Jumping the fine LINE between species:

Horizontal transfer and evolution of repetitive elements in eukaryotic species

By

ATMA MARIA IVANCEVIC

THE UNIVERSITY of ADELAIDE

Department of Genetics and Evolution

School of Biological Sciences

A thesis presented for the degree of DOCTOR OF PHILOSOPHY

DECEMBER 2016

Abstract

Transposable elements (TEs) are mobile DNA sequences, colloquially known as 'jumping genes' because of their ability to replicate to new genomic locations. Active TEs have the potential to transform genome structure by inserting into regulatory regions or accumulating within the genome. Mammals are particularly susceptible to TE expansion; TEs account for significant proportions of all eukaryotic genomes we see today.

Horizontal transfer (HT) is the transmission of genetic material between non-mating species. HT is frequently observed in prokaryotes but rarely occurs in multicellular eukaryotes. As TEs are autonomous elements, they have the capability to move into another genome and immediately commence replicating, making them the perfect candidate for eukaryotic HT. Growing evidence indicates that this phenomenon is more widespread than current literature suggests, although questions still remain concerning the frequency of HT and whether all TEs are capable of moving between species.

In this thesis, I describe large-scale phylogenomic analyses of eukaryotic species in order to identify and characterise TEs, particularly BovB and L1 (predominantly found in mammals). Past studies on this topic were limited by the scarce availability of genome sequences, which were mainly model organisms. I addressed this limitation by comprehensively screening more than 500 species, demonstrating the remarkable and overlooked diversity of L1s across the eukaryotic tree of life. The rapid explosion of L1s in mammals provides a striking contrast to the diverged L1 lineages found in other metazoans and plants. Even within individual genomes there are marked differences between ancient, degraded L1s and young, intact L1s that are potentially still active.

L1s are only believed to vertically inherited; with my plethora of data, I challenged this perception by mining for L1 HT candidates. For comparison, I used BovB retrotransposons as an exemplar of obvious and rampant eukaryotic HT. I extended the current BovB paradigm to include more species, find new vectors of transfer, and refine the estimated times of insertion. Similarities between the distributions of L1 and BovB led me to postulate that the presence of L1s in therian mammals is due to an ancient HT event. Similar L1 HT events can be observed in plants. Given the extent of L1 colonisation in today's mammals, the idea that L1s were initially introduced as foreign DNA has wide-reaching implications for our perception of genome evolution.

Repetitive elements are often discarded from analyses because they are deemed 'junk' DNA. However, a genome's junk is a bioinformatician's treasure. Chapter 4 details a novel method for resolving species differences by using the repetitive intervals in a genome to identify binary variance (presence versus absence). We were able to infer the evolutionary relationships of 21 modern and ancient elephants and compare the results to an established phylogeny from single nucleotide polymorphisms (SNP). Repeats can thus be used as informative genetic markers, particularly useful for datasets with no known SNP variants.

Altogether, this thesis presents *in silico* approaches for handling large and highly repetitive datasets. By characterising millions of repetitive elements from 503 eukaryotic species, we provide evidence of their impact and importance in eukaryotic evolution.

Dedication and Acknowledgements

The completion of this thesis would not be possible without the support and guidance of some truly exceptional people. I would like to express my gratitude to all of them, particularly the following.

First of all, I would like to thank David Adelson, Dan Kortschak and Terry Bertozzi. To call you supervisors would be a gross understatement of the counsel and inspiration you have given me these last few years. Dave, your perpetual optimism makes it a joy to come to the lab and I cannot even begin to thank you for all the opportunities you have provided. Dan, you have always seemed like more of a friend than a supervisor and I am eternally grateful for your advice on all matters, both personal and professional. Terry, your unwavering enthusiasm and boundless knowledge has been a much needed source of comfort during times of doubt. The fact that I wish to pursue science, even academia, attests to the impact all of you have had on my life.

To the wonderful people in the Adelson lab, you have made me look forward to uni every day. Lu, you are and always will be my sister. It is impossible to feel down with you around and I look forward to many more bubble tea catchups for years to come. Reuben, I'd like to thank you for graciously accepting your role as lab punching bag - the rivalry between you and Lu has been an endless source of amusement. Brittany and James, it has been an honour to work with you and watch you grow as scientists. Zhipeng, your wisdom and advice always came at the most crucial times. Hien, your ready smile and computational experience has been a lifesaver these past few months. This list could be twice as long, and it would still be a poor representation of my PhD experience.

I would like to thank my family, for introducing me to Netflix and bringing me cake when things got hard. Kali and Nick, you are the sweetest siblings anyone could ever hope for. Your constant support, in whatever I do, means the world to me.

Finally, to Joel. Thank you for being there through the bad days and celebrating with me on the good days. Your frequent reminder to "work smart, not hard" is probably the reason this thesis is complete. Thank you for reminding me to enjoy life every once in a while, despite my protestations.

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: .....20/2/17.....

Table of Contents

Chapter 1

Introduction

"Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less."

— Marie Curie

Studies in evolutionary biology have predominantly focused on coding regions such as genes, which have known function and structure. Only recently have scientists started looking into the 'dark matter' of the genome, formerly dismissed as 'junk' DNA because of its highly repetitive nature. Rapid advances in genome sequencing techniques mean that we are now able to distinguish and categorise repeats based on their sequence composition. Of particular interest are transposable elements: mobile, parasitic sequences which are able to perpetually replicate themselves within genomes. Given a vector of transfer (e.g. tick or virus), these elements are able to jump further, between organisms or species in a process known as horizontal transfer. Transferred elements can interrupt existing genomic structures and thus have a huge impact on the new host. Analysing the likelihood of transfer and rate of expansion can help us predict how likely these elements are to affect future generations.

Statement of Authorship

Title of Paper	Jumping the fine LINE between species: Horizontal transfer of transposable elements in animals catalyses genome evolution
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Atma M. Ivancevic, Ali M. Walsh, R. Daniel Kortschak, David L. Adelson (2013). Jumping the fine LINE between species: Horizontal transfer of transposable elements in animals catalyses genome evolution. BioEssays 35: 1071-1082. DOI: 10.1002/bies.201300072

Principal Author

Name of Principal Author (Candidate)	Atma M. Ivancevic		
Contribution to the Paper	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	6/12/16

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	Ali M. Walsh		
Contribution to the Paper	Helped to interpret and design one of the figures, and assisted in writing the manuscript.		
Signature		Date	9/12/16

Name of Co-Author	R. Daniel Kortschak		
Contribution to the Paper	Supervised the development of work and assisted in writing the manscuript.		
Signature		Date	6/12/16

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised the development of work and assisted in writing the manscuript.		
Signature		Date	7 Dec. 2016

Jumping the fine LINE between species: Horizontal transfer of transposable elements in animals catalyses genome evolution

*Atma M. Ivancevic, Ali M. Walsh, R. Daniel Kortschak and David L. Adelson**

Horizontal transfer (HT) is the transmission of genetic material between non-mating species, a phenomenon thought to occur rarely in multicellular eukaryotes. However, many transposable elements (TEs) are not only capable of HT, but have frequently jumped between widely divergent species. Here we review and integrate reported cases of HT in retrotransposons of the BovB family, and DNA transposons, over a broad range of animals spanning all continents. Our conclusions challenge the paradigm that HT in vertebrates is restricted to infective long terminal repeat (LTR) retrotransposons or retroviruses. This raises the possibility that other non-LTR retrotransposons, such as L1 or CR1 elements, believed to be only vertically transmitted, can horizontally transfer between species. Growing evidence indicates that the process of HT is much more general across different TEs and species than previously believed, and that it likely shapes eukaryotic genomes and catalyses genome evolution.

Keywords:
lateral transfer; repetitive DNA; retrotransposon; transposon

DOI 10.1002/bies.201300072

School of Molecular and Biomedical Science, University of Adelaide, Adelaide, SA, Australia

***Corresponding author:**
David L. Adelson
E-mail: david.adelson@adelaide.edu.au

Abbreviations:
BLAST, basic local alignment search tool; **BLASTn,** nucleotide BLAST; **dN/dS,** number of synonymous (dS) and non-synonymous (dN) substitutions per site; **EN domain,** endonuclease domain; **HT,** horizontal transfer; **LINE,** long interspersed element; **LTR,** long terminal repeat; **MUSCLE,** multiple sequence comparison by log-expectation; **ORF,** open reading frame; **SINE,** short interspersed element; **TE,** transposable element; **YAC,** yeast artificial chromosome

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Introduction

The evolution of eukaryotic genomes is strongly driven by repetitive DNA derived from dynamic transposable elements (TEs). While TEs are still considered 'junk' DNA because they provide no clear adaptive advantage [1], their ability to invade the genome of their host can have potential mutagenic or speciation effects [2]. They are also the largest class of repeats found in eukaryotes, occupying at least 45% of the human genome [3, 4] and even more in other mammalian species [5]. TEs are typically inherited vertically, passing from parent to offspring with subsequent duplications, but growing evidence suggests that the passage of TEs is not restricted to vertical inheritance. Instead, it is becoming clear that these dynamic elements are able to move horizontally between different organisms as they do within genomes. We hence define horizontal transfer (HT) as the transmission of genetic material between non-mating species [6]. Because TEs influence the duplication and rearrangement of regulatory DNA, and HT increases the occurrence of TE invasions, HT can be regarded as a catalyst for actively transforming genome structure and biological information [7].

All TEs can be divided into two major classes based on how they transpose: (1) retrotransposons, which 'copy and paste' via an RNA intermediate; and (2) DNA transposons, which use a 'cut and paste' mechanism. Retrotransposons can be further categorised as long terminal repeat (LTR) or non-LTR, the latter including long interspersed elements (LINEs) and short interspersed elements (SINEs) [8]. It is believed that LTR retrotransposons evolved from a non-LTR retrotransposon fused with a DNA transposon in early eukaryotes [9]. LTR retrotransposons are similar to retroviruses in terms of replication mechanism and structural organisation (e.g. common *LTR*, *gag* and *pol* genes), but differ because they lack a functional envelope (*env*) gene, present in retroviruses and allowing those elements to move

easily between species by infecting new cells [10, 11]. It has, however, been found that a small group of LTR retrotransposons encode an extra open reading frame in the same position as the retroviral *env* gene and thus have infectious properties (e.g. the *gypsy* element in *Drosophila* [12]). Without virus-like envelope proteins, other TEs require a vector to facilitate HT. DNA transposons and LTR retrotransposons, unlike non-LTR retrotransposons, also have a more stable double-stranded DNA intermediate, and so are more likely to be capable of HT [7].

The most extensively studied case of HT is that of DNA transposons in *Drosophila* and other insects [13–16]. More recently there have been investigations into the HT of DNA transposons such as *SPIN* and *OC1* in vertebrates [17–19]. But perhaps most surprising is the emerging evidence of widespread HT of BovB non-LTR retrotransposons, complete with two plausible arthropod vectors [20]. These studies effectively eliminate the assumption that HT is restricted to LTR retrotransposons or retroviruses. They also raise the possibility that other non-LTR retrotransposons, such as L1 or CR1 elements, have undergone similar HT events. Recent evidence shows that HT of TEs (including non-LTR retrotransposons) is much more widespread and frequent than previously believed, affecting a broad range of organisms through numerous potential vectors.

Autonomous elements are more likely to transfer horizontally than non-autonomous elements

Autonomy in transposition refers to whether or not an element encodes the factors required for its own mobilisation. Both class 1 and class 2 TEs can be either autonomous or non-autonomous. For example, in terms of non-LTR retrotransposons, there are autonomous elements such as L1 LINEs with two open reading frames, which, respectively, encode an RNA binding protein and a protein with endonuclease and reverse transcriptase properties [21]. Other autonomous LINEs include L2, MIR, CR1 and BovB. In contrast, non-LTR SINEs such as *Alu* repeats are non-autonomous and thus rely on LINE-encoded retrotransposition machinery. Similarly, LTR retrotransposons are considered autonomous if they encode all the protein-coding domains necessary for transposition, or non-autonomous if they lack some or all protein-coding domains but display evidence of amplification capability [22]. The same principle applies to DNA transposons: autonomous elements have an intact gene encoding an active transposase enzyme, while non-autonomous elements require transposase from a functional TE [22]. In each case, the non-autonomous element transposes by hijacking the transposition machinery of its autonomous partner. This means that although both autonomous and non-autonomous TEs can be activated given the necessary proteins and intact *cis*-acting sequences, autonomous elements are more likely to transfer horizontally because they encode their own proteins [7].

L1 elements influence genome function and evolution

L1 elements are abundant in mammals, comprising up to 20% of a typical mammalian genome and contributing another 30% or so through SINE amplification and pseudogene processing [23]. In humans, full-length L1s are about 6 kb long and consist of a conserved internal promoter for RNA polymerase II in the 5′ UTR region, two retro-transposing open reading frames (ORF1 and ORF2) separated by an intergenic spacer, and a 3′ UTR ending with a poly-A tail [24]. Most L1 copies are truncated at the 5′ end following transposition [25] or have accumulated various mutations over time, leading to inactivation.

Highly active L1s are rare but account for the majority of retrotransposition in humans

These retrotransposons are considered one of the most active elements in the human genome [21], either because they are too young to have acquired mutations [25] or they have somehow escaped mutational and epigenetic suppression [26]. Out of approximately 7,000 full-length L1 elements in the human reference genome [26], it is believed that about 80–100 copies are potentially still active in any human [27]. But as Brouha et al. [27] showed, only a small minority of these active L1s, known as 'hot' L1s, are highly active in the genome. By cloning 90 intact L1s from the human reference genome and assaying them for activity, they found that six hot L1s accounted for an exceptional 84% of total retrotransposition capability. It was further found that four of the five known disease-causing L1s had activity matching these hot L1s. So while there are many active L1 copies, it is the few hot L1s that contribute most to retrotransposition in the human population [27].

L1s are capable of somatic retrotransposition

Recent studies have shown that L1 retrotransposition is not restricted to the germline, but can also occur in early development. Evidence of this emerged by looking at mice carrying L1 transgenes: for example, Kano et al. [28] created L1 transgenic rodent models to demonstrate that although both germ cells and embryos contain abundant L1 RNA, most insertion events are somatic and not inherited. This phenomenon has further been studied in humans; Rangwala et al. [26] examined how L1 elements affect the transcriptome of human somatic cells by cloning out expressed sequence tags corresponding to 5′ and 3′ L1 flanking regions. Using human lymphoblastoid cell lines, they were able to isolate expressed sequence tags for 692 distinct L1 element sites (410 full-length); verifying the large number of L1 sites expressed in human somatic cells. Possibly the most remarkable evidence of somatic insertions is that characterising L1 retrotransposition in human neural progenitor cells (NPCs). Coufal et al. [29] used qPCR to detect increased L1 copy numbers (approx. 80 extra L1 insertions/NPC) in the hippocampus and other regions of the adult brain, compared to L1 copy numbers in the heart and liver. More recently, over 7,000 potential somatic L1 insertions were identified in human hippocampus samples via retrotransposon capture arrays and sequencing, along with the first reported *Alu* and SVA insertions [30]. Such studies confirm that the hippocampus is predisposed to

somatic L1 activity. However, more research is needed to understand the effects of these somatic L1 insertion events, particularly in the brain [31].

Immobile L1s are thought to affect chromatin and transcription regulation

The importance of immobile L1 retrotransposons should not be overlooked either. Being an autonomous element, L1 is capable not only of its own retrotransposition (in *cis*) but also the *trans*-mobilisation of non-autonomous elements such as *Alus* and SVAs, or cellular mRNAs to form processed pseudogenes [26]. And it has been proposed that if a truncated or mutated L1 sequence can still be transcribed, then this *trans*-mobilisation may not need an active ORF1 [32]. For instance, transcriptionally active L1 elements are thought to be involved in neocentromere activity regulation: LINE RNA contributes to the structure and function of neocentromeric chromatin, possibly acting as an epigenetic determinant in chromatin modification [33]. Even L1s without function in either ORF can provide promoter or polyadenylation sites, affecting transcriptional regulation in different parts of the genome [26, 34]. Both active and inactive L1 elements can change the structure and function of human genomes.

L1 distributions vary within genomes and species

Within a genome, intact active and inactive L1 elements are usually present in similar genomic regions [27]. They are often said to congregate at AT-rich and gene-poor genomic regions [23], though we have not observed this to be a general correlation (unpublished data). A correlation analysis of several different species groups revealed that, while this seems to hold for humans and rodents, L1s in other species do not seem to preferentially home to these areas. In fact, L1 elements in horse and elephant showed the opposite correlation: preference towards GC-rich and high gene density areas. These differences may be due to different epigenetic factors in each species (e.g. chromatin state), which are known to influence L1 insertional preference [35]. Monotremes such as platypus do not contain L1s, but in regards to their L2 content, there did not seem to be a consistent bias towards AT-rich and gene-poor regions. Interestingly, BovB elements in elephant showed an opposite bias to L1 elements, which was different again from that in the bovine and opossum. So while L1 elements are assumed to show distribution bias to particular regions based on nucleotide content and gene density, this does not hold for LINEs across or even within the same species.

