	Single	NCBI	Marin	IH2500	
SRR5412250	*Gallus gallus*	Chicken	Female	Kidney	Single	NCBI	Marin	IH2500	
SRR5412251	*Gallus gallus*	Chicken	Male	Kidney	Single	NCBI	Marin	IH2500	
SRR5412252	*Gallus gallus*	Chicken	Male	Kidney	Single	NCBI	Marin	IH2500	
SRR306716	*Gallus gallus*	Chicken	Female	Kidney	Single	NCBI	Brawand	IGA IIX	
SRR306717	*Gallus gallus*	Chicken	Male	Kidney	Single	NCBI	Brawand	IGA IIX	
SRR5412253	*Gallus gallus*	Chicken	Female	Liver	Single	NCBI	Marin	IH2500	
SRR5412254	*Gallus gallus*	Chicken	Female	Liver	Single	NCBI	Marin	IH2500	
SRR5412255	*Gallus gallus*	Chicken	Male	Liver	Single	NCBI	Marin	IH2500	
SRR5412256	*Gallus gallus*	Chicken	Male	Liver	Single	NCBI	Marin	IH2500	
SRR306718	*Gallus gallus*	Chicken	Female	Liver	Single	NCBI	Brawand	IGA IIX	
SRR306719	*Gallus gallus*	Chicken	Male	Liver	Single	NCBI	Brawand	IGA IIX	
SRR306720	*Gallus gallus*	Chicken	Male	Liver	Single	NCBI	Brawand	IGA IIX	
SRR5412257	*Gallus gallus*	Chicken	Female	Ovary	Single	NCBI	Marin	IH2500	
SRR5412258	*Gallus gallus*	Chicken	Female	Ovary	Single	NCBI	Marin	IH2500	
SRR5412259	*Gallus gallus*	Chicken	Male	Testis	Single	NCBI	Marin	IH2500	
SRR5412260	*Gallus gallus*	Chicken	Male	Testis	Single	NCBI	Marin	IH2500	
SRR306721	*Gallus gallus*	Chicken	Male	Testis	Single	NCBI	Brawand	IGA IIX	
SRR306722	*Gallus gallus*	Chicken	Male	Testis	Single	NCBI	Brawand	IGA IIX	
SRR306723	*Gallus gallus*	Chicken	Male	Testis	Single	NCBI	Brawand	IGA IIX	
ERR753525	*Pogona Vitticeps*	Bearded dragon	Male	Brain	Paired	NCBI	Georges	IH2000	
ERR413064	*Pogona Vitticeps*	Bearded dragon	Male	Brain	Paired	NCBI	Georges	IH2000	
ERR753526	*Pogona Vitticeps*	Bearded dragon	Female	Brain	Paired	NCBI	Georges	IH2000	
ERR413071	*Pogona Vitticeps*	Bearded dragon	Female	Brain	Paired	NCBI	Georges	IH2000	
ERR413072	*Pogona Vitticeps*	Bearded dragon	Female	Heart	Paired	NCBI	Georges	IH2000	
ERR413065	*Pogona Vitticeps*	Bearded dragon	Male	Heart	Paired	NCBI	Georges	IH2000	
ERR413073	*Pogona Vitticeps*	Bearded dragon	Female	Kidney	Paired	NCBI	Georges	IH2000	
ERR413066	*Pogona Vitticeps*	Bearded dragon	Male	Kidney	Paired	NCBI	Georges	IH2000	
ERR413074	*Pogona Vitticeps*	Bearded dragon	Female	Liver	Paired	NCBI	Georges	IH2000	
ERR413067	*Pogona Vitticeps*	Bearded dragon	Male	Liver	Paired	NCBI	Georges	IH2000	
ERR413070	*Pogona Vitticeps*	Bearded dragon	Male	Testis	Paired	NCBI	Georges	IH2000	
ERR753529	*Pogona Vitticeps*	Bearded dragon	Male	Testis	Paired	NCBI	Georges	IH2000	
ERR753530	*Pogona Vitticeps*	Bearded dragon	Female	Ovary	Paired	NCBI	Georges	IH2000	
ERR413082	*Pogona Vitticeps*	Bearded dragon	Female	Ovary	Paired	NCBI	Georges	IH2000	
SRR5412222	*Ornithorhynchus anatinus*	Platypus	Female	Brain	Single	NCBI	Marin	IH2500	
SRR5412223	*Ornithorhynchus anatinus*	Platypus	Female	Brain	Single	NCBI	Marin	IH2500	
SRR5412224	*Ornithorhynchus anatinus*	Platypus	Male	Brain	Single	NCBI	Marin	IH2500	
SRR5412225	*Ornithorhynchus anatinus*	Platypus	Male	Brain	Single	NCBI	Marin	IH2500	
SRR306724	*Ornithorhynchus anatinus*	Platypus	Female	Brain	Single	NCBI	Brawand	IGA IIX	
SRR306725	*Ornithorhynchus anatinus*	Platypus	Female	Brain	Single	NCBI	Brawand	IGA IIX	
SRR306726	*Ornithorhynchus anatinus*	Platypus	Male	Brain	Single	NCBI	Brawand	IGA IIX	
SRR306727	*Ornithorhynchus anatinus*	Platypus	Male	Brain	Single	NCBI	Brawand	IGA IIX	
SRR5412226	*Ornithorhynchus anatinus*	Platypus	Female	Heart	Single	NCBI	Marin	IH2500	