L1 elements also exhibit a ubiquitous distribution across species, present in all eutherian mammals examined to date and hence believed to have been introduced in the genome before mammalian radiation [36]. However, whilst all mammals have ancestral L1 elements, some species (e.g. rat, Tasmanian Devil) do not have mobile L1s [36, 37]. Furthermore, most mammalian species examined phylogenetically only seem to have a single lineage of L1 families [38–40], although Casavant et al. [41] supported the persistence of more than one L1 lineage in deer mice. Khan et al. [42] investigated this restricted distribution of a single lineage in humans, and discovered that from about 70 million years ago (Mya) to 40 Mya, there were in fact three distinct L1 lineages simultaneously active in ancestral primates. It is only in the last 40 million years that one family has evolved to dominate the replicative process [42]. They deduced that only families with different 5′ UTR could coexist for long time periods, presumably because they do not compete for the same host-encoded transcription factors. In contrast, there are over 30 distinct and active L1 lineages in fish [40]. This suggests that L1 copy number is strictly controlled in fish, as opposed to the thousands of copies fixed in an L1 mammalian family. Novick et al. [43] further showed that the lizard *Anolis carolinensis* had an L1 length distribution more similar to fish than mammals, indicating that mammals and non-mammal vertebrates react differently to retrotransposition. Thus despite originating from the same ancient L1 clade [8], there are now highly divergent L1 sequences among different species.

Is there evidence of L1 horizontal transfer across species?

L1 retrotransposons are currently not believed to have been horizontally transferred. On the contrary, Schaack et al. [7] used mammalian L1 elements as an example of exceptional vertical endurance over the past 100 million years, supported by evolutionary analyses of mammals [40, 42]. Waters et al. [25] came to a similar conclusion after examining sequences from the 3′ region of the reverse transcriptase from 21 mammalian species. They noticed that there were active autapomorphic groups of L1 in Afrotheria, Xenarthra and Boreoeutheria (i.e. AfroLINEs, XenaLINEs and BoreoLINEs) forming three major clades of L1, but each clade corresponded to a main placental lineage. So the observed active L1 lineages followed expected species relationships [25].

But there are inconsistencies in L1 studies that have not yet been addressed. Because of the ubiquitous distribution of L1, most of the current data comes from human or mouse genomes and is simply assumed to hold for all placentals [25]. Many studies [25, 42, 44] also use sequences only from the 3′ end, rather than full-length elements, and have relatively small sample sizes of species. Without extracting sufficient L1 data from many host species, it is difficult to create an accurate L1 phylogeny. A further discrepancy in the results of Waters et al. [25] is that some very distantly related species show high L1 sequence similarity (Fig. 1). For example, three orthologous L1 sequences (with 98–99% identity) were identified between human and chimpanzee: HSA_3 and PTR_2 elements fell within the primate L1, as expected; but the older HSA_2 and PTR_1 fell at the base of Boreoeutheria; and ancient HSA_4 and PTR_3 fell with other ancient elements near the root of the tree. Even more surprising is the unexplained phenomenon where MAMU_1 (Rhesus monkey) appears most closely related to LAF (African elephant) elements on their phylogenetic L1 tree. This raises the possibility of HT. Growing evidence supports the

Review essays

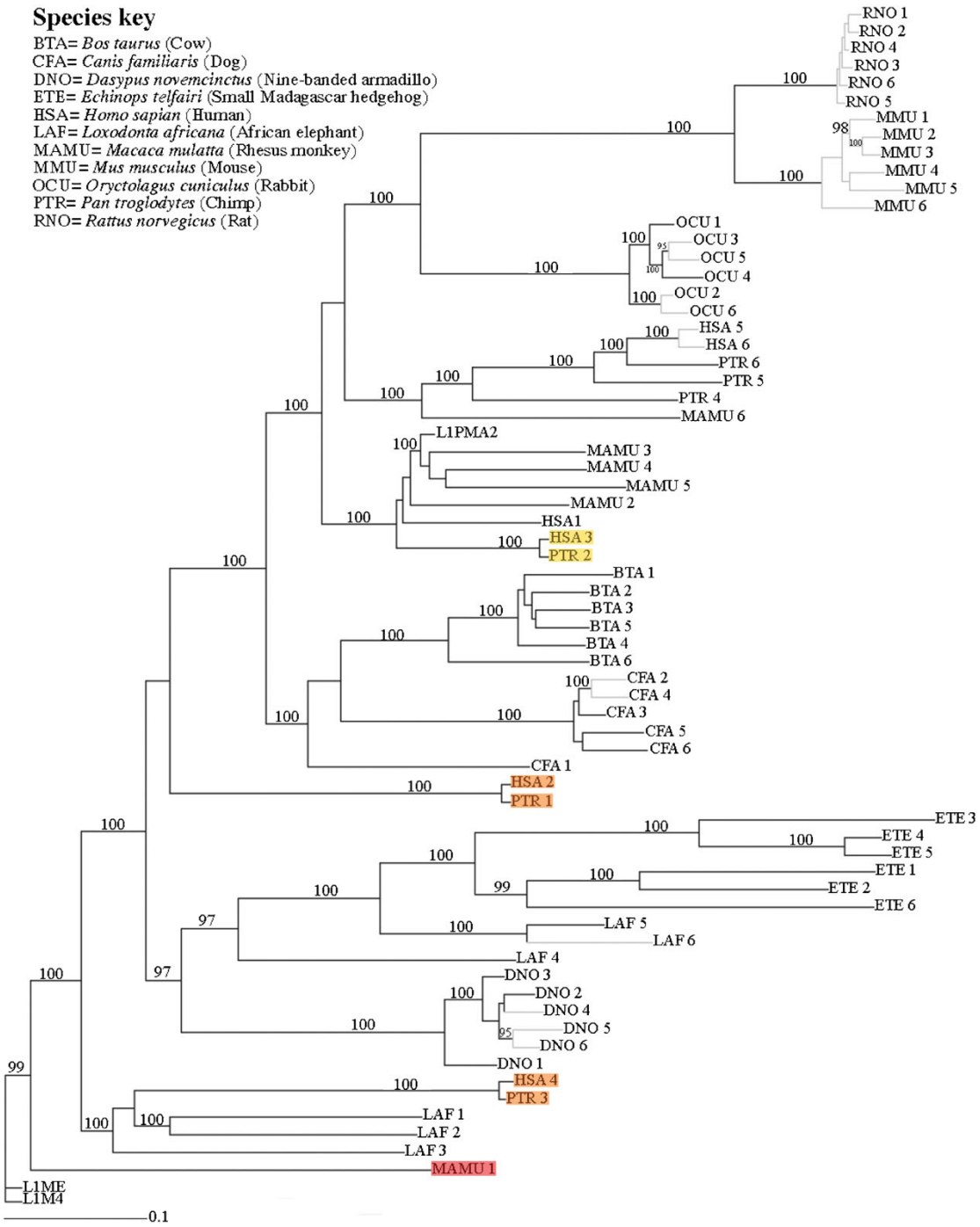


Figure 1. Phylogenetic tree of L1 reproduced from Fig. 2 [25]. Bayesian consensus tree generated by a GTR invariant-sites plus Γ model, applied to 69 (long) sequences. Grey branches indicate sequences with >98% homology to other L1 copies in their respective genomes. Highlighted sequences HSA_3 and PTR_2 (yellow) fall within the primate L1, as expected. Anomalies are shown in orange or red: HSA_2 and PTR_1 (orange) fall at the base of Boreoeutheria; ancient HSA_4 and PTR_3 fall with other ancient elements (orange); and MAMU_1 (red) falls next to LAF elements. A species key shows the abbreviated names, scientific names and common names. Note that L1M4, L1ME and L1PMA2 are consensus sequences.

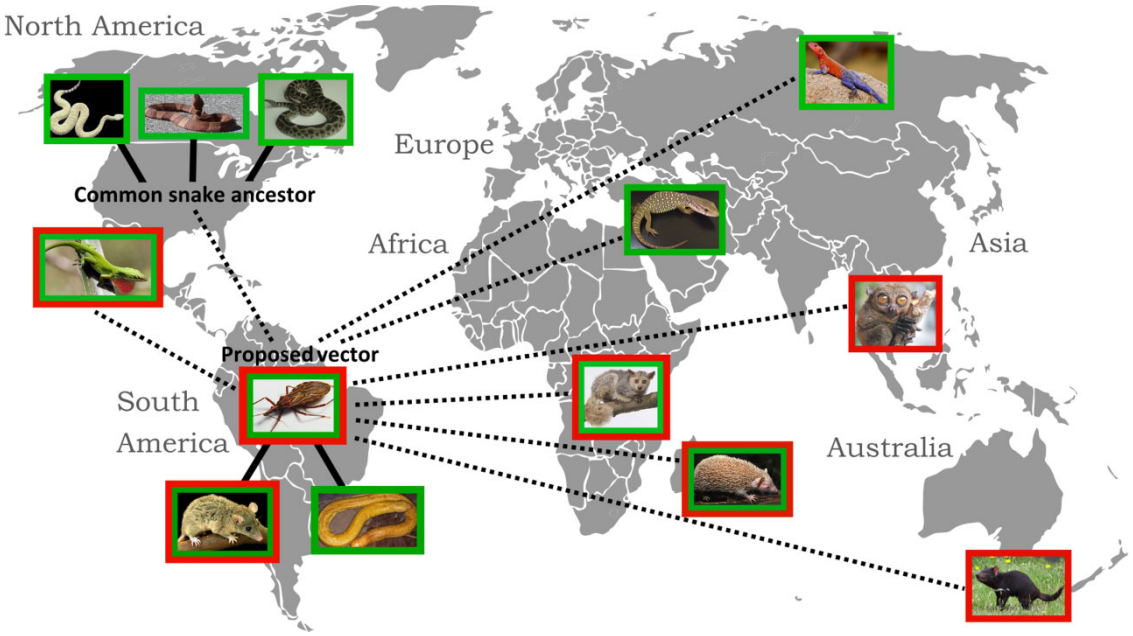


Figure 2. Map depicting the pandemic-like horizontal transfer of DNA transposons SPIN and OC1 across species and continents, thought to have occurred in the last 50 million years. Outlined colours indicate which element is present in that species: green for SPIN elements, red for OC1 elements. *Rhodnius prolixus* has been identified as a possible vector for these HT events because it contains SPIN and OC1 elements with >98% identity and clusters phylogenetically with the distantly related opossum and red worm lizard. This suggests a host-parasite type of HT, which can spread through any of the hosts that the bug feeds on. However, if the current distribution of *R. prolixus* reflects its origin, given that it is only found in Central and South America, other intermediate vectors are needed to explain the transoceanic movement of SPIN and OC1 transposons to Asia, Africa and Australia. Categorised by continent, the species included on the map are: *R. prolixus*, *Monodelphis domestica*, *Amphisbaena alba* (South America); *Agkistrodon contortrix*, *Sistrurus catenatus*, *Crotalus atrox*, *Anolis carolinensis* (North America); *Agama agama*, *Varanus exanthematicus*, *Tarsius* (Asia); *Otolemur garnettii*, *Echinops telfairi* (Africa); and *Sarcophilus harrisii* (Australia) [6, 17–19, 57]. This map is for illustrative purposes only and not meant to serve as a comprehensive phylogeographic reference.

presence of more ubiquitous lateral transfer in genome evolution and diversification.

Retroviruses encode infective machinery conducive for horizontal transfer

LTR retrotransposons (including retroviruses) and DNA transposons are thought to have a greater likelihood of horizontally transferring between species than non-LTR retrotransposons [7, 45]. Retroviruses are infective because they have encoded envelope-like proteins which can recognise host surface receptor proteins and thus penetrate the plasma membrane [46]. They are often used to reveal phylogenetic relationships; populations sharing a retrovirus in the same genomic location must be related, because infection of an endogenous retrovirus into the host genome is irreversible [47]. Retroviruses are also known to be highly active in mice, with continued retrotransposition resulting in very high levels of insertional polymorphism [48]. Because of their infective machinery, there is no doubt that retroviruses undergo HT from one genome to another before they are vertically inherited.

DNA transposons are known to transfer horizontally in insects

HT of transposons, such as *P* and *mariner* elements, has been extensively examined in insects. Daniels et al. [13] described the patchy distribution of *P* elements found in *Drosophila melanogaster* but otherwise absent from the *melanogaster* subgroup, indicating HT from donor species *D. willistoni* (with almost identical *P* elements) specifically into *D. melanogaster*. This is supported by the proposition of mite *Proctolaelaps regalis* as a potential HT vector, as *P. regalis* samples were shown to contain both the required *P*-sequences and *Drosophila* ribosomal DNA sequences [14]. Since then, this case of HT has been confirmed a number of times by molecular analyses [49, 50]. Further research showed that HT of *P* elements is by no means restricted to *D. willistoni* and *D. melanogaster*. For example, Loreto et al. [16] identified a canonical *P* element in *D. mediopunctata* homologous to that in *D. melanogaster*; the first to be found outside the subgenus *Sophophora*. The most likely explanation for this is that the *P* element entered *D. mediopunctata* around the time it infiltrated the *saltans* and *willistoni* groups, again via HT.

Perhaps the best example to demonstrate the ability and range of HT in insects is the *mariner* transposon. Maruyama and Hartl [15] explored the presence of *mariner* elements in the genus *Zaprionus*, the only instance found outside of the *D. melanogaster* species group. They found support for HT by observing that there was higher similarity of *mariner* sequences between distantly related species (*Zaprionus tuberculatus* and *D. mauritiana*) than between closely related species within the *melanogaster* species group (*D. tsacasi* and *D. mauritiana)*. They then built a species phylogeny using conserved *alcohol dehydrogenase (Adh*) sequences and compared it with the *mariner* gene phylogeny; the only inconsistency was in the position of the *Z. mariner* sequence. Knowing that the *Adh* gene is vertically inherited, this strongly indicates HT of the *mariner* element. Many subsequent studies presented evidence of *mariner* elements being horizontally transferred between different insect species: for example, Lohe et al. [51] investigated *mariner* HT between flies and fleas, and Lampe et al. [52] reported recent HTs involving *mariner* elements from insects of four separate orders (European earwig, European honey bee, Mediterranean fruit fly, and a blister beetle). As such, it is thought that HT plays a crucial part in the *mariner* replication cycle, preventing its extinction by introducing it to new hosts [53].

There is abundant evidence indicating that transposons have been horizontally transferred in and between insect species. For *Drosophila* alone, the literature presents over 100 recognised HT events spanning 21 different elements (52.4% DNA transposons, 42.6% LTR retrotransposons and 5% non-LTR retrotransposons) [54]. However, this list of cases does not accurately represent the importance of such events in genome evolution, as it draws on a sample of relatively simple organisms [55]. Recent research has progressed to investigating the impact of HT in vertebrates.

DNA transposons show repeated horizontal transfer between animal species

There are many different ways to test for lateral transfer of TEs. Schaack et al. [7] summarised this by defining three criteria for inferring HT: (i) sporadic distribution of the TE within a set of taxa; (ii) exceptional degree of nucleotide similarity over widely divergent species; and (iii) differences in TE and host phylogenies. While combinations of these three criteria have been used in the past, ideally there should also be evidence disproving that the observed patterns resulted from some other evolutionary process or mechanism [7]. For instance, comparison of synonymous to non-synonymous substitutions can reveal how the TE is evolving after insertion in the genome, as used by Pace et al. [17] to dispute vertical inheritance of SPIN (space invader) sequences in tetrapods. SPIN elements are DNA transposons of the *h*AT (*hobo/Activator/Tam3*) superfamily, which are known for their ability to move among a wide range of heterologous species and different conditions [56]. Pace et al. [17] were the first to identify these transposons and thus provide substantial evidence of lateral transfer in seven tetrapod lineages, including the mammalian tenrec, opossum, bushbaby, little brown bat, murine rodents and non-mammalian anole lizard and African clawed frog. They initially inferred HT from the patchy distribution and high sequence similarity of full-length *SPIN* transposons in these tetrapods (ranging from 84 to 99% and averaging 96% pairwise nucleotide identity between any two species). But while this suggests preservation by purifying selection, the SPIN evolutionary pattern displayed a lack of selective constraint or bias towards synonymous substitutions, indicating neutral evolution. Rather than vertical transmission from a common ancestor, this suggests that active SPIN progenitors were horizontally transferred and then amplified within these lineages [17].

Horizontal transfer of transposons spans species and continents

Recent studies have expanded on this research to report large-scale HT of DNA transposons in vertebrates. Gilbert et al. [18] tripled the number of known HT cases in tetrapods by showing that, as a result of at least 13 independent HT events, SPIN has colonised 17 species of reptiles representative of nearly every major lineage of squamates (Fig. 4A). They were also able to increase the geographic range of SPIN HTs: earlier estimates [6, 17, 57] placed the HT events in Africa, Eurasia and South America, but with these results it seems feasible that there was at least one transoceanic transfer extending to North America and possibly Asia (Fig. 2). A DNA transposon frequently associated with SPIN is the OposCharlie1 (OC1), which has previously had reported HT events in Asia, Africa and South America [6]. But Gilbert et al. [19] added Australia as the fourth continent of OC1 HT by examining a new case of lateral movement in the Tasmanian devil and other marsupials. They were able to deduce that OC1 had infiltrated a total of 12 distinct animal lineages. Both of these studies [18, 19] used the same method of inferring HT as Pace et al. [17]: a high degree of nucleotide identity across the respective full-length elements and neutral evolution after genome insertion, seen by the dN/dS values and lack of evidence for purifying selection. More importantly, their results provide further evidence for the most widespread cases of HT in eukaryotes, accentuating the pandemic-like effect of transposon invasions [19].