SRR5412227	*Ornithorhynchus anatinus*	Platypus	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412228	*Ornithorhynchus anatinus*	Platypus	Male	Heart	Single	NCBI	Marin	IH2500
SRR5412229	*Ornithorhynchus anatinus*	Platypus	Male	Heart	Single	NCBI	Marin	IH2500
SRR306730	*Ornithorhynchus anatinus*	Platypus	Female	Heart	Single	NCBI	Brawand	IGA IIX
SRR306731	*Ornithorhynchus anatinus*	Platypus	Male	Heart	Single	NCBI	Brawand	IGA IIX
SRR5412230	*Ornithorhynchus anatinus*	Platypus	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412231	*Ornithorhynchus anatinus*	Platypus	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412232	*Ornithorhynchus anatinus*	Platypus	Male	Kidney	Single	NCBI	Marin	IH2500
SRR5412233	*Ornithorhynchus anatinus*	Platypus	Male	Kidney	Single	NCBI	Marin	IH2500
SRR306732	*Ornithorhynchus anatinus*	Platypus	Female	Kidney	Single	NCBI	Brawand	IGA IIX
SRR306733	*Ornithorhynchus anatinus*	Platypus	Male	Kidney	Single	NCBI	Brawand	IGA IIX
SRR306734	*Ornithorhynchus anatinus*	Platypus	Male	Kidney	Single	NCBI	Brawand	IGA IIX
SRR5412234	*Ornithorhynchus anatinus*	Platypus	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412235	*Ornithorhynchus anatinus*	Platypus	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412236	*Ornithorhynchus anatinus*	Platypus	Male	Liver	Single	NCBI	Marin	IH2500
SRR5412237	*Ornithorhynchus anatinus*	Platypus	Male	Liver	Single	NCBI	Marin	IH2500
SRR306735	*Ornithorhynchus anatinus*	Platypus	Female	Liver	Single	NCBI	Brawand	IGA IIX
SRR306736	*Ornithorhynchus anatinus*	Platypus	Female	Liver	Single	NCBI	Brawand	IGA IIX
SRR306737	*Ornithorhynchus anatinus*	Platypus	Male	Liver	Single	NCBI	Brawand	IGA IIX
SRR306738	*Ornithorhynchus anatinus*	Platypus	Male	Liver	Single	NCBI	Brawand	IGA IIX
SRR5412238	*Ornithorhynchus anatinus*	Platypus	Female	Ovary	Single	NCBI	Marin	IH2500
SRR5412239	*Ornithorhynchus anatinus*	Platypus	Female	Ovary	Single	NCBI	Marin	IH2500
SRR5412240	*Ornithorhynchus anatinus*	Platypus	Male	Testis	Single	NCBI	Marin	IH2500
SRR5412241	*Ornithorhynchus anatinus*	Platypus	Male	Testis	Single	NCBI	Marin	IH2500
SRR306739	*Ornithorhynchus anatinus*	Platypus	Male	Testis	Single	NCBI	Brawand	IGA IIX
SRR306741	*Ornithorhynchus anatinus*	Platypus	Male	Testis	Single	NCBI	Brawand	IGA IIX
SRR5412205	*Monodelphis domestica*	Opossum	Female	Brain	Single	NCBI	Marin	IH2500
SRR5412206	*Monodelphis domestica*	Opossum	Male	Brain	Single	NCBI	Marin	IH2500
SRR306742	*Monodelphis domestica*	Opossum	Female	Brain	Single	NCBI	Brawand	IGA IIX
SRR306743	*Monodelphis domestica*	Opossum	Female	Brain	Single	NCBI	Brawand	IGA IIX
SRR306744	*Monodelphis domestica*	Opossum	Male	Brain	Single	NCBI	Brawand	IGA IIX
SRR5412207	*Monodelphis domestica*	Opossum	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412208	*Monodelphis domestica*	Opossum	Female	Heart	Single	NCBI	Marin	IH2500
SRR5412209	*Monodelphis domestica*	Opossum	Male	Heart	Single	NCBI	Marin	IH2500
SRR5412210	*Monodelphis domestica*	Opossum	Male	Heart	Single	NCBI	Marin	IH2500
SRR306747	*Monodelphis domestica*	Opossum	Female	Heart	Single	NCBI	Brawand	IGA IIX
SRR306748	*Monodelphis domestica*	Opossum	Female	Heart	Single	NCBI	Brawand	IGA IIX
SRR306749	*Monodelphis domestica*	Opossum	Male	Heart	Single	NCBI	Brawand	IGA IIX
SRR306750	*Monodelphis domestica*	Opossum	Male	Heart	Single	NCBI	Brawand	IGA IIX
SRR5412211	*Monodelphis domestica*	Opossum	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412212	*Monodelphis domestica*	Opossum	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412213	*Monodelphis domestica*	Opossum	Male	Kidney	Single	NCBI	Marin	IH2500
SRR5412214	*Monodelphis domestica*	Opossum	Male	Kidney	Single	NCBI	Marin	IH2500
SRR306751	*Monodelphis domestica*	Opossum	Female	Kidney	Single	NCBI	Brawand	IGA IIX
SRR306752	*Monodelphis domestica*	Opossum	Male	Kidney	Single	NCBI	Brawand	IGA IIX
SRR5412215	*Monodelphis domestica*	Opossum	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412216	*Monodelphis domestica*	Opossum	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412217	*Monodelphis domestica*	Opossum	Male	Liver	Single	NCBI	Marin	IH2500
SRR5412218	*Monodelphis domestica*	Opossum	Male	Liver	Single	NCBI	Marin	IH2500
SRR306753	*Monodelphis domestica*	Opossum	Female	Liver	Single	NCBI	Brawand	IGA IIX