New evidence supporting widespread horizontal transfer of BovB non-LTR retrotransposons

HT is thought to occur rarely in non-LTR retrotransposons [45, 58]. Unlike DNA transposons, retrotransposons have a relatively unstable RNA intermediate that is reverse-transcribed directly into the chromosomal target site, so transfer outside the cell nucleus is decreased [58]. Nonetheless, there are several possibilities for HT of non-LTR retrotransposons: (1) RNA-mediated HT, involving the use of a virus as a vector for RNA transcript packaging [7, 59–62]; (2) DNA-mediated HT, where the retrotransposon inserts into a DNA transposon and the resulting construct (chimeric element) is horizontally

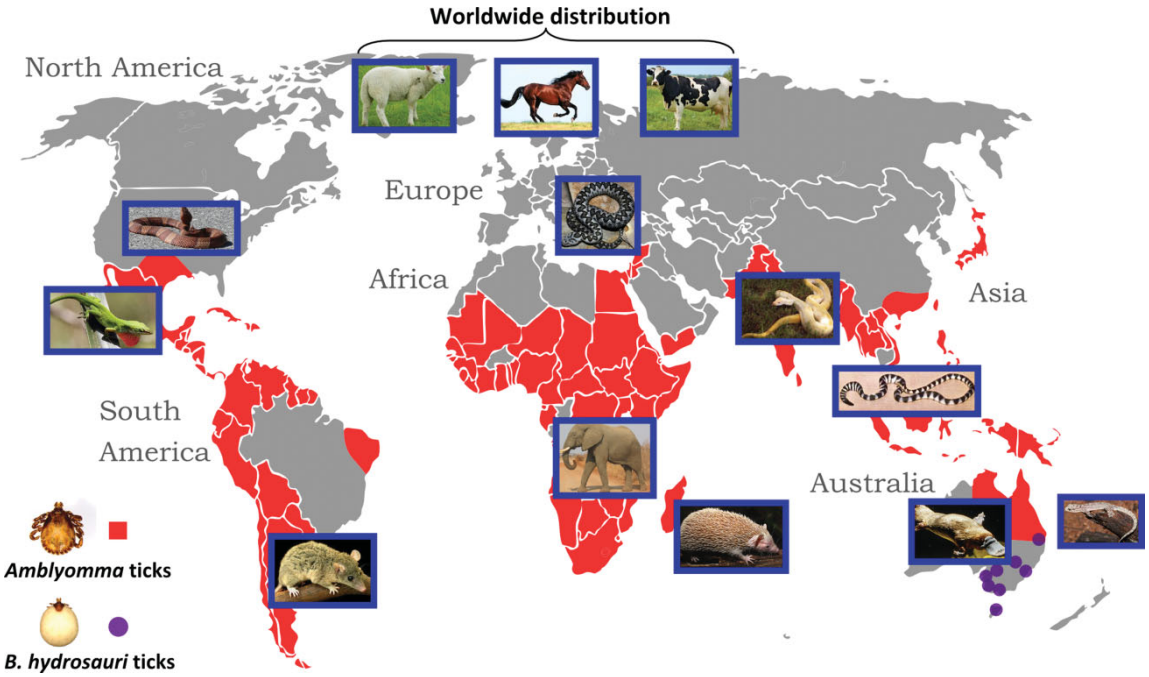


Figure 3. Map showing the overlap of potential vectors and species bearing BovB retrotransposons. The blue outlines indicate the presence of BovB in species. Shading represents the location of reported tick cases: red for *Amblyomma* tick species [70–72], purple for *Bothriocroton hydriosauri* ticks [73]. Categorised by continent, the species depicted on the map are *Monodelphis domestica* (South America); *Anolis carolinensis*, *Agkistrodon contortrix* (North America); *Loxodonta africana*, *Echinops telfairi* (Africa); *Hydrophis spiralis*, *Python molurus* (Asia); *Vipera ammodytes* (Europe); *Ornithorhynchus anatinus*, *Christinus guentheri* (Australia) [20]. Domestic animals such as *Equus caballus*, *Bos taurus* and *Ovis aries* are also shown, with a worldwide distribution. This map is for illustrative purposes only and not meant to serve as a comprehensive phylogeographic reference.

transferred [63, 64]; or (3) transfer of 'naked' DNA and RNA circulating in animal bodily fluids through some sort of vector [7]. Many reported cases have shown that both RNA-mediated and DNA-mediated elements are able to effectively cross species boundaries [7, 59–64].

Introduction to BovB and its distribution

HT is the best explanation for the sporadic distribution of BovB retrotransposons. BovB is a LINE about 3.2 kb long, originating in squamates [65] but nowadays found in a wide range of genomes including ruminants, marsupials, monotremes and mammals [5, 20, 66, 67]. The first suggestion of potential HT was the observation that the BovB sequences found in ruminants are very similar to those in snakes (especially vipers) and lizards [68]. A wider analysis of vertebrate classes [65] revealed that BovB was absent from most mammals and reptile species, yet present and highly similar in monotremes (platypus) and marsupials (opossum and tammar wallaby). Such a discontinuous phylogenetic distribution cannot be explained by vertical inheritance. Instead, it was proposed that three independent HT events had occurred between squamate, ruminant and metatherian ancestors to produce the observed BovB topology [69].

Widespread horizontal transfer of BovB across taxa

Walsh et al. [20] recently showed that lateral transfer of BovB is much more widespread than previously believed. By analysing all publicly available genomes for full-length BovB sequences, they were able to build the most extensive phylogenetic tree of BovB sequences to date (Fig. 4B). It was found that the extent of the differences between BovB and species phylogenies (Fig. 5) could not be explained without at least nine HT events, far surpassing previous estimates for BovB [65, 66, 69]. The BovB tree showed that distantly related species, e.g. snake and opossum, or tick and lizard, displayed an unusually high percentage identity. More closely related species, such as cow and horse, did not show as much BovB sequence similarity; in fact, it was found that the horse BovB grouped with the BovB subfamily from the Howe Island Gecko instead, and that both clustered with the Afrotheria and monotremes. This is indicative of lateral transfer. In the squamate lineage, there was evidence that BovB was moving both horizontally and vertically: all reptile species examined showed significant BLAST BovB hits and generally grouped together as expected (e.g. skinks formed a robust group, as did most snakes), yet the presence of two tick species in this clade with squamate-like BovBs supported the occurrence of HT. So Walsh et al. [20] were able to infer HT of BovB in the evolution of life, even presenting two plausible HT vectors in reptiles and possibly ruminants and marsupials (Fig. 3).

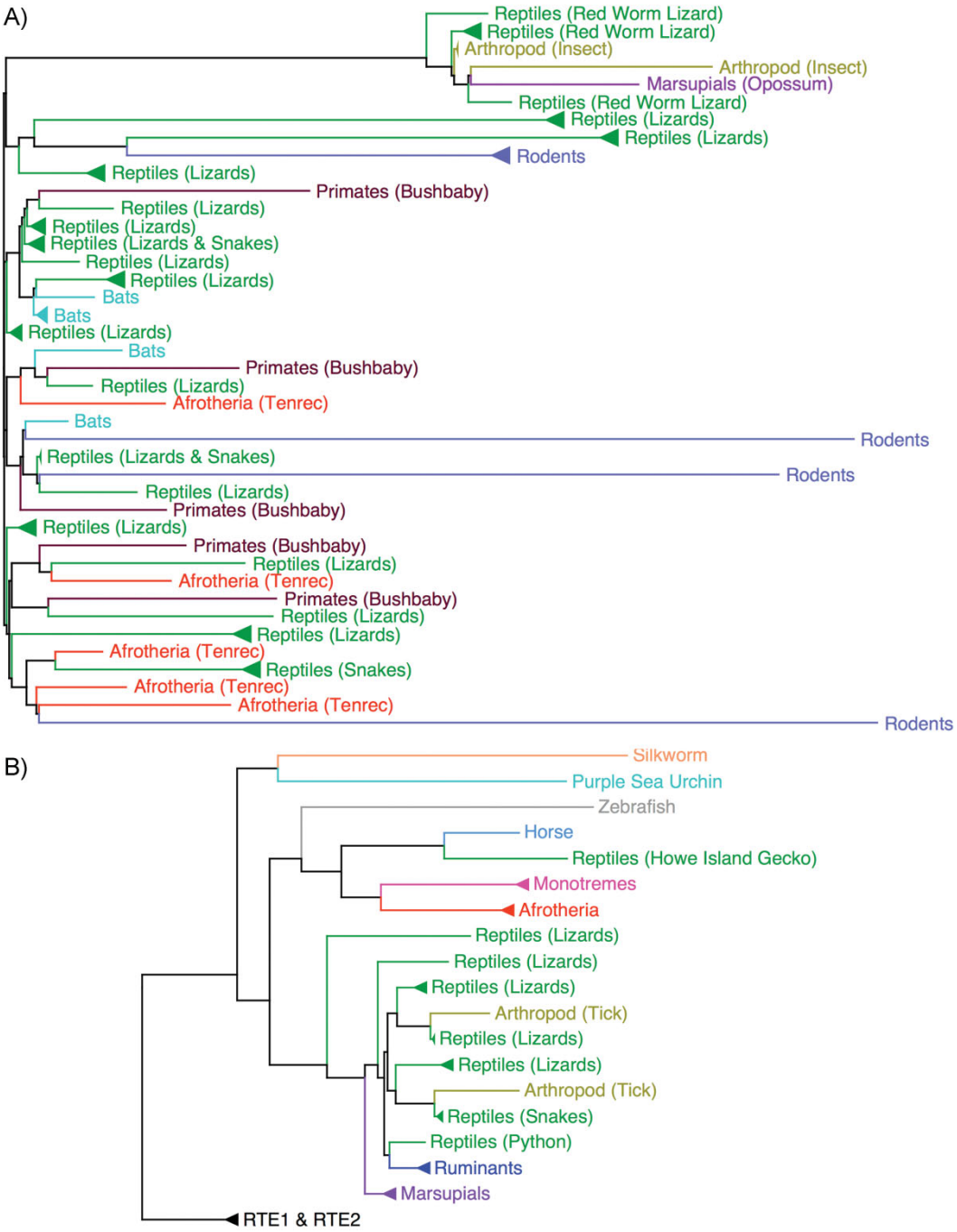


Figure 4. Phylogenetic trees showing the respective distributions of SPIN and BovB across taxa. **A:** Tree of SPIN elements, reconstructed using the publicly available data from Gilbert et al. [18]. Includes all autonomous squamates SPIN sequences sequenced in the study plus five previously characterised SPIN species sequences. Phylogenetic analyses were carried out using MUSCLE for sequence alignment and FastTree for maximum likelihood tree construction. Note that a BLASTn of the SPIN superconsensus sequence (supplied by Gilbert et al. [18]) with sensitive algorithm parameters came up with a total of 254 Blast hits in various Afrotheria species (not just Tenrec) and 116 Marsupial hits, suggesting that the SPIN distribution is more diverse than this tree indicates. **B:** Maximum likelihood tree showing the distribution of BovB across taxa, reproduced from Walsh et al. [20]. Also built using MUSCLE for alignments, Gblocks for processing to limit the effect of indels, and FastTree for tree construction from full-length BovB sequences extracted from full genome sequence and those constructed from low coverage reads. Taxa and branches across both trees are coloured taxonomically, with marsupials in purple, reptiles in green, arthropods in yellow, Afrotheria in red, ruminants in dark blue, etc.

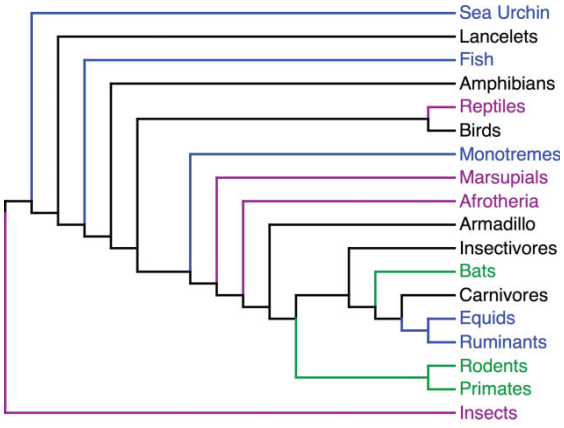


Figure 5. Tree built from orthologues for comparison to the phylogenetic trees built from SPIN and BovB sequences. Colours indicate the taxonomic groups that have both SPIN and BovB (purple), only BovB (blue), only SPIN (green) or neither (black).

TE insertion and amplification is species-specific

Another significant finding was that the abundance of BovB elements varied greatly among different species [20]. The highest BovB percentages were found in cow (18% BovB coverage, although this is an underestimate because it does not include BovB SINE sequences derived from other sources), sheep (15%) and Afrotherian genomes (11% in elephant, 7% in rock hyrax and 8% in tenrec). In contrast, the anole (reptile), marsupials and monotremes all showed BovB coverage of around 1%. Even more surprising is the extremely low BovB copy number seen in horse (0.11%) and zebrafish. Walsh et al. [20] stated that only 31 full-length instances of BovB were found in the horse genome, despite the presence of horse-specific SINEs in some full-length BovBs indicating that BovB had been in horse for some time. Due to limited genomic data available, they were unable to test other equine species. BovB was not found at all in camelids, tuatara, turtles, mosquitos, birds and other mammals. Given that no SPIN transposons were found in some of these species either (e.g. turtles or crocodiles), it may be that these organisms have developed a better genomic defence against the insertion or amplification of TEs, or lack interaction with putative HT vectors [18]. There is also the question of whether animals should defend against TE insertion at all – perhaps some organisms allow HT of TEs because it gives them an evolutionary selective advantage (while suppressing the disadvantages). More research is needed to understand why BovB and other TEs successfully colonise some genomes but not others.

Horizontal transfer of TEs is complicated by CR1 repeats

Chicken repeat 1 (CR1) is an ancient non-LTR retrotransposon, abundant in reptiles and birds. It is of particular interest in regards to the HT of BovB because BovB_VA, a sequence extracted from the horn-nosed viper (*Vipera ammodytes*), has CR1 fragments on both ends of the full-length BovB element [20]. This means that at some point in the evolution of BovB, it was inserted into a CR1 type element and since then has copied itself throughout genomes with the CR1 ends attached. As Walsh et al. [20] discovered, this complicates the construction of other squamate BovB consensus sequences and can result in false positive BovB hits in bird or reptile genomes. The only way around this was to remove all CR1 fragments before assembling BovB consensus sequences for these species, meaning that it could not be determined when the squamate BovB lineage had acquired the CR1 ends. However some squamates, namely the python and copperhead snakes, did not have BovB elements with CR1-like ends [20]. Although additional sequencing in a wider range of reptiles would be needed to confirm a discontinuous distribution, this does present the possibility that the CR1-flanked BovB elements were horizontally transferred.

Some TEs appear to act as vectors for others

The fact that CR1 elements by themselves are known to transfer horizontally among insect species supports this possibility. Novikova et al. [74] notably found that a CR1 family in lycaenid butterflies of the genus *Maculinea* was remarkably similar to CR1 elements in the distantly related Bombycidae moths: silkworm *Bombyx mori* and *Oberthueria caeca*. However, no similar CR1-like elements were found in the taxa closely related to *Maculinea*. Divergence-versus-age analyses confirmed that these CR1 elements did not diverge at the same time as their host taxa, so the most likely explanation is HT [74]. This study was recently extended to investigate whether DNA transposons, such as *mariner* elements, can act as natural HT vectors for these CR1 elements [63]. The results indicated that recurrent lateral transfer of *mariner* and CR1 elements recently occurred between Lepidoptera species. This may be because the CR1 elements are located next to *mariners*, as they are with BovBs, and thus get transferred between butterflies and moths as a single DNA fragment. Or it could be that the *mariner* elements are actually facilitating the HT of CR1 elements: a theory supported by several chimeric CR1/*mariner* sequences found in *Maculinea* and *Bombyx* genomes, which could be left over from the transposon-based vectors. Either way, Sormacheva et al. [63] demonstrated that frequent and possibly simultaneous HT of TEs can occur between distantly related insects. This provides another example where CR1 elements are involved in the HT of another retrotransposon, and by extension, suggests that different types of TEs may be able to use each other as vectors.

Proposed vectors that facilitate horizontal transfer of TEs

The HT of DNA transposons and retrotransposons is important to look at because it is becoming increasingly clear that HT is much more widespread than originally thought. Even though replication mechanisms differ between DNA transposons and retrotransposons, the vectors used might be the same. For HT

to be successful, DNA needs to be transported from donor to host cell (and into the germline for multicellular organisms) and then integrated into the host genome [7]. However, despite numerous proposed vectors in the literature (Table 1), the exact mechanisms that TEs use to move between genomes are still unknown. It is also important to note that the proposed vectors are not mutually exclusive, because HT events are not restricted to any single mechanism [54].

How can exaptation of transposable elements affect gene regulation?

Exaptation is the term used to describe the phenomenon where TEs, usually viewed as 'junk' DNA, actually acquire a new function in the genome [79]. For example, studies using YAC transgenic mice and reporter assays in cell cultures have shown that L1 elements have exapted as enhancers of the human *apoa* gene [80]. Similarly, part of the BovB retrotransposon (the EN domain) has been exapted into the *Bucentaur (Craniofacial development protein 2)* gene, providing a protein coding function in all ruminants [81]. In fact, Lowe and Haussler [82] assert that about 20% of gene regulatory sequence in the human genome were co-opted from SINE, LINE, LTR and DNA transposon insertions, with recent reviews summarising these exaptations [21, 83]. The exaptation of TEs serves as a good reminder that HT is just the beginning; much more research is needed to determine the full impact of these TEs on the genome.