SRR306754	*Monodelphis domestica*	Opossum	Male	Liver	Single	NCBI	Brawand	IGA IIX
SRR5412219	*Monodelphis domestica*	Opossum	Female	Ovary	Single	NCBI	Marin	IH2500
SRR5412220	*Monodelphis domestica*	Opossum	Male	Testis	Single	NCBI	Marin	IH2500
SRR5412221	*Monodelphis domestica*	Opossum	Male	Testis	Single	NCBI	Marin	IH2500
SRR306755	*Monodelphis domestica*	Opossum	Male	Testis	Single	NCBI	Brawand	IGA IIX
SRR306756	*Monodelphis domestica*	Opossum	Male	Testis	Single	NCBI	Brawand	IGA IIX
SRR5412174	*Homo sapiens*	Human	Female	Brain	Single	NCBI	Marin	IH2500
SRR5412175	*Homo sapiens*	Human	Male	Brain	Single	NCBI	Marin	IH2500
SRR306838	*Homo sapiens*	Human	Female	Brain	Single	NCBI	Brawand	IGA IIX
SRR306839	*Homo sapiens*	Human	Male	Brain	Single	NCBI	Brawand	IGA IIX
SRR306841	*Homo sapiens*	Human	Male	Brain	Single	NCBI	Brawand	IGA IIX
SRR306843	*Homo sapiens*	Human	Male	Brain	Single	NCBI	Brawand	IGA IIX
SRR5412176	*Homo sapiens*	Human	Female	Heart	Paired	NCBI	Marin	IH2500
SRR5412177	*Homo sapiens*	Human	Male	Heart	Single	NCBI	Marin	IH2500
SRR5412178	*Homo sapiens*	Human	Male	Heart	Paired	NCBI	Marin	IH2500
SRR306847	*Homo sapiens*	Human	Female	Heart	Single	NCBI	Brawand	IGA IIX
SRR306848	*Homo sapiens*	Human	Male	Heart	Single	NCBI	Brawand	IGA IIX
SRR306849	*Homo sapiens*	Human	Male	Heart	Single	NCBI	Brawand	IGA IIX
SRR306850	*Homo sapiens*	Human	Male	Heart	Single	NCBI	Brawand	IGA IIX
SRR5412179	*Homo sapiens*	Human	Female	Kidney	Single	NCBI	Marin	IH2500
SRR5412180	*Homo sapiens*	Human	Male	Kidney	Single	NCBI	Marin	IH2500
SRR5412181	*Homo sapiens*	Human	Male	Kidney	Single	NCBI	Marin	IH2500
SRR306851	*Homo sapiens*	Human	Female	Kidney	Single	NCBI	Brawand	IGA IIX
SRR306852	*Homo sapiens*	Human	Male	Kidney	Single	NCBI	Brawand	IGA IIX
SRR306853	*Homo sapiens*	Human	Male	Kidney	Single	NCBI	Brawand	IGA IIX
SRR5412182	*Homo sapiens*	Human	Female	Liver	Single	NCBI	Marin	IH2500
SRR5412183	*Homo sapiens*	Human	Male	Liver	Single	NCBI	Marin	IH2500
SRR306854	*Homo sapiens*	Human	Male	Liver	Single	NCBI	Brawand	IGA IIX
SRR306855	*Homo sapiens*	Human	Male	Liver	Single	NCBI	Brawand	IGA IIX
SRR306856	*Homo sapiens*	Human	Male	Liver	Single	NCBI	Brawand	IGA IIX
SRR5412184	*Homo sapiens*	Human	Male	Testis	Single	NCBI	Marin	IH2500
SRR5412185	*Homo sapiens*	Human	Male	Testis	Single	NCBI	Marin	IH2500
SRR306857	*Homo sapiens*	Human	Male	Testes	Single	NCBI	Brawand	IGA IIX
SRR306858	*Homo sapiens*	Human	Male	Testes	Single	NCBI	Brawand	IGA IIX