Conclusions and outlook

Growing evidence indicates that HT has played a significant part in the evolution of animal genomes. Virtually all classes of TEs can undergo HT across widely divergent species, from retroviruses with envelope-like proteins to DNA transposons and even non-LTR retrotransposons. Recently, studies have progressed to exploring the large-scale HT of elements in vertebrates. For example, Gilbert et al. [18, 19] found that the combined HT of DNA transposons SPIN and OC1 spanned many animal species worldwide. The new evidence showing widespread HT of BovB [20] further accentuates how frequently this phenomenon can occur and counters the belief that non-LTR retrotransposons are incapable of HT. This suggests that our understanding of the full impact of HT on genomic change has not yet been realised. Further research

Table 1. Proposed vectors of HT

Type of vector		Name of vector(s)	Type of TE that is horizontally transferred	Name of TE(s)	Species involved	Reference
Arthropod	Reptile ticks	*B. hydrosauri* and *A. limbatum*	Non-LTR retrotransposon	BovB	Reptiles, and possibly ruminants and marsupials	[20]
	Insect	*R. prolixus*	DNA transposon	SPIN, OC1, hAT1, ET	Invertebrates and vertebrates	[6]
	Insect	*D. hydei*	DNA transposon	*Minos*	Drosophila species (*repleta* and *saltans*)	[75]
	Mite	*P. regalis*	DNA transposon	*P* element	*Drosophila* species (*D. willistoni* and *D. melanogaster*)	[14]
Virus	Poxvirus	TATV	Non-LTR retrotransposon	Sauria SINE	Snake (*Echis ocellatus*) and West African rodents	[61]
	Baculovirus	*Autographa californica* nuclear polyhedrosis virus (NPV)	LTR retrotransposon	TED	Moth (*Trichoplusia ni*)	[76]
	Baculovirus	*Cydia pomonella* granulovirus (CpGV)	DNA transposon	TCp3.2	*Caenorhabditis elegans* and *Cydia pomonella* granulovirus	[59]
	dsDNA virus	Polydnaviruses	DNA transposon	*Mariner*	Parasitoid *braconid* wasps and *lepidopteran* hosts	[77, 78]
Other	Freshwater snail	*Lymnaea stagnalis*	DNA transposon	SPIN	Invertebrates and vertebrates	[6]
	DNA transposon	*Mariner*	Non-LTR retrotransposon	CR1	Butterflies and moths (i.e. *Maculinea* and *Bombyx)*	[63]
No vector	–	–	Retrovirus	*gypsy*	*Drosophila* species	[12]

A brief overview of different eukaryotic and viral HT vectors presented in the literature, classified according to type and name of vector, type and name of transposable element that is horizontally transferred, the species involved and source citation. The numerous vector types suggested show that much is still unknown regarding the mechanisms by which TEs are transferred.

should look at using comprehensive, genome-wide scans to investigate the distribution of other retrotransposons, particularly since there are unexplained discrepancies in past studies (e.g. [25]). This work would help determine how general the process of HT is across different types of TEs and species, and narrow down the potential vectors that facilitate the spread of TEs. Finally, at present we are only able to detect HT after it reaches the germline. But it is intriguing to consider the possibility of widespread somatic HT, especially since significant somatic L1 retrotransposition has been detected in neural cells [29, 30].

References

1. **Wong GK, Passey DA, Huang Y, Yang Z**, et al. 2000. Is "junk" DNA mostly intron DNA? *Genome Res* **10**: 1672–8.
2. **Jurka J, Bao W, Kojima KK**. 2011. Families of transposable elements, population structure and the origin of species. *Biol Direct* **6**: 44.
3. **de Koning AP, Gu W, Castoe TA, Batzer MA**, et al. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**: e1002384.
4. **Lander ES, Linton LM, Birren B, Nusbaum C**, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
5. **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**: 12855–60.
6. **Gilbert C, Schaack S, Pace JK, II, Brindley PJ**, et al. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* **464**: 1347–50.
7. **Schaack S, Gilbert C, Feschotte C**. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol* **25**: 537–46.
8. **Eickbush TH, Malik HS**. 2002. Origins and evolution of retrotransposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, eds; *Mobile DNA II*. Washington, DC: ASM Press. p 1111–44.
9. **Bao W, Kapitonov VV, Jurka J**. 2010. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. *Mobile DNA* **1**: 3.
10. **Havecker ER, Gao X, Voytas DF**. 2004. The diversity of LTR retrotransposons. *Genome Biol* **5**: 225.
11. **Lerat E, Capy P**. 1999. Retrotransposons and retroviruses: analysis of the envelope gene. *Mol Biol Evol* **16**: 1198–207.
12. **Song SU, Gerasimova T, Kurkulos M, Boeke JD**, et al. 1994. An env-like protein encoded by a *Drosophila* retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev* **8**: 2046–57.
13. **Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG**, et al. 1990. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics* **124**: 339–55.
14. **Houck MA, Clark JB, Peterson KR, Kidwell MG**. 1991. Possible horizontal transfer of *Drosophila* genes by the mite *Proctolaelaps regalis*. *Science* **253**: 1125–8.
15. **Maruyama K, Hartl DL**. 1991. Evidence for interspecific transfer of the transposable element mariner between *Drosophila* and *Zaprionus*. *J Mol Evol* **33**: 514–24.
16. **Loreto EL, Valente VL, Zaha A, Silva JC**, et al. 2001. *Drosophila mediopunctata* P elements: a new example of horizontal transfer. *J Hered* **92**: 375–81.
17. **Pace JK, II, 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**: 17023–8.
18. **Gilbert C, Hernandez SS, Flores-Benabib J, Smith EN**, et al. 2012. Rampant horizontal transfer of SPIN transposons in squamate reptiles. *Mol Biol Evol* **29**: 503–15.
19. **Gilbert C, Waters P, Feschotte C, Schaack S**. 2013. Horizontal transfer of OC1 transposons in the Tasmanian devil. *BMC Genomics* **14**: 134.
20. **Walsh AM, Kortschak RD, Gardner MG, Bertozzi T**, et al. 2013. Widespread horizontal transfer of retrotransposons. *Proc Natl Acad Sci USA* **110**: 1012–6.
21. **Cowley M, Oakey RJ**. 2013. Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet* **9**: e1003234.
22. **Wicker T, Sabot F, Hua-Van A, Bennetzen JL**, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–82.
23. **Graham T, Boissinot S**. 2006. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *J Biomed Biotechnol* **2006**: 75327.
24. **Penzkofer T, Dandekar T, Zemojtel T**. 2005. L1Base: from functional annotation to prediction of active LINE-1 elements. *Nucleic Acids Res* **33**: D498–D500.
25. **Waters PD, Dobigny G, Waddell PJ, Robinson TJ**. 2007. Evolutionary history of LINE-1 in the major clades of placental mammals. *PLoS One* **2**: e158.
26. **Rangwala SH, Zhang L, Kazazian HH, Jr.** 2009. Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol* **10**: R100.
27. **Brouha B, Schustak J, Badge RM, Lutz-Prigge S**, et al. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* **100**: 5280–5.
28. **Kano H, Godoy I, Courtney C, Vetter MR**, et al. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev* **23**: 1303–12.
29. **Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW**, et al. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**: 1127–31.
30. **Baillie JK, Barnett MW, Upton KR, Gerhardt DJ**, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534–7.
31. **Hancks DC, Kazazian HH, Jr.** 2012. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* **22**: 191–203.
32. **Dewannieux M, Esnault C, Heidmann T**. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**: 41–8.
33. **Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KH**, et al. 2009. LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genet* **5**: e1000354.
34. **Mourier T, Willerslev E**. 2008. Does selection against transcriptional interference shape retroelement-free regions in mammalian genomes? *PLoS One* **3**: e3760.
35. **Cost GJ, Golding A, Schlissel MS, Boeke JD**. 2001. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res* **29**: 573–7.
36. **Casavant NC, Scott L, Cantrell MA, Wiggins LE**, et al. 2000. The end of the LINE?: lack of recent L1 activity in a group of South American rodents. *Genetics* **154**: 1809–17.
37. **Grahn RA, Rinehart TA, Cantrell MA, Wichman HA**. 2005. Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. *Cytogenet Genome Res* **110**: 407–15.
38. **Furano AV, Hayward BE, Chevret P, Catzeflis F**, et al. 1994. Amplification of the ancient murine Lx family of long interspersed repeated DNA occurred during the murine radiation. *J Mol Evol* **38**: 18–27.
39. **Boissinot S, Furano AV**. 2001. Adaptive evolution in LINE-1 retrotransposons. *Mol Biol Evol* **18**: 2186–94.
40. **Boissinot S, Entezam A, Young L, Munson PJ**, et al. 2004. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res* **14**: 1221–31.
41. **Casavant NC, Sherman AN, Wichman HA**. 1996. Two persistent LINE-1 lineages in *Peromyscus* have unequal rates of evolution. *Genetics* **142**: 1289–98.
42. **Khan H, Smit A, Boissinot S**. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* **16**: 78–87.
43. **Novick PA, Basta H, Floumanhaft M, McClure MA**, et al. 2009. The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis carolinensis* shows more similarity to fish than mammals. *Mol Biol Evol* **26**: 1811–22.
44. **Smit AF, Toth G, Riggs AD, Jurka J**. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246**: 401–17.
45. **Silva JC, Loreto EL, Clark JB**. 2004. Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol* **6**: 57–71.
46. **Vicient CM, Kalendar R, Schulman AH**. 2001. Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Res* **11**: 2041–9.
47. **Chessa B, Pereira F, Arnaud F, Amorim A**, et al. 2009. Revealing the history of sheep domestication using retrovirus integrations. *Science* **324**: 532–6.
48. **Zhang Y, Maksakova IA, Gagnier L, van de Lagemaat LN**, et al. 2008. Genome-wide assessments reveal extremely high levels of

polymorphism of two active families of mouse endogenous retroviral elements. *PLoS Genet* **4**: e1000007.
49. **Clark JB, Altheide TK, Schlosser MJ, Kidwell MG**. 1995. Molecular evolution of P transposable elements in the genus *Drosophila*. I. The *saltans* and *willistoni* species groups. *Mol Biol Evol* **12**: 902–13.
50. **Clark JB, Kidwell MG**. 1997. A phylogenetic perspective on P transposable element evolution in *Drosophila*. *Proc Natl Acad Sci USA* **94**: 11428–33.
51. **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**: 62–72.
52. **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**: 554–62.
53. **Hartl DL, Lozovskaya ER, Nurminsky DI, Lohe AR**. 1997. What restricts the activity of mariner-like transposable elements. *Trends Genet* **13**: 197–201.
54. **Loreto EL, Carareto CM, Capy P**. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. *Heredity* **100**: 545–54.
55. **Bartolome C, Bello X, Maside X**. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol* **10**: R22.
56. **Kawakami K**. 2007. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol* **8** (Suppl 1): S7.
57. **Gilbert C, Pace JK, Feschotte C**. 2009. Horizontal SPINning of transposons. *Commun Integr Biol* **2**: 117–9.
58. **Malik HS, Burke WD, Eickbush TH**. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**: 793–805.
59. **Jehle JA, Nickel A, Vlak JM, Backhaus H**. 1998. Horizontal escape of the novel Tc1-like lepidopteran transposon TCp3.2 into Cydia pomonella granulovirus. *J Mol Evol* **46**: 215–24.
60. **Malik HS, Henikoff S, Eickbush TH**. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**: 1307–18.
61. **Piskurek O, Okada N**. 2007. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. *Proc Natl Acad Sci USA* **104**: 12046–51.
62. **Thomas J, Schaack S, Pritham EJ**. 2010. Pervasive horizontal transfer of rolling-circle transposons among animals. *Genome Biol Evol* **2**: 656–64.
63. **Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A**, et al. 2012. Vertical evolution and horizontal transfer of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons in Lepidoptera species. *Mol Biol Evol* **29**: 3685–702.
64. **Takasaki N, Park L, Kaeriyama M, Gharrett AJ**, et al. 1996. Characterization of species-specifically amplified SINEs in three salmonid species–chum salmon, pink salmon, and kokanee: the local environment of the genome may be important for the generation of a dominant source gene at a newly retroposed locus. *J Mol Evol* **42**: 103–16.
65. **Kordis D, Gubensek F**. 1998. The Bov-B lines found in *Vipera ammodytes* toxic PLA2 genes are widespread in snake genomes. *Toxicon* **36**: 1585–90.
66. **Gentles AJ, Wakefield MJ, Kohany O, Gu W**, et al. 2007. Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Res* **17**: 992–1004.
67. **Kordis D**. 2009. Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenet Genome Res* **127**: 94–111.
68. **Kordis D, Gubensek F**. 1997. Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A2 genes and in genomes of Viperidae snakes. *Eur J Biochem* **246**: 772–9.
69. **Zupunski V, Gubensek F, Kordis D**. 2001. Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons. *Mol Biol Evol* **18**: 1849–63.
70. **Voltzit O**. 2007. A review of neotropical *Amblyomma* species (Acari: Ixodidae). *Acarina* **15**: 3–134.
71. **Voltzit OaK, JE**. 2002. A review of Asian *Amblyomma* species (Acari, Ixodida, Ixodidae). *Acarina* **10**: 95–136.
72. **Voltzit OaK, JE**. 2003. A review of African *Amblyomma* species (Acari, Ixodida, Ixodidae). *Acarina* **11**: 135–214.
73. **Guzinski J, Bull CM, Donnellan SC, Gardner MG**. 2009. Molecular genetic data provide support for a model of transmission dynamics in an Australian reptile tick, *Bothriocroton hydrosauri*. *Mol Ecol* **18**: 227–34.
74. **Novikova O, Sliwinska E, Fet V, Settele J**, et al. 2007. CR1 clade of non-LTR retrotransposons from *Maculinea* butterflies (Lepidoptera: Lycaenidae): evidence for recent horizontal transmission. *BMC Evol Biol* **7**: 93.
75. **de Almeida LM, Carareto CM**. 2005. Multiple events of horizontal transfer of the Minos transposable element between *Drosophila* species. *Mol Phylogenet Evol* **35**: 583–94.
76. **Friesen PD, Nissen MS**. 1990. Gene organization and transcription of TED, a lepidopteran retrotransposon integrated within the baculovirus genome. *Mol Cell Biol* **10**: 3067–77.
77. **Yoshiyama M, Tu Z, Kainoh Y, Honda H**, et al. 2001. Possible horizontal transfer of a transposable element from host to parasitoid. *Mol Biol Evol* **18**: 1952–8.
78. **Turnbull M, Webb B**. 2002. Perspectives on polydnavirus origins and evolution. *Adv Virus Res* **58**: 203–54.
79. **Brosius J, Gould SJ**. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci USA* **89**: 10706–10.
80. **Yang Z, Boffelli D, Boonmark N, Schwartz K**, et al. 1998. Apolipoprotein(a) gene enhancer resides within a LINE element. *J Biol Chem* **273**: 891–7.
81. **Iwashita S, Osada N, Itoh T, Sezaki M**, et al. 2003. A transposable element-mediated gene divergence that directly produces a novel type bovine Bcnt protein including the endonuclease domain of RTE-1. *Mol Biol Evol* **20**: 1556–63.
82. **Lowe CB, Haussler D**. 2012. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS One* **7**: e43128.
83. **de Souza FS, Franchini LF, Rubinstein M**. 2013. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol* **30**: 1239–51.

Chapter 2

Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life

"If you want to understand function, study structure."

— Francis Crick

L1 elements have been studied largely using model organisms. Due to limits on available genome data, the results from model organism studies have been applied to all eukaryotes, despite the huge diversity in the group. As such, there is this preconceived notion that L1 structure is tightly constrained and ubiquitous across eukaryotes. But L1s are ancient, they are present in both plants and animals, meaning that they have had millions of years to diverge. It is reasonable to expect variation due to accumulated changes over time. Before I could start assessing the possibility of horizontal transfer, I had to understand the constitution of L1s: the difference between young and ancient elements, domains or motifs that are linked to retrotranspositional capability, species-specific differentiation, and underlying characteristics which distinguish them from other retrotransposons.

Statement of Authorship

Title of Paper	LINEs between species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Atma M. Ivancevic, R. Daniel Kortschak, Terry Bertozzi, David L. Adelson (2016). LINEs between species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. Genome Biology and Evolution, Advance Access publication. DOI: 10.1093/gbe/evw243

Principal Author

Name of Principal Author (Candidate)	Atma M. Ivancevic		
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	6/12/16

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. Daniel Kortschak		
Contribution to the Paper	Supervised the development of work and assisted in analysing the results and writing the manuscript.		
Signature		Date	6/12/16

Name of Co-Author	Terry Bertozzi		
Contribution to the Paper	Supervised the development of work and assisted in analysing the results and writing the manuscript.		
Signature		Date	14.xii.16

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	7 Dec 2016

LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life

Atma M. Ivancevic[1], R. Daniel Kortschak[1], Terry Bertozzi[1,2], and David L. Adelson[1,*]

[1]School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia

[2]Evolutionary Biology Unit, South Australian Museum, Adelaide, South Australia, Australia

*Corresponding author: E-mail: david.adelson@adelaide.edu.au.