Table 2: **Assembly dataset**. Shows the systematic name, common name, genome version, source and submitter for all the genomes tested with our *ab initio* method.
The Following abbreviations are used for submitters:
Genome Sequencing Platform, The Genome Assembly Team = GAT;
Genome Reference Consortium = GRC;
International Chicken Genome Consortium = ICGS;
Washington University = WashU.

No	Systematic Name	Common Name	RefSeq Assembly Accession	Source	Submitter
1	*Homo sapiens*	Human	GCF_000001405.25	NCBI	GRC
2	*Pogona Vitticeps*	Bearded Dragon	GCF_900067755.1	NCBI	BRAEMBL
3	*Anolis Carollinensis*	Anole lizard	GCF_000090745.1	NCBI	Broad
4	*Gallus gallus*	Chicken	GCF_000002315.3	NCBI	ICGS
5	*Monodelphis domestica*	Opossum	GCF_000002295.2	NCBI	GAT
6	*Ornithorhynchus anatinus*	Platypus	GCF_000002275.2	NCBI	WashU

Table 3: **Comparsion of orthologs with ssTE *vs* orthologs with nsTE and ∅ TE**. Shows the number of sample genes used in the bootstrap approach. Test sample is ortholog genes containing recent species-specific TE (ssTE), reference sample is ortholog genes with no ssTE.

	Chicken		**Anole**		**Bearded dragon**		**Platypus**		**Opossum**		**Human**	
	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference
LINE	1,580	5,015	4,135	2,640	3,613	2,982	1,854	4,741	3,274	3,321	2,048	4,547
SINE	0	NA	1,566	5,029	5,660	935	513	NA	317	NA	3,388	3,207
ERV	143	NA	2,340	4,255	104	NA	16	NA	3,064	3,531	994	5,601
DNA	5	NA	3,436	3,159	496	NA	6	NA	236	NA	45	NA

Table 4: **Comparsion of orthologs with nsTE *vs* orthologs with ssTE and ∅ TE**. Shows the number of sample genes used in the bootstrap approach. Test sample is ortholog genes containing non-recent species specific TE (nsTE), reference sample is ortholog genes with no nsTE.

	Chicken		**Anole**		**Bearded dragon**		**Platypus**		**Opossum**		**Human**	
	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference
LINE	4,320	2,275	2,221	4,374	2,805	3,790	4,516	2,079	3,013	3,582	4,340	2,255
SINE	1,174	5,421	4,369	2,226	4,667	1,928	5,830	765	6,076	NA	3,070	3,525
ERV	5,797	NA	3,652	2,943	6,106	NA	5,374	NA	2,871	3,724	5,470	1,125
DNA	5,894	NA	3,066	3,529	5,931	NA	5,525	NA	5,819	NA	6,455	NA

Table 5: **Comparsion of non-orthologs with ssTE *vs* non-orthologs with nsTE and ∅ TE**. Shows the number of sample genes used in bootstrap approach. Test sample is non-ortholog genes containing recent species-specific TE (ssTE), reference sample is non-ortholog genes with no ssTE.

	Chicken		**Anole**		**Bearded dragon**		**Platypus**		**Opossum**		**Human**	
	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference
LINE	1,488	9,025	8,337	10,988	5,671	9,728	2,677	16,844	5,025	12,279	6,065	45,076
SINE	0	NA	2,670	16,655	1,203	14,196	553	NA	413	NA	8,603	42,538
ERV	211	NA	4,528	14,797	142	NA	24	NA	4,439	12,865	3,401	47,740
DNA	11	NA	5,593	13,372	560	NA	5	NA	344	NA	220	NA

Table 6: **Comparsion of non-orthologs with nsTE *vs* non-orthologs with ssTE and ∅ TE**. Shows the number of sample genes used in the bootstrap approach. Test sample is non-ortholog genes containing non-recent species specific TE (nsTE), reference sample is non-ortholog genes with no nsTE.

	Chicken		**Anole**		**Bearded dragon**		**Platypus**		**Opossum**		**Human**	
	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference	Test	Reference
LINE	6,402	4,111	8,186	11,139	8,428	6,971	14,113	5,408	9,423	7,881	32,875	18,266
SINE	1,191	9,322	11,472	7,853	9,147	6,252	16,175	3,346	13,241	4,063	32,869	18,272
ERV	7,453	3,060	8,690	10,635	12,671	2,728	9,778	9,743	7,360	9,944	34,177	18,202
DNA	7,397	3,116	11,320	8,005	13,190	2,209	9,878	NA	10,232	7,072	32,939	16,964

Table 7: **Difference in the gene expression of orthologs/non-orthologs with a TE insertion.** Shows the species, TE element, Tissue, gene expression comparison sets , lowest gene expression level, median gene expression level, highest gene expression level, bonferroni-adjusted 95% CI lowest gene expression, bonferroni-adjusted 95% CI highest expression level and the significance indicator. TPM counts were log2 transformed.