Accepted: September 26, 2016

Abstract

LINE-1 (L1) retrotransposons are dynamic elements. They have the potential to cause great genomic change because of their ability to 'jump' around the genome and amplify themselves, resulting in the duplication and rearrangement of regulatory DNA. Active L1, in particular, are often thought of as tightly constrained, homologous and ubiquitous elements with well-characterized domain organization. For the past 30 years, model organisms have been used to define L1s as 6–8 kb sequences containing a 5′-UTR, two open reading frames working harmoniously in *cis*, and a 3′-UTR with a polyA tail. In this study, we demonstrate the remarkable and overlooked diversity of L1s via a comprehensive phylogenetic analysis of elements from over 500 species from widely divergent branches of the tree of life. The rapid and recent growth of L1 elements in mammalian species is juxtaposed against the diverse lineages found in other metazoans and plants. In fact, some of these previously unexplored mammalian species (e.g. snub-nosed monkey, minke whale) exhibit L1 retrotranspositional 'hyperactivity' far surpassing that of human or mouse. In contrast, non-mammalian L1s have become so varied that the current classification system seems to inadequately capture their structural characteristics. Our findings illustrate how both long-term inherited evolutionary patterns and random bursts of activity in individual species can significantly alter genomes, highlighting the importance of L1 dynamics in eukaryotes.

Key words: transposable element; retrotransposon; LINE; eukaryotes; evolution.

Introduction

Transposable elements (TEs) are repetitive DNA sequences found in genomes scattered across the tree of life, and are often called 'jumping genes' because of their ability to replicate and move to new genomic locations. As such, they provide an important source of genome variation at both the species and individual level (Lynch 2006). Eukaryotic TEs are categorized based on their mechanism of retrotransposition. Class I retrotransposons use a copy-and-paste mechanism via an RNA intermediate, allowing massive amplification of copy number, which has the potential to cause substantial genomic change. Class II DNA transposons are more restricted because of their cut-and-paste mechanism. Retrotransposons are further divided into elements with (LTR) and without (non-LTR) long terminal repeats. Non-LTR elements comprise long interspersed elements (LINEs) and short interspersed elements (SINEs). LINEs are autonomous because they encode their own proteins for retrotransposition, whereas SINEs are non-autonomous and depend (in *trans*) on LINE-expressed proteins.

Long interspersed element 1 (LINE-1 or L1) is a well-known group of non-LTR retrotransposons found primarily in mammals (Kazazian 2000). Given their presence in both plant and animal species, L1s are very ancient elements; and it is assumed that they are ubiquitous across eukaryotes. More importantly, they are one of the most active autonomous elements in mammals, covering as much as 18% of the human genome (Furano 2000; Lander et al. 2001) and accountable for about 30% through amplification of processed pseudogenes and *Alu* SINEs (Esnault et al. 2000; Dewannieux et al. 2003; Graham and Boissinot 2006). This means that L1s are major drivers of evolution, capable of wreaking havoc on the genome through gene disruption (Kazazian 1998), alternative splicing (Kondo-Iida et al. 1999) and overexpression leading to cancer development and progression (Chen et al. 2005; Kaer and Speek 2013).

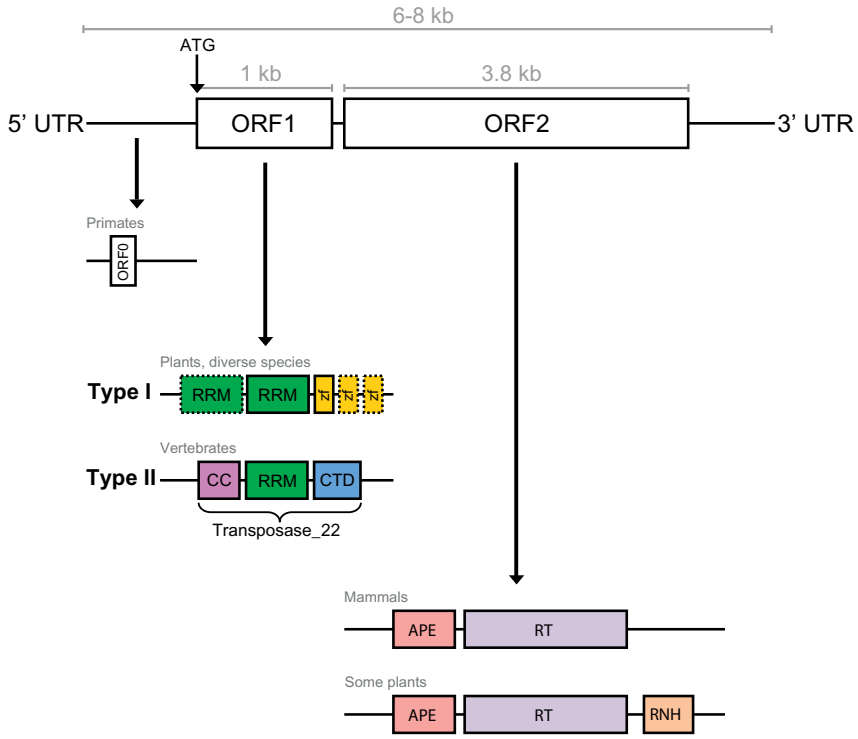


Fig. 1.—Conventional L1 structure and known variants. A functional L1 retrotransposon is 6–8 kb in length and contains two ORFs, both of which encode proteins for retrotransposition. ORF0 has recently been discovered in primates and is thought to facilitate retrotransposition. L1 ORF1 sequences are divided into two types: Type II is widespread throughout vertebrates, while Type I has only been found in diverse plants and non-mammalian animals such as amphibians and fish. Likewise, domain variants of ORF2 with an additional ribonuclease domain have been found in some plant species (described in the main text). UTR, untranslated region; ORF, open reading frame; RRM, RNA recognition motif; zf, gag-like $Cys_2HisCys$ zinc knuckle; CC, coiled-coil; CTD, C-terminal domain; APE, apurinic endonuclease; RT, reverse transcriptase; RNH, ribonuclease H domain.

In the literature, active L1s are defined as 6–8 kb elements 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 polyA tail (Furano 2000) (see fig. 1). ORF2 is around 3.8 kb in length, translating to a 150-kDa protein (ORF2p) which encodes an apurinic endonuclease and reverse transcriptase (RT) necessary for retrotransposition. ORF1 is much smaller (1 kb nucleotide sequence; ORF1p is only 40 kDa) and thought to have RNA-binding functionality (Furano 2000; Cost et al. 2002). This widely accepted structure has been used for over 30 years to identify putatively active elements in mammalian genomes (Scott et al. 1987). More recently, however, L1s with significant structural variations have been discovered – to the extent that the current terminology on what constitutes an L1 seems inadequate and limiting.

For example, some plant species have been shown to contain an additional ribonuclease H domain (RNH) in ORF2p downstream of the RT domain, possibly acquired from domain shuffling between plants, bacteria, and Archaea (Smyshlyaev et al. 2013). The domains located within ORF1p can also vary drastically. Khazina and Weichenrieder (2009) classified retrotransposon ORF1 proteins into five types based on the presence and grouping of different domains, and indicated in which species/transposons each type was most commonly found. Type I ORF1p contains at least one RNA recognition motif (RRM) with a $Cys_2HisCys$ (CCHC) zinc knuckle, and is found in some plant L1s. Type II is the typical mammalian L1 ORF1p 'Transposase 22' (Finn et al. 2010), consisting of a coiled-coil (CC), single RRM and C-terminal domain. Type III and IV ORF1s are supposedly restricted to archaic elements such as CR1s (Chicken repeat 1) (Kapitonov and Jurka 2003) and L2s (Nakamura et al. 2012) and Type V are unclassified. However, even these classifications are insufficient. Metcalfe and Casane (2014) found that Jockey superfamily elements (especially CR1s and L2s) contain every possible type described by Khazina and Weichenrieder (2009), as well as further subtypes. This raises the question of whether L1s are also diverse in their structure, rather than being confined to Type II or I.

Some L1s do not appear to have an ORF1 region (Odon et al. 2013). For a long time, it was thought that co-expression of both ORF1p and ORF2p in *cis* was necessary for retrotransposition (Moran et al. 1996). However, L1 copies containing a disrupted ORF1p but intact ORF2p retain the ability to mobilise SINEs within the genome, as shown by Dewannieux et al. (2003) with a defective ORF1p mutant. Perhaps most intriguingly of all, recent evidence suggests the possibility of a third ORF in L1 elements: ORF0, an antisense open reading frame upstream of ORF1 (Denli et al. 2015). This ORF0 is very short, encoding a 71 amino acid peptide, and is thought to be primate-specific. Overexpression of ORF0p leads to a significant increase in L1 mobility, which may help explain the high retrotransposition activity of L1 in some primates (e.g. humans).

Growing evidence (Kordis et al. 2006; Waters et al. 2007; Blass et al. 2012; Tollis and Boissinot 2013; Heitkam et al. 2014) suggests that the current model of L1 activity is insufficient. The idea that ORF1p + ORF2p in *cis* = retrotransposition fails to capture variation between different organisms, particularly beyond the mammalian lineage. In this study, we provide a definitive and comprehensive phylogenetic analysis of L1 content and activity in over 500 species from widely divergent branches of the tree of life. The genomes selected include plants, arthropods, sauropsids, mammals, and other, more primitive eukaryotic species. We also include several cases of closely related organisms (within the same genus or species) to look for L1 differences between individuals, and the effects of different genome assembly methods. For each genome, we searched for the presence of L1 elements; and if found, characterized the elements as active or inactive and identified the domains in each of the ORF proteins. Our findings effectively illustrate the overlap between inherited evolutionary patterns and random individual bursts of activity, allowing a much broader understanding of TE dynamics in eukaryotes.

Materials and Methods

Extraction and Characterization of L1 Repeats from Taxa with Full Genome Data

Almost all of the genomes used in this study (499 out of 503) are publicly available from the National Center for Biotechnology Information (NCBI) (Sayers et al. 2012) or UCSC Genome Browser (Kent et al. 2002). Supplementary table S1, Supplementary Material online lists the systematic name, common name, version, source and submitter of each genome assembly, and marks which genomes were privately acquired. If there was both a GenBank and RefSeq version for the genome, the GenBank version was used by default. Supplementary table S2, Supplementary Material online shows the total genome sequence length and scaffold/contig N50 values, giving an approximation of the assembly quality. Supplementary table S3, Supplementary Material online compares the different sequencing technologies and methods. A phylogenetic representation of the genomic dataset was inferred using Archaeopteryx (Zmasek 2015) to download the Tree of Life (Maddison and Schulz 2007) topology for all Eukaryota (node identifier 3, ~76,000 species). The tree was extended (e.g. descendants added where necessary) to include all of the 503 genomes, and species not included in this study were removed. Out-dated branches were changed using OrthoDB (Kriventseva et al. 2015), OrthoMaM (Douzery et al. 2014), NCBI Taxonomy (Sayers et al. 2012) and recent publications (Murphy et al. 2001; Beck et al. 2006; Janecka et al. 2007) as references (see supplementary fig. S1, Supplementary Material online).

L1 hits were initially identified in each genome using an iterative query-driven method based on sequence similarity, as seen in Walsh et al. (2013). The original query L1 sequences were obtained from Repbase (Jurka et al. 2005) by searching for anything listed as 'L1' or 'Tx1' (subgroup of the L1 clade) for all taxa. Cow and horse L1s were also obtained from past analyses (Adelson et al. 2009, 2010). All of the accumulated query sequences were concatenated into one file, which was used as the input query to run LASTZ v1.02.00 (Harris 2007) with at least 80% length coverage. BEDTools v2.17.0 (Quinlan and Hall 2010) was used to merge overlapping hit intervals from different queries and extract a non-redundant set of L1 sequences in FASTA format. For each genome, the output hits were globally aligned with MUSCLE v3.8.31 (Edgar 2004) to produce a species consensus with Geneious v7.0.6 (Kearse et al. 2012). Genomes with a substantial number of hits required clustering with UCLUST v7.0.959_i86linux32 (Edgar 2010) before aligning. The species consensus sequences were then added to the query file (see supplementary fig. S2, Supplementary Material online). This process was repeated three times, to accommodate inclusion of new genomes at various stages in the pipeline and to include diverse L1s to the set of queries.

To control for difference in genome assembly quality, we also used the TBLASTN program (Altschul et al. 1990) to search the non-redundant NCBI nucleotide database (NR) and high throughput genomic sequences (HTGS) (Sayers et al. 2012). TBLASTN search parameters were default except the e-value was changed to $1e^{-5}$. Input was the concatenated ORF1p and ORF2p from 13 full-length L1-clade elements from Repbase (Jurka et al. 2005), spanning each order/clade (where available), and consisting of mammalian L1/diverse L1/diverse Tx1 elements (see supplementary table S4, Supplementary Material online for exact queries and TBLASTN results). To determine the reliability of low-scoring hits, each hit was extracted as a nucleotide sequence and screened with CENSOR (Kohany et al. 2006) against the entire Repbase library of known repeats. This provided a 'reciprocal best-hit' check: Hits were kept if the best hit from CENSOR was an L1, and discarded if the best hit was another repetitive sequence (e.g. retrotransposons BovB or CR1).

Confirmed L1 sequences from the TBLASTN approach were used as species-specific queries to re-run LASTZ on each genome. Then, the sequences from each species were concatenated into a final query file (>3 million L1s, both fragment and full-length copies) for the last round of LASTZ extraction. The Repbase library (with CENSOR) was again used to verify L1s with a reciprocal best-hit check. Supplementary table S5, Supplementary Material online shows the results from the final LASTZ extraction, with notes comparing the number of L1s found to previous studies. Sample code for each step is available online (https://github.com/AdelaideBioinfo/L1-dynamics).

Both the LASTZ and TBLASTN approaches are limited by the quality and quantity of available nucleotide data whether it is from the genome assembly or nucleotide databases (NR/HTGS). As such, the L1 status of each species (e.g. L1 presence versus absence) was determined based on the union of the two methods (see Supplementary table S7, Supplementary Material online).

Identification of Intact Open Reading Frames

BEDTools (Quinlan and Hall 2010) was used to extend each L1 hit by 1kb either side before the ORF analysis, to overcome incomplete 5′ and 3′ ends that may be missing crucial start/stop codons. Geneious (Kearse et al. 2012) was then used to scan for open reading frames that were at least 80% of the expected length (≥ 800 bp for ORF1 and ≥3 kb for ORF2 – see supplementary fig. S4, Supplementary Material online). ORF sequences which satisfied the length requirements were subjected to a series of tests to confirm their functionality: Each ORF had to be complete with a start codon, stop codon and no debilitating mutations in between (such as premature stop codons or too many ambiguous nucleotides). For ORF1, the start codon had to be a methionine (ATG) (Penzkofer et al. 2005) and ORF2p sequences had to have a confirmed RT domain. After translation, both ORF1p and ORF2p candidates were checked for similarity to known domains using HMM–HMM comparison (Finn et al. 2011) against the Pfam 28.0 database (Finn et al. 2010) as at May 2015 (includes 16,230 families).

ORF1p sequences were initially screened for known L1 ORF1p domains (e.g. Transposase_22, RRM, zf-CCHC). Sequences containing at least one of these domains were kept as 'confirmed' ORF1p. Confirmed ORF1p sequences often contained other, associated domains: 'probable' ORF1p domains, such as DUF4283 in plants. A library was generated containing probable ORF1p-associated domains and used to re-screen the unconfirmed ORF1p candidates. Matching sequences were categorized as 'probable ORF1p' (see supplementary fig. S7, Supplementary Material online). This resulted in three categories of L1 ORF proteins: Confirmed ORF2p, confirmed ORF1p, and probable ORF1p. Nucleotide L1 sequences were given label prefixes according to their ORF composition: ORF1_ (confirmed ORF1p), ORF2_ (confirmed ORF2p), probORF1_ (probable ORF1p), ORF1_ORF2_ (both ORF proteins confirmed), or probORF1_ORF2_ (confirmed ORF2p, probable ORF1p). Supplementary table S6, Supplementary Material online summarizes the ORF content in each genome. Only ORF sequences that passed all the tests were included in subsequent analyses.

Classification of Potentially Active L1 Elements

An L1 was defined as a potentially active candidate if it contained an intact ORF2 (regardless of the state of ORF1), as this means that it is either fully capable of retrotransposing itself (Moran et al. 1996; Heras et al. 2006) or it can cause activity in the genome by mobilizing SINEs (Dewannieux et al. 2003). The ORF2 sequence had to satisfy the criteria listed above (≥ 3kb nucleotide sequence, complete with start and stop codons and no inactivating mutations, and confirmed RT domain). L1 elements containing intact ORF2, and thus potentially active, were typically full-length or near full-length (e.g. >4.5 kb). Genomes with low copy number were further checked for contamination: For example, the potentially active L1s were not considered valid if they came from short, isolated scaffolds or showed suspiciously high similarity to another (divergent) species.

Dendrogram Construction from Nucleotide L1 Sequences

Full-length L1 sequences (or near full-length, as long as they included an intact ORF2) were globally aligned using MUSCLE (Edgar 2004). Mammalian species required iterative clustering with UCLUST (Edgar 2010) before aligning, due to the huge number of hits. Clustering identities ranged from 70 to 95%. Alignments were trimmed with Gblocks (Castresana 2000) to remove large gaps (default parameters, allowed gap positions: with half). The dominant active clusters for each species were represented as dendrograms, or unrooted tree diagrams, using FastTree v2.1.8, double-precision version (i.e. compiled with –DUSE_DOUBLE) (Price et al. 2010). Archaeopteryx v0.9901 beta (Zmasek 2015) was used to visualise and annotate each tree based on the ORF labels.

Phylogenetic Analysis of Conserved L1 Amino Acid Residues

Two methods were tested to depict the evolutionary dynamics of potentially active L1 elements. First, we inferred an ORF2p consensus tree: All confirmed ORF2 sequences in each species were extracted, translated and globally aligned with MUSCLE (Edgar 2004). The consensus for each species was generated in Geneious (Kearse et al. 2012) using majority rule (most common bases, fewest ambiguities) and a base was regarded ambiguous if coverage at that position was < 3 sequences (unless the alignment had ≤ 3 sequences, in which case this was changed to < 2 sequences). This produced a single L1 ORF2p consensus for each species. These consensus

sequences were globally aligned using MUSCLE (Edgar 2004) and a phylogeny was inferred with maximum likelihood using FastTree, double precision compilation (Price et al. 2010).

Another phylogeny was inferred using just the RT domains within ORF2p. For each confirmed ORF2p sequence, the RT domain was extracted using the envelope coordinates from the HMMer domain hits table (–domtblout) (Finn et al. 2011), with minimum length 200 amino acid residues. RT domains from all species were collated into one file (37,994 sequences total), which was then clustered with USEARCH (Edgar 2010) at 90% identity. Each cluster was defined as a L1 RT-family (3508 families total). Only RT-families containing more than five members were included in the phylogenetic analysis. Two RT domains from Repbase (Jurka et al. 2005) were also included: A CR1 element from *Anopheles gambiae* (Ag-CR1-22), to act as the outgroup, and Zepp from *Chlorella vulgaris*, as a sister element to the L1s found in *Coccomyxa subellipsoidea*. As before, alignments were performed using MUSCLE, Geneious was used to extract a consensus for each family, and FastTree was used to infer a maximum likelihood phylogeny. A second tree was built using the neighbor-joining method and tested with bootstrapping (1,000 replicates).

Clustering Analysis of L1 ORF1 Proteins

A reliable phylogeny could not be inferred from ORF1p sequences because of the high variation in non-mammalian species. Instead, ORF1p sequences were clustered using an all-against-all BLAST (Altschul et al. 1990) approach. The BLAST was performed using BLAST v2.2.24 and NCBI-BLAST v2.2.27+ (Altschul et al. 1990) with the following parameters: -p blastp, -e 1e−10, -m 8 (for tabular output). Based on the BLAST results, the ORFs were then clustered using SiLiX software (Miele et al. 2011) with default parameters and –net to create a net file which contains all the pairs taken into account after filtering.

Results

Ubiquity of L1 across Plants and Animals

To simplify discussion of the results, we define three different states that a genome can be in, in terms of L1 content: Absent ($L1^-$), meaning that no L1s were detected in the genome; present ($L1^+$), meaning that L1s were found in partial or full-length form; and potentially active (L1*), meaning that at least one putatively active L1 was found in the genome (using either the TBLASTN or LASTZ method). $L1^-$ and $L1^+$ are mutually exclusive (a genome cannot have both presence and absence of L1s), whereas L1* is the potentially active subset of $L1^+$. Using this ternary system, we screened 503 eukaryotic species representing key clades of the tree of life (125 plants, 145 protostomes, 98 mammals, 74 sauropsids, 22 neopterygians, 11 flatworms, and 28 other species) (fig. 2; see supplementary fig. S1, Supplementary Material online). Of these, 407 species were found to be $L1^+$. L1 copy number was highest in mammals, with thousands of full-length L1 sequences found in almost every mammalian species analysed (with the exception of monotremes, which are $L1^-$).

L1s also appeared frequently in plants (118/125 $L1^+$ plant species), but colonized far less of each genome (e.g. typical copy number between 10 and 1,000 L1s). Fish, non-avian reptiles and amphibians showed consistent presence but similarly low copy numbers compared with mammals. Birds had an exceptionally low (yet consistent) L1 copy number: Only one full-length L1 element was found in most of the bird species analysed (and multiple fragments), yet this element was conserved through enough species that it is likely an ancient remnant of L1 from a common ancestor.

In the protostomes, L1 presence was verified in all mosquito and fly species, but appeared sporadically elsewhere. Fragments were found in all *Schistosoma* flatworms, as well as *Clonorchis sinensis*. The remaining 'primitive' orders contained multiple full-length L1 families, with the exception of Tentaculata (*Mnemiopsis leidyi*), Placozoa (*Trichoplax adhaerens*), and Porifera (*Amphimedon queenslandica*). Supplementary table S5, Supplementary Material online contains a summary of the L1 sequences found in each genome and the length distribution of the hits.

Dead or Alive – How Many L1s Have Retained Their Activity?

Of the 407 $L1^+$ eukaryotes, 206 species were further determined to be L1*: 92 plants, 67 mammals, and 47 non-mammalian animal species. This is illustrated in fig. 2 (full tree, no node labels – see supplementary fig. S5, Supplementary Material online), fig. 3 (mammals) and fig. 4 (plants). Although all coloured branches indicate presence ($L1^+$), the potentially active subset (L1*) is coloured magenta, so in this case the blue branches ($L1^+$—L1*) indicate species that only contain 'extinct' L1s (i.e. present but inactive). Because the L1 state of each genome is only observable at the tree tips, the phylogeny was annotated based on the notion that the most parsimonious explanation is a loss of activity, not a gain (hence ancestral branches are coloured 'active' if any of the descendants display activity). Noticeably, despite the ubiquitous presence of L1 across the mammalian lineage, L1 in quite a few mammalian species or subgroups (e.g. megabats, some rodents, and Afrotherian mammals) appear extinct. In contrast, other mammals seem to be bursting with L1 activity: Including several species (e.g. minke whale, antelope, snub-nosed monkey, panda, baiji) which have not been studied before in the context of L1 retrotransposition.

Previously, the human genome has been used as a model for high retrotranspositional activity. Numerous studies have found that L1 retrotransposition rates differ substantially between primate lineages, for example, human versus chimpanzee (Gregory et al. 2002; Mathews et al. 2003; Lee et al.

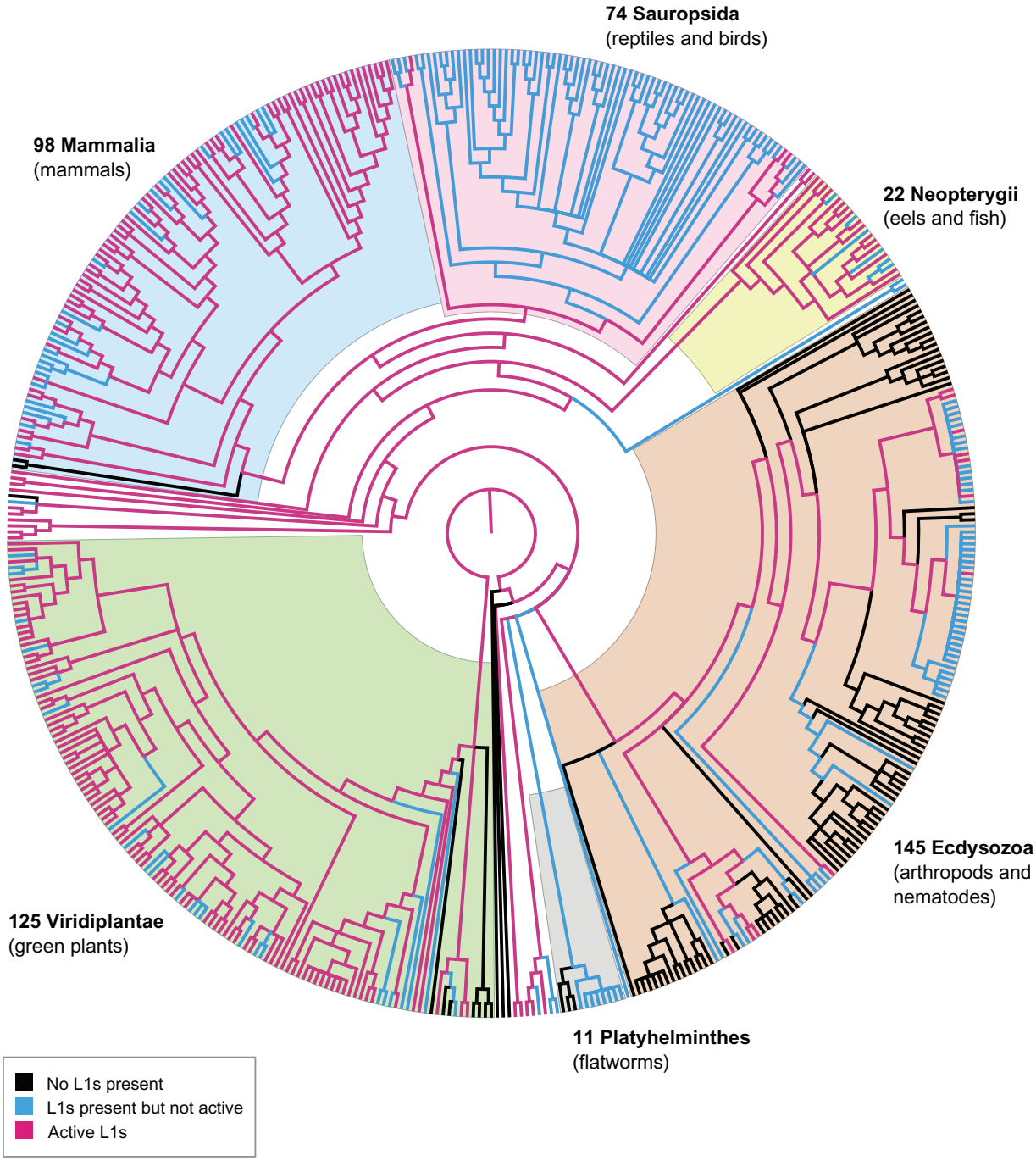


Fig. 2.—Phylogenetic representation of genomic dataset. Species relationships between the 503 representative genomes used in this study were depicted using Archaeopteryx to download the Tree of Life topology for all Eukaryota (node id 3) and extract the 503 species of interest. Out-dated branches were updated using OrthoDB, OrthoMaM, NCBI Taxonomy and recent publications as references. Labels indicate the major groups present in this dataset. Branches are colored to indicate the L1 state of each genome, as shown in the legend.

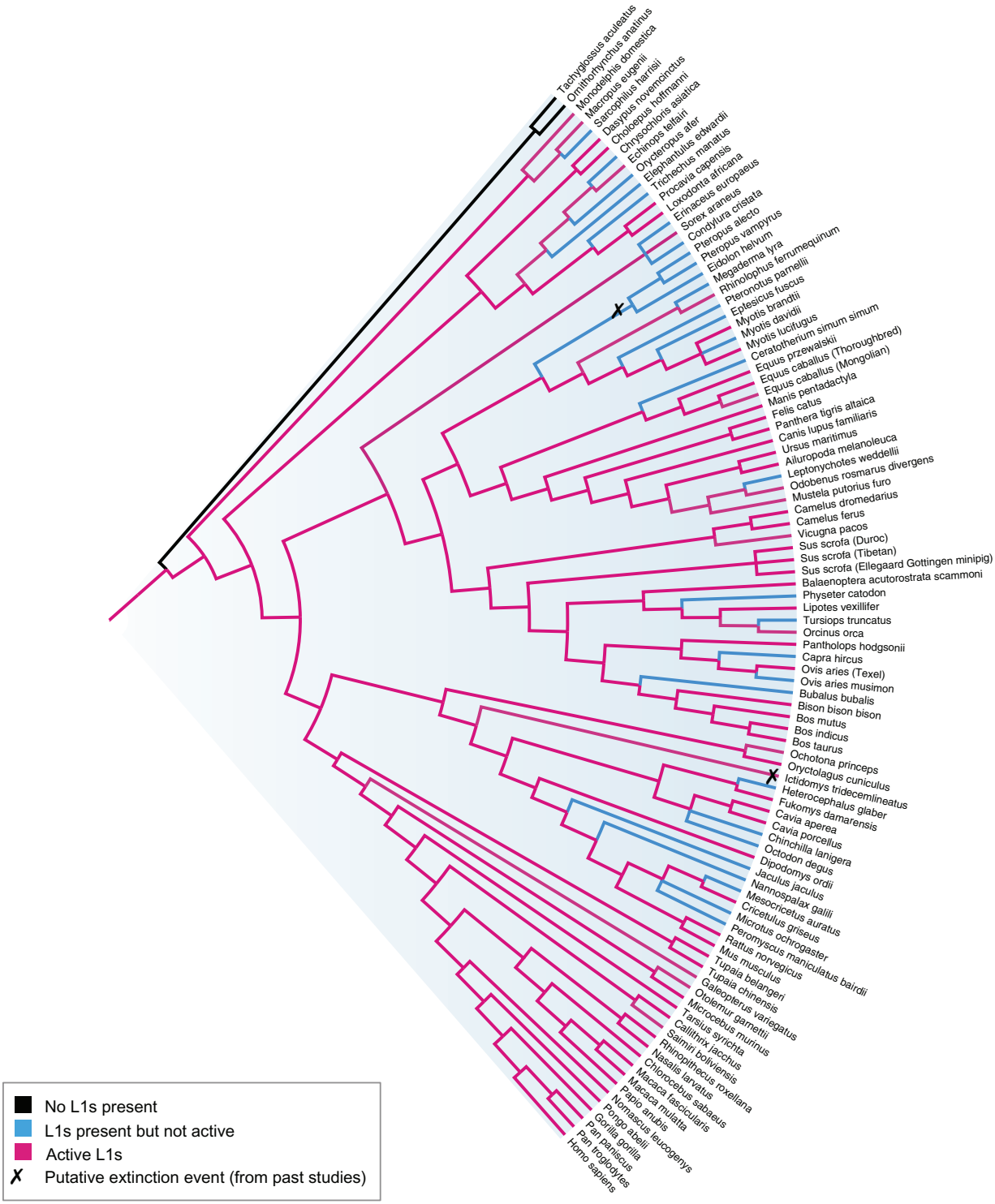


Fig. 3.—Mammalian phylogeny reveals ubiquitous L1 presence (except for monotremes) and possible extinction events. Genomes are classified as L1 absent ($L1^{-}$) (black), L1 present but inactive ($L1^{+}$–L1*) (blue) or L1 active (L1*) (red). Putative extinction events from past studies are marked.

Fig. 4.—Plant phylogeny showing the sporadic distribution of active L1 and the L1 state of each genome (colored branches). Brassicales and Poales stand out as the dominant L1* families. Orders containing more than three representative genomes are named.

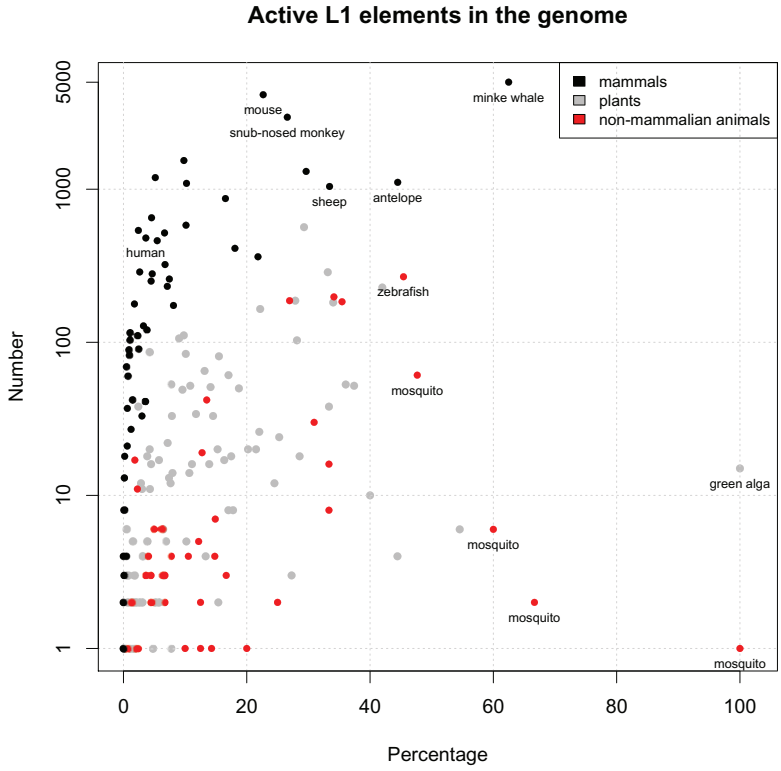


Fig. 5.—Distribution of active L1 elements reveals several 'hyperactive' mammalian species. The *y*-axis shows the number of active L1 in the genome; the *x*-axis shows the percentage of active L1s in the genome (i.e. # active L1/# near full-length L1 × 100, as described in supplementary table S8, Supplementary Material online). Non-mammalian animal species (red) and plants (gray) appear to have high retrotranspositional potential but low observable L1 activity in the genome. In contrast, mammals (black) typically have a very high L1 copy number, but the majority of these are inactive. The labelled mammalian species stand out as L1 'hyperactive' species because they are the most likely to be currently replicating and expanding within the genome.