Species	Element	Tissue	Data	lwr	med	upr	lwr95	upr95	Sig
Platypus	LINE	Heart	ssTE-ortholog	-0.859	-0.443	-0.045	-0.769	-0.131	T
Bearded dragon	LINE	Heart	ssTE-ortholog	-0.704	-0.387	-0.064	-0.626	-0.147	T
Bearded dragon	LINE	Kidney	ssTE-ortholog	-0.815	-0.485	-0.160	-0.732	-0.242	T
Bearded dragon	LINE	Liver	ssTE-ortholog	-0.864	-0.525	-0.188	-0.771	-0.276	T
Bearded dragon	LINE	Ovary	ssTE-ortholog	-0.975	-0.629	-0.278	-0.883	-0.373	T
Bearded dragon	LINE	Testes	ssTE-ortholog	-0.677	-0.369	-0.041	-0.599	-0.129	T
Bearded dragon	SINE	Heart	ssTE-ortholog	-0.946	-0.584	-0.218	-0.851	-0.322	T
Bearded dragon	SINE	Kidney	ssTE-ortholog	-1.035	-0.672	-0.325	-0.933	-0.406	T
Bearded dragon	SINE	Liver	ssTE-ortholog	-0.965	-0.566	-0.194	-0.851	-0.283	T
Bearded dragon	SINE	Ovary	ssTE-ortholog	-1.022	-0.592	-0.212	-0.901	-0.300	T
Bearded dragon	SINE	Testes	ssTE-ortholog	-0.939	-0.601	-0.259	-0.854	-0.351	T
Chicken	LINE	Brain	ssTE-ortholog	0.013	0.323	0.667	0.089	0.579	T
Anole	LINE	Heart	ssTE-ortholog	-0.792	-0.444	-0.112	-0.701	-0.197	T
Anole	LINE	Kidney	ssTE-ortholog	-0.830	-0.459	-0.128	-0.724	-0.212	T
Anole	LINE	Liver	ssTE-ortholog	-0.832	-0.481	-0.120	-0.743	-0.211	T
Anole	LINE	Ovary	ssTE-ortholog	-0.730	-0.402	-0.087	-0.647	-0.168	T
Anole	SINE	Heart	ssTE-ortholog	-0.847	-0.509	-0.191	-0.758	-0.266	T
Anole	SINE	Kidney	ssTE-ortholog	-0.824	-0.506	-0.185	-0.745	-0.267	T
Anole	SINE	Ovary	ssTE-ortholog	-0.838	-0.493	-0.170	-0.751	-0.252	T
Anole	SINE	Testes	ssTE-ortholog	-0.762	-0.459	-0.154	-0.684	-0.232	T
Anole	ERV	Heart	ssTE-ortholog	-0.760	-0.426	-0.104	-0.669	-0.187	T
Anole	ERV	Kidney	ssTE-ortholog	-0.968	-0.653	-0.348	-0.886	-0.425	T
Anole	ERV	Liver	ssTE-ortholog	-1.011	-0.680	-0.347	-0.924	-0.433	T
Anole	ERV	Ovary	ssTE-ortholog	-0.709	-0.391	-0.066	-0.628	-0.151	T
Human	LINE	Kidney	ssTE-ortholog	-1.179	-0.746	-0.041	-1.079	-0.126	T
Human	LINE	Liver	ssTE-ortholog	-1.283	-0.738	-0.116	-1.179	-0.216	T
Human	LINE	Testes	ssTE-ortholog	-0.917	-0.562	-0.221	-0.840	-0.296	T
Human	SINE	Brain	ssTE-ortholog	0.119	0.429	0.750	0.196	0.671	T
Human	SINE	Testes	ssTE-ortholog	0.274	0.619	0.933	0.357	0.858	T
Bearded dragon	LINE	Kidney	ssTE-nonOrtholog	-0.836	-0.433	-0.063	-0.695	-0.172	T
Bearded dragon	LINE	Ovary	ssTE-nonOrtholog	-0.929	-0.510	-0.091	-0.797	-0.226	T
Bearded dragon	SINE	Heart	ssTE-nonOrtholog	-0.859	-0.498	-0.136	-0.748	-0.250	T
Bearded dragon	SINE	Kidney	ssTE-nonOrtholog	-0.846	-0.463	-0.074	-0.725	-0.199	T
Bearded dragon	SINE	Ovary	ssTE-nonOrtholog	-1.063	-0.616	-0.200	-0.933	-0.334	T
Bearded dragon	SINE	Testes	ssTE-nonOrtholog	-0.720	-0.375	-0.026	-0.619	-0.136	T
Chicken	LINE	Brain	ssTE-nonOrtholog	0.008	0.388	0.832	0.125	0.694	T
Anole	ERV	Heart	ssTE-nonOrtholog	-0.855	-0.456	-0.076	-0.730	-0.191	T
Anole	ERV	Kidney	ssTE-nonOrtholog	-0.764	-0.378	-0.020	-0.640	-0.127	T
Anole	ERV	Liver	ssTE-nonOrtholog	-0.865	-0.458	-0.062	-0.737	-0.191	T
Anole	ERV	Ovary	ssTE-nonOrtholog	-0.876	-0.461	-0.073	-0.743	-0.194	T
Human	SINE	Brain	ssTE-nonOrtholog	0.350	1.099	1.629	0.470	1.478	T
Human	SINE	Heart	ssTE-nonOrtholog	0.206	1.005	1.538	0.315	1.375	T
Human	SINE	Kidney	ssTE-nonOrtholog	0.386	1.109	1.590	0.511	1.451	T
Human	SINE	Liver	ssTE-nonOrtholog	0.085	0.759	1.292	0.218	1.147	T
Human	SINE	Testes	ssTE-nonOrtholog	0.685	1.310	1.765	0.815	1.621	T
Platypus	SINE	Liver	nsTE-ortholog	0.089	0.622	1.168	0.251	1.001	T
Platypus	SINE	Ovary	nsTE-ortholog	0.027	0.473	0.924	0.162	0.780	T
Platypus	SINE	Testes	nsTE-ortholog	0.114	0.567	1.031	0.250	0.890	T
Bearded dragon	LINE	Heart	nsTE-ortholog	0.009	0.391	0.786	0.134	0.649	T