2007). That is particularly evident with this new comparison of human versus snub-nosed monkey. For example: In the human genome, we identified 266 potentially active, both-ORF-intact L1s, and other studies have quoted similar numbers [e.g. Penzkofer et al. (2005) estimate ~150 on L1 Base]. Of such L1* candidates, <50% are active in cell culture: Brouha et al. (2003) predict that there are only about 80–100 active L1s in the average human, although this varies between individuals (Seleme et al. 2006; Beck et al. 2010). The snub-nosed monkey genome, on the other hand, contains 2549 both-ORF-intact L1* candidates. More than 95% of these would have to be determined inactive upon experimental analysis to obtain a comparable number to human; so the retrotransposition potential of snub-nosed monkey is substantially higher than that of human or any other primate.

L1 activity persists beyond the mammalian lineage as well. Almost every order that exhibits L1 presence contains L1* species (the two exceptions being Platyhelminthes and Chondrichthyes, where the presence is solely due to L1 fragments). Birds similarly contain L1 fragments or low copy number full-length elements, yet the ORF2 region is heavily degraded and mutated.

In plants, the L1 state of species seems to mirror mammalian genomes. Brassicales and Poales stand out as the most dominant orders, with each member bearing a significant number of active L1s. Another notable L1* species is *Coccomyxa subellipsoidea*, which only contains 15 L1 elements but every single one of these elements is putatively active and almost identical, suggesting recent retrotransposition. This genome also appears as a discrepancy in our tree; it is one of the only instances where a L1* species is phylogenetically placed next to a $L1^{-}$ species (fig. 4). However, given that our dataset does not contain all species, this could be a result of incomplete sampling and hence incorrect placement of the species. The ancestral branch was coloured red (L1*) despite the absence of L1s in several descendent species, because another study shows that *Chlorella vulgaris* (sister to *Chlorella variabilis*, which is marked $L1^{+}$) contains active L1-like Zepp elements 98% identical to *Coccomyxa subellipsoidea* (Higashiyama et al. 1997).

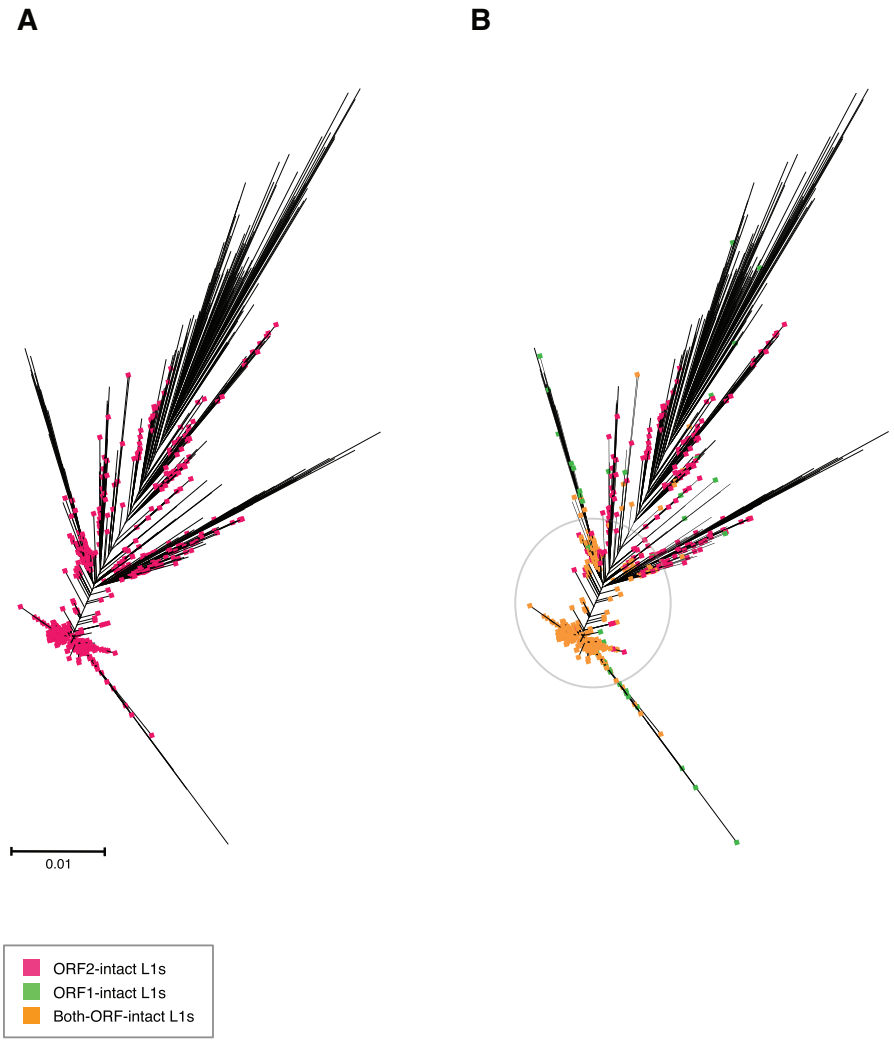


Fig. 6.—Master lineage model predominant in most mammalian species, including snub-nosed monkey *Rhinopithecus roxellana*. (*a*) Maximum likelihood dendrogram inferred using FastTree double precision version, from full-length L1 nucleotide sequences extracted from genomic data. Sequences were clustered with UCLUST and globally aligned with MUSCLE. Species with a clearly dominant L1* cluster were classified as master lineage models, as shown in Supplementary table 9. Sequences in the alignment were tagged to indicate which ORFs were intact and visualized using Archaeopteryx. This figure highlights the ORF2-intact L1s. (*b*) Same as (*a*), but here the highlighting also shows ORF1-intact L1s and both-ORF-intact L1s. Both-ORF-intact L1s are tightly clustered on the short branches in the middle.

Finally, the number of potentially active L1s found in each genome was compared with the total number of near full-length L1s in that genome, to get a percentage estimate of L1 activity per species (fig. 5; see supplementary table S8, Supplementary Material online). We found that mammalian species often contain a large number of inactive elements, so the percentage of active L1s is relatively low (e.g. <20%). In contrast, non-mammalian species (animals and plants) seem to have a higher proportion of active L1s in the genome despite the lower copy number; so the centroid of the graph is shifted to the right.

Mammalian Species Typically Have a Dominant Active Cluster

The clustering and dendrogram construction of L1 nucleotide sequences revealed that most mammals contain one large, dominant active cluster of closely related elements. As mentioned before, snub-nosed monkey is a remarkably active species in a comparatively inactive subgroup (i.e. primates). The cluster depicted in figure 6 contains 1742 full-length L1 (1337 both-ORF-intact and another 195 ORF2-intact) with 95.2% pairwise identity, which was used to construct an unrooted

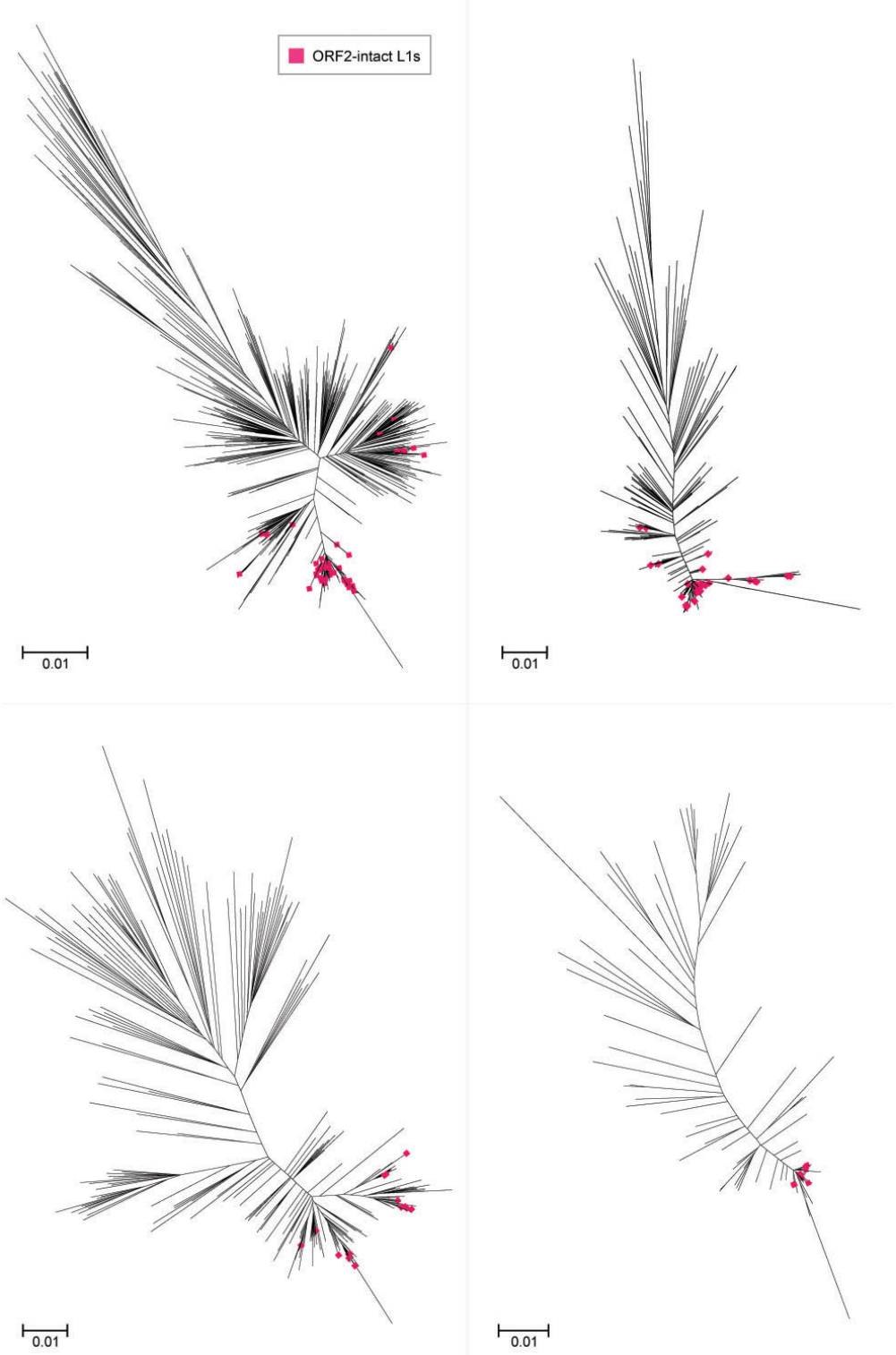


Fig. 7.—Multiple L1 lineages present in the *Myotis lucifugus* genome. Maximum likelihood dendrogram inferred using FastTree from full-length L1 nucleotide sequences extracted from full genome species data. As in Fig. 6, sequences were clustered with UCLUST, aligned with MUSCLE, annotated with Geneious and visualized with Archaeopteryx. Only ORF2-intact L1s are highlighted.

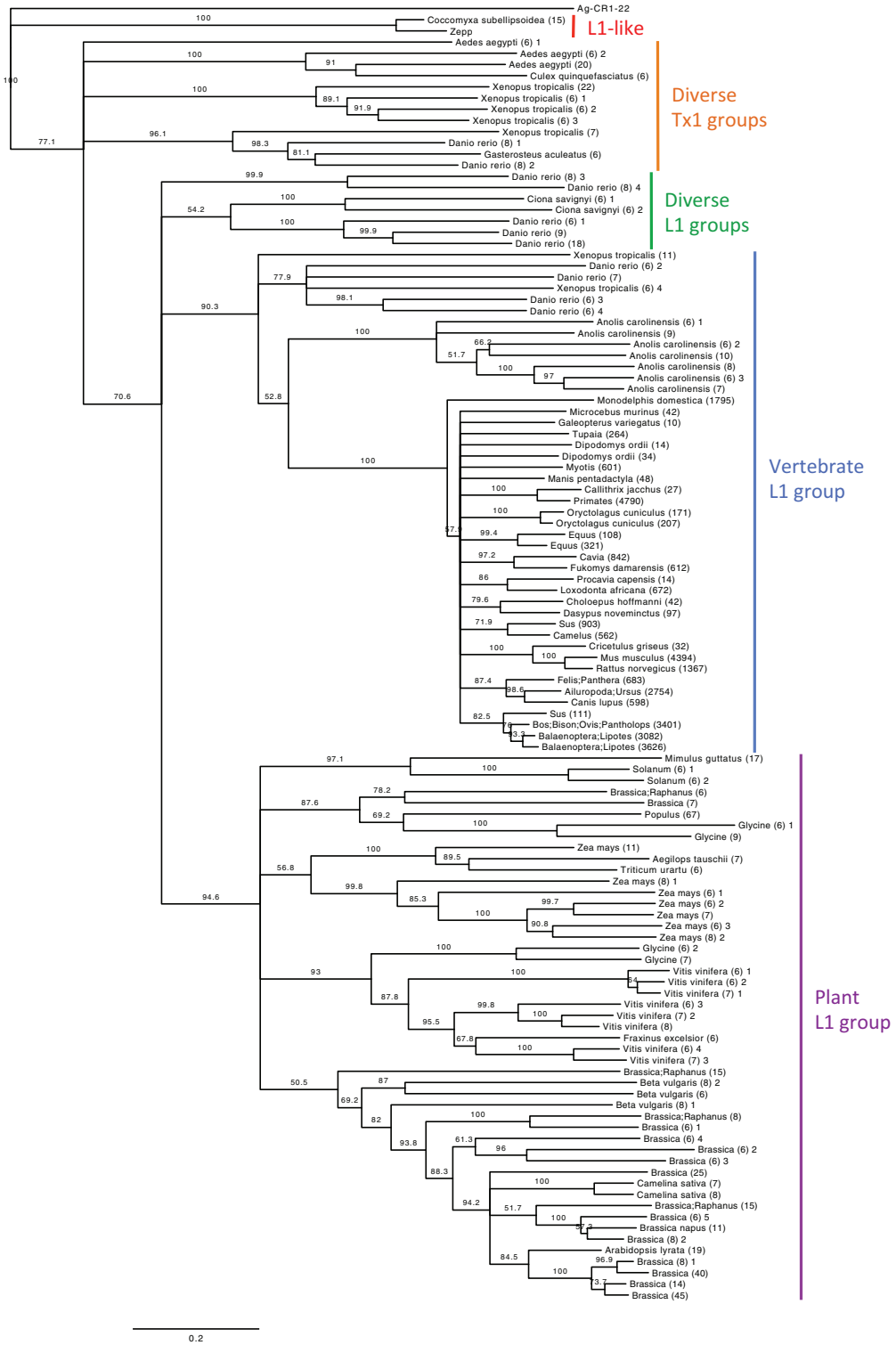


Fig. 8.—Phylogenetic analysis of RT families shows the overall hierarchy of L1/Tx1 groups. Rooted Neighbor-Joining tree based on amino acid RT domains. This tree represents the bootstrap consensus after 1,000 replicates, with nodes that have confidence values over 50% labelled. CR1 from *Anopheles gambiae* (outgroup) and Zepp from *Chlorella vulgaris* (98% identical to *Coccomyxa subellipsoidea* L1s) were obtained from Repbase. Only

maximum likelihood tree highlighting elements with ORF1 intact, ORF2 intact, or both ORFs intact. Almost all of the L1s in this cluster have both ORFs intact and are clustered on the shorter branches, indicating very recent activity.

However, in some species it is obvious that there is more than one significant active cluster. Horse (*Equus caballus*) is a well-known example of a species with five L1 (equine) subfamilies, two of which contain active elements (Adelson et al. 2010). Megabats are also known to have harboured multiple contemporaneous L1 lineages, although those lineages are now extinct (Yang et al. 2014). Nonetheless, this multiple lineage phenomenon seems to extend to the microbat subgroup as well: figure 7 depicts the clustering and dendrogram construction for *Myotis lucifugus*, where there is no discernible dominant cluster. The elements in each cluster are >70% similar to each other, but the clusters themselves are distinct at this level (see supplementary table S9, Supplementary Material online). Once again, we see a tendency for active L1s to converge on the short branches.

RT Domain Reveals Distinct L1 Groups

The phylogenetic analysis of RT families (fig. 8) clearly illustrated differences between L1 groups. Two L1 clades are immediately obvious: Vertebrate L1s, with the shortest observed branch lengths, and plant L1s, displaying significantly longer branches and lower support values. The rest of the phylogeny is made up of diverse L1 and Tx1 groups from combinations of fish, amphibians, mosquitos, sea squirts, and green algae.

Mammalian species form a hard polytomy, vaguely reflecting expected species relationships but without accurate subclass structure. This is most likely due to the sporadic sampling of species (based on data availability). In addition, the mammalian RT-families all have a large number of shared amino acids, making it difficult to reliably distinguish subfamilies. This is especially true for primates, which all grouped together as a single RT-family (4790 members with >90% identity) except for the strepsirrhine primate *Microcebus murinus*. The striking lack of diversity supports the idea of a rapid L1 explosion in the mammalian lineage following a severe population bottleneck (Kordis et al. 2006).

In contrast, non-mammalian animals contain multiple distinct L1 lineages and are not restricted to a single group or clade. This phenomenon has been explored in depth for fish (Duvernell et al. 2004; Furano et al. 2004; Blass et al. 2012), Anole lizard (Novick et al. 2009; Tollis and Boissinot 2013), *Xenopus* frogs (Kojima and Fujiwara 2004; Kordis et al. 2006) and African mosquitos (Biedler and Tu 2003). Fish and amphibians are the only known species to contain both mammalian-like vertebrate L1s, and diverse L1/Tx1 families (representatives *Danio rerio* and *Xenopus tropicalis* shown in fig. 8). Note that figure 8 only shows RT families within confirmed ORF2p, ≥200 amino acids in length, and containing >5 members at 90% identity, to reduce the dataset to a manageable number for visualization.