Bearded dragon	LINE	Kidney	nsTE-ortholog	0.050	0.428	0.795	0.172	0.687	T
Bearded dragon	LINE	Liver	nsTE-ortholog	0.125	0.523	0.916	0.250	0.790	T
Bearded dragon	LINE	Ovary	nsTE-ortholog	0.186	0.577	0.972	0.300	0.840	T
Bearded dragon	LINE	Testes	nsTE-ortholog	0.007	0.368	0.733	0.125	0.610	T
Chicken	SINE	Heart	nsTE-ortholog	-1.039	-0.624	-0.250	-0.905	-0.365	T
Chicken	SINE	Kidney	nsTE-ortholog	-1.134	-0.743	-0.356	-1.012	-0.472	T
Chicken	SINE	Liver	nsTE-ortholog	-1.424	-0.935	-0.451	-1.272	-0.583	T
Chicken	SINE	Ovary	nsTE-ortholog	-1.051	-0.685	-0.340	-0.935	-0.447	T
Chicken	SINE	Testes	nsTE-ortholog	-1.025	-0.715	-0.412	-0.925	-0.508	T
Anole	LINE	Heart	nsTE-ortholog	0.149	0.514	0.895	0.260	0.777	T
Anole	LINE	Kidney	nsTE-ortholog	0.212	0.567	0.974	0.320	0.844	T
Anole	LINE	Liver	nsTE-ortholog	0.175	0.578	0.978	0.303	0.852	T
Anole	LINE	Ovary	nsTE-ortholog	0.170	0.534	0.919	0.282	0.795	T
Anole	SINE	Heart	nsTE-ortholog	0.035	0.393	0.758	0.148	0.643	T
Anole	SINE	Ovary	nsTE-ortholog	0.109	0.474	0.836	0.222	0.726	T
Anole	SINE	Testes	nsTE-ortholog	0.084	0.431	0.773	0.196	0.665	T
Human	LINE	Liver	nsTE-ortholog	0.033	0.584	1.159	0.155	1.002	T
Human	LINE	Testes	nsTE-ortholog	0.102	0.474	0.844	0.209	0.732	T
Human	SINE	Testes	nsTE-ortholog	-0.776	-0.411	-0.063	-0.664	-0.169	T
Bearded dragon	LINE	Ovary	nsTE-nonOrtholog	0.012	0.446	0.879	0.155	0.743	T
Bearded dragon	DNA	Kidney	nsTE-nonOrtholog	0.023	0.464	0.892	0.186	0.751	T
Chicken	SINE	Kidney	nsTE-nonOrtholog	-0.915	-0.502	-0.102	-0.775	-0.234	T
Chicken	SINE	Liver	nsTE-nonOrtholog	-1.459	-0.960	-0.421	-1.292	-0.585	T
Chicken	SINE	Ovary	nsTE-nonOrtholog	-0.913	-0.522	-0.086	-0.781	-0.255	T
Opossum	SINE	Ovary	nsTE-nonOrtholog	0.053	0.449	0.830	0.178	0.709	T
Human	SINE	Brain	nsTE-nonOrtholog	-0.955	-0.481	-0.071	-0.803	-0.198	T
Human	SINE	Heart	nsTE-nonOrtholog	-0.937	-0.451	-0.033	-0.772	-0.156	T
Human	SINE	Kidney	nsTE-nonOrtholog	-0.955	-0.485	-0.054	-0.804	-0.191	T
Human	SINE	Testes	nsTE-nonOrtholog	-0.879	-0.473	-0.044	-0.740	-0.202	T
Human	ERV	Brain	nsTE-nonOrtholog	0.382	1.149	1.651	0.521	1.482	T
Human	ERV	Heart	nsTE-nonOrtholog	0.278	1.097	1.638	0.437	1.469	T
Human	ERV	Kidney	nsTE-nonOrtholog	0.631	1.235	1.806	0.787	1.611	T
Human	ERV	Liver	nsTE-nonOrtholog	0.493	1.159	1.683	0.636	1.521	T
Human	ERV	Testes	nsTE-nonOrtholog	0.624	1.321	1.815	0.765	1.656	T
Human	DNA	Testes	nsTE-nonOrtholog	0.020	0.433	0.891	0.162	0.717	T