The plant L1 group (excluding *Coccomyxa subellipsoidea*) is divided into five subclades: The largest of which is made up of Brassicales species plus *Beta vulgaris* (Caryophyllales) (fig. 8). Brassicales is one of the most L1-active orders (fig. 4; see supplementary table S5, Supplementary Material online) and contains multiple L1 lineages. This is evident by the ORF2p analysis: Excluding *Carica papaya* ($L1^{-}$), all *Brassicales* species contain both the typical RT (RVT_1), as well as diverse RT and ribonuclease combinations (e.g. RVT_1 + RVT_3/RNH, see supplementary table S10, Supplementary Material online). The ORF1p analysis similarly revealed novel L1 lineages within Brassicales species *Camelina sativa*, *Aethionema arabicum*, and *Arabis alpina*, characterized by the presence of N-terminal RRMs (see supplementary table S11, Supplementary Material online). *Beta vulgaris* contains these same RRM-ORF1p, known as the BNR lineage (Heitkam and Schmidt 2009) – which is probably why *Beta vulgaris* is the only non-Brassicales species to appear in this L1 subgroup (fig. 8). Heitkam et al. (2014) suggested that the RRM domain substitutes the RNA-binding function of the zinc finger. A number of other plant species were found to include RRM-ORF1p (see supplementary table S11, Supplementary Material online), supporting the idea that L1s can recruit functional domains from their host to contribute to retrotransposition (Heitkam et al. 2014).

Variation of ORF1 Proteins across Species

The variability found in ORF1 sequences, from both plants and animals, is staggering. Khazina and Weichenrieder (2009) defined Type II ORF1p as the Transposase_22 domain, and Type I ORF1p as a combination of RRM and zf-CCHC domains (fig. 1). Mammalian species are dominated by Transposase_22 ORF1 proteins (fig. 9a); as expected from the Type II classification. However, some mammalian species also contain ORF1 proteins with RRM or zf-CCHC domains – which are more characteristic of Type I, and are likely very ancient. There was even a Type II variant found: Several ORF1p in *Myotis lucifugus* display an RRM domain

Fig. 8.—Continued

RT-families with >5 members at > 90% identity are shown in this tree. Node are labelled as follows: By species name if there is only one species in the family (e.g. Loxodonta africana); by genus name if there are multiple species of the same genus (e.g. Sus); by multiple genus names if there are multiple genera in the family (e.g. Ailuropoda; Ursus); and by clade name if there are more than five genera (e.g. Primates). The number in parentheses after the node name indicates the number of elements in the family.

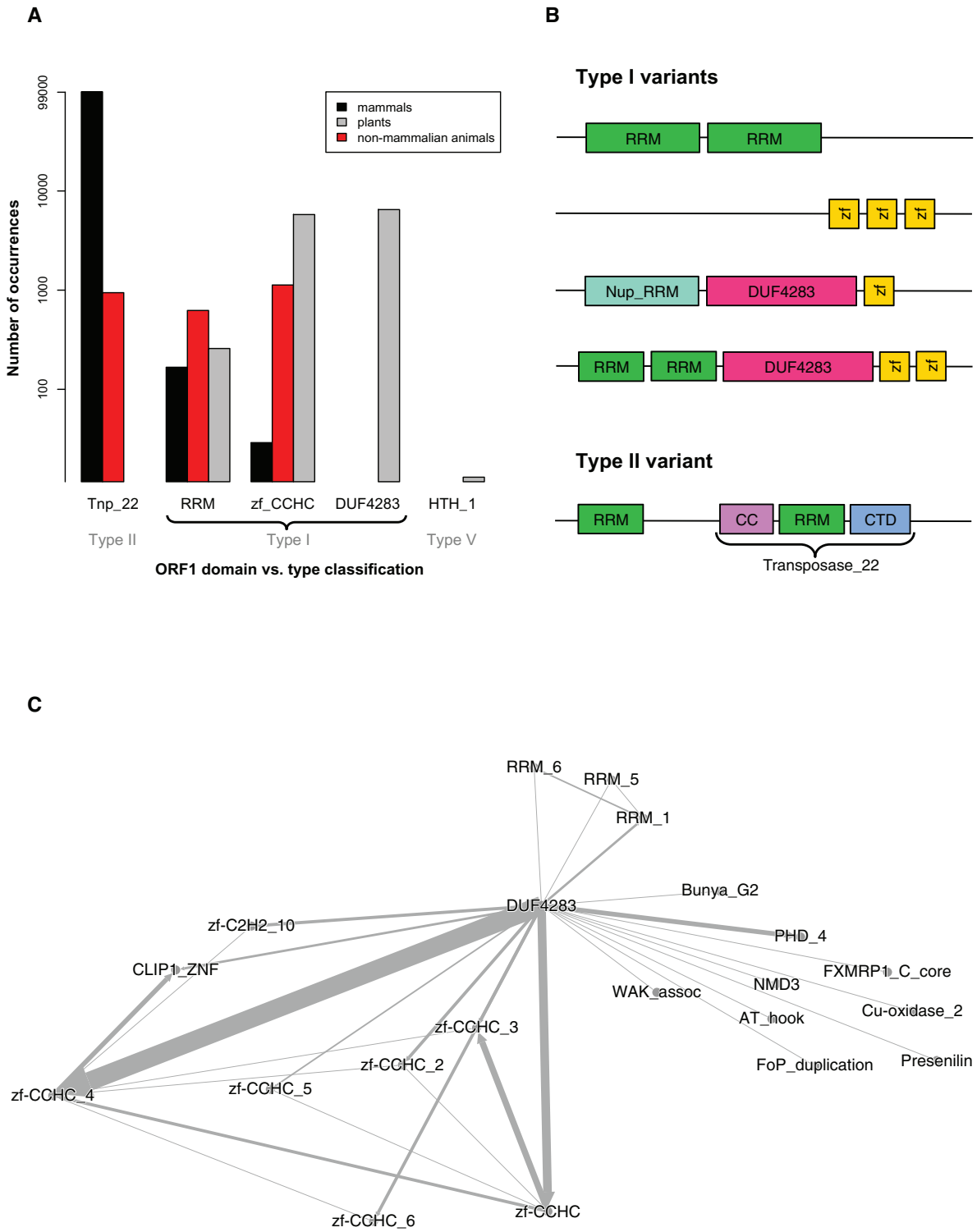


Fig. 9.—ORF1p clustering and domain identification analysis. (*a*) ORF1p domain summary from HMM–HMM comparison. Transposase_22 (Tnp_22), RNA recognition motifs (RRM), and zinc fingers (zf-CCHC) are known ORF1p domains. The *y*-axis shows the number of times these appeared in each group of species (mammals, non-mammalian animals, plants), on a log scale. Several unknown domains also appeared frequently; for example, DUF4283 was

before the expected Transposase_22 (fig. 9b), which has not been previously documented.

Non-mammalian animals contain the typical Type II ORF1p, Type I ORF1p, and assorted combinations of RRM/zf-CCHC domains. These appear as variants of Type I ORF1p (fig. 9b) but are consistent with the Tx1 clade of retrotransposons and RT-based phylogeny (fig. 8). There are numerous studies that describe these domains in depth, for example, Kojima and Fujiwara (2004) and Kordis et al. (2006).

In plants there were many ORF1p with RRM or zf-CCHC domains, indicative of Type I proteins. As mentioned above, several species harboured novel Nup_RRM or RRM domains. However, the overwhelmingly dominant plant ORF1p domain was DUF4283: An uncharacterized domain of unknown function (Finn et al. 2010). Figure 9c shows a directed network graph of the most frequently seen ORF1p domains across Viridiplantae. For all other species, this graph is centred around Transposase_22, RRM or zf-CCHC domains (see supplementary fig. S7a–f, Supplementary Material online). In plants, DUF4283 appears to act as the primary ORF1p classifier, strongly associated with zf-CCHC_4 (fig. 9c).

Coccomyxa subellipsoidea does not contain any of these domains – instead, the entire ORF1p region is enveloped by HTH_1 (fig. 9a): A bacterial regulatory helix-turn-helix protein of the LysR family (Finn et al. 2010). *Coccomyxa subellipsoidea* L1s are 98% identical to Zepp (fig. 8), a LINE-like retrotransposon found in *Chlorella vulgaris* (Higashiyama et al. 1997). *Chlorella vulgaris* was not included in this study as the assembly is only available in contig form. However, another *Chlorella* species (*C. variabilis*) was included and showed minimal, fragmented L1 presence (fig. 4). Given that *Coccomyxa subellipsoidea* and *C. vulgaris* share such high L1 identity, yet this is missing from the closely related *C. variabilis* species, it is possible that a horizontal transfer event occurred between the first two species. Alternatively, TEs have a tendency to take necessary proteins directly from their host (Abrusan et al. 2013; Heitkam et al. 2014); this may also explain the newly acquired HTH_1 proteins.

Antisense Characteristics of Active L1s

The analysis of ORF1 and ORF2 sequences across genomes led to the discovery of an antisense open reading frame overlapping ORF1. This novel ORF was initially noticed in the panda genome (*Ailuropoda melanoleuca*), where it is present in almost every L1 element that has both ORFs intact (1157/1200). As a result, we screened each genome for strictly active L1s (i.e. both ORF1 and ORF2 intact) to determine whether other species contained similar antisense ORFs (i.e. overlapping ORF1 in the reverse direction and about 1 kb in length). Apart from panda, only eight other mammalian species contained anything remotely similar (fig. 10a), albeit at lower copy number. No such reverse ORFs were found in any of the non-mammalian animal or plant species. Interestingly, these ORFs only appeared in mammalian species with a substantial number of active L1s (e.g. minke whale, baiji, dog, rat), suggesting that they might somehow contribute to L1 retrotransposition; yet they are noticeably absent from all of the primates, including snub-nosed monkey. They are also clearly distinct from the primate-specific antisense ORF0 (Denli et al. 2015), which is much shorter and upstream of ORF1.

Using the same procedure as previously described for ORF2p, we extracted and aligned the reverse ORF proteins in each species to generate a representative consensus sequence, then aligned the consensus sequences and inferred maximum likelihood and Neighbor-Joining phylogenies (fig. 10b shows the maximum likelihood tree). The only difference between the trees was the position of *Myotis brandtii* (outgroup to minke whale/baiji on NJ tree, with low support). The reverse ORF proteins found in dog *Canis lupus* and Siberian tiger *Panthera tigris* appear to be a distinct type of reverse ORFp, denoted r2. Both r1 and r2 ORFs were found in the rat genome (*Rattus norvegicus*). All reverse ORF proteins were checked for similarity to known domains using HMMer (Finn et al. 2011). The most significant hits came from *Myotis brandtii* (r1 ORF, only 19/68 non-redundant sequences), which showed homology to the Pico_P1A picornavirus coat protein; and *Canis lupus* (r2 ORF, all 81/81 non-redundant sequences), which showed a range of hits from various transporter and initiation molecules (e.g. ZIP: Zinc transporter, Rrn6: RNA polymerase I-specific transcription–initiation factor, Afi1: Docking domain of Afi1 for Arf3 in vesicle trafficking).

Fig. 9.—Continued

found in every plant species except *Coccomyxa subellipsoidea*, which harboured HTH_1 ORF1 proteins instead. (*b*) Variants of Type I ORF1 proteins. Type I ORF1p typically has at least 1 RRM and 1 zf-CCHC; Type II ORF1p is characterized as the Transposase_22 domain. This figure highlights type variants found in the analyzed species: for example, lack of zf-CCHC motifs, seen in mosquitos; lack of RRM domains, seen in sea squirts; Nup_RRM instead of RRM, seen in some plants; over-representation of unknown DUF4283 domain in almost all plants; and an additional RRM before the Transposase_22 in some mammals, for example, bat *Myotis lucifigus*. Supplementary table S11, Supplementary Material online shows the ORF1p domains in each species. (*c*) Directed network graph of Type I ORF1 protein domains found in plants. Each ORF1p in each L1 (in each plant species) was screened using HMMer against the Pfam database. The highest-scoring domain hit was ranked first; other domains also found within that ORF1p sequence were listed next, by decreasing score. This was used to construct a network graph of the associated domain. DUF4283 was the most frequently seen, highest scoring domain – it is the centroid of the graph. RRM and zf-CCHC domains are associated with this domain (especially zf-CCHC_4), but it is the unknown domain that acts as the vital ORF1p identifier in plants.

Discussion

Extinction of L1s in Mammalian Taxa – Known Versus New Events

An L1 element is called 'extinct' if it completely loses its ability to retrotranspose. If there is very low (but still extant) activity in the genome, this has been referred to as 'quiescence' rather than extinction (Yang et al. 2014). Figure 3 shows all of the known cases of L1 extinction (not quiescence) out of the 98 mammalian species analysed in this study: Three pteropodid bats (Cantrell et al. 2008; Yang et al. 2014) and the thirteen-lined ground squirrel *Ictidomys tridecemlineatus* (Platt and Ray 2012). Interestingly, the TBLASTN analysis found intact ORF2 in nucleotide sequences from squirrel – so in figure 3, this species is annotated L1-active. It is possible that squirrel is a case of quiescence rather than extinction, or the ORF2 regions are structurally conserved rather than functional. Other confirmed cases of L1 extinction include the spider monkey (Boissinot et al. 2004) and all studied Sigmondontinae rodents except for the Sigmodontini tribe (Casavant et al. 2000; Grahn et al. 2005), which were not included in this study because there are no public genome assemblies available.

Novel L1 extinction species candidates found in this study include eight rodents, five cetartiodactyls, one carnivore, one perissodactyl, four bats, two Insectivora, four Afrotherian mammals and one marsupial (fig. 3). Gallus et al. (2015) recently investigated L1 dynamics in Tasmanian devil – their results also suggest that this marsupial has lost L1 functionality. To our knowledge, the remaining species have not been previously studied as L1 extinction candidates, although some closely related species have been, for example, *Peromyscus californicus* (Casavant et al. 1998).

Evidence of a retro-element extinction event is often difficult to confirm, because we cannot determine whether it occurred in the individual genome or at the species level. The easiest extinction event to observe is one that is ancestral, such that a large monophyletic group of species all lack evidence of recent L1 activity (Grahn et al. 2005). For example, Cantrell et al. (2008) confirmed L1 extinction of the Pteropodidae megabat family by showing that the event had been inherited in 11 sampled genera. There are no other monophyletic extinction events shown in the mammalian phylogeny (fig. 3). Instead, all of the new L1 extinction candidate species appear paraphyletic or polyphyletic.

There are several possible explanations for these occurrences. First, these may be individual organism-specific changes – as with the putative extinction of L1s in the ground squirrel, which corresponded to a steady decline of all TE classes in that genome (Platt and Ray 2012), or the similar scenario seen in Tasmanian devil (Gallus et al. 2015). Second, the re-emergence or persistence of L1 activity in closely related species suggests that these are examples of quiescence rather than extinction. This may especially be true for rodents, where we already know of several extinct/quiescent species (Casavant et al. 1998, 2000). Such a scenario suggests that there is a fine line between calling an L1 active or extinct, and a lot of these rodents may have only recently become inactive. The fact that numerous rodent species (eight in fig. 3 alone, not including previous studies) have no intact ORF2 argues that the entire group may be headed towards L1 extinction (disregarding mouse and rat, which are extraordinarily L1-active). The naked mole rat (*Heterocephalus glaber*) and blind mole rat (*Nannospalax galili*) are among these putatively 'L1-extinct' species: Two species renowned for their cancer resistance. Given the deleterious effects that L1 activity can cause, if these rodents are truly L1-extinct, it would likely be a consequent of robust host suppression mechanisms (Deininger et al. 2003; Han and Boeke 2005).

Lastly, it is possible that these supposedly extinct species appear so because of the draft quality of the genome assemblies used. There are several cases (e.g. wallaby *Macropus eugenii*) where intact ORF2 could only be found in the NR/HTGS NCBI databases, not in the genome assembly. Indeed, many of the species colored in blue (e.g. *Leptonychotes weddellii, Bubalus bubalis*) have short Illumina read assemblies with low contig N50 values – making it virtually impossible to find perfectly intact ORF2 sequences. Gallus et al. (2015) experienced the same problem when mining the Tasmanian devil genome for intact L1s. More reliable analyses such as long read Sanger sequencing or *in situ* hybridization would be needed to confirm complete loss or presence of L1 activity (Grahn et al. 2005; Cantrell et al. 2008).

The Difference between Retrotransposition Potential and Activity

The majority of this study focuses on identifying L1 elements that have retrotransposition potential, and therefore may be active within the genome and causing change. But what does it mean for an L1 to be active? We can label an element as having the potential to be active by looking for intact open reading frames, or calculating the proportion of intact full-length L1s in the genome. But to be truly active, the element must provide evidence that it is doing something in the genome, not just that it has the potential to. So for L1 elements, effective activity should be confirmable by substantial replication and propagation of the element throughout the genome.

The distribution of L1* proportions shown in figure 5 clearly illustrates this concept. There are three things that are immediately obvious in this figure: (1) non-mammalian animal species (shown in red) and plant species (e.g. green alga) have a surprisingly high proportion of potentially active elements but low copy number; (2) the majority of mammals have a huge number of potentially active L1s, but a consistently low (<20%) proportion; (3) several mammalian species (e.g. minke whale, antelope, snub-nosed monkey, mouse,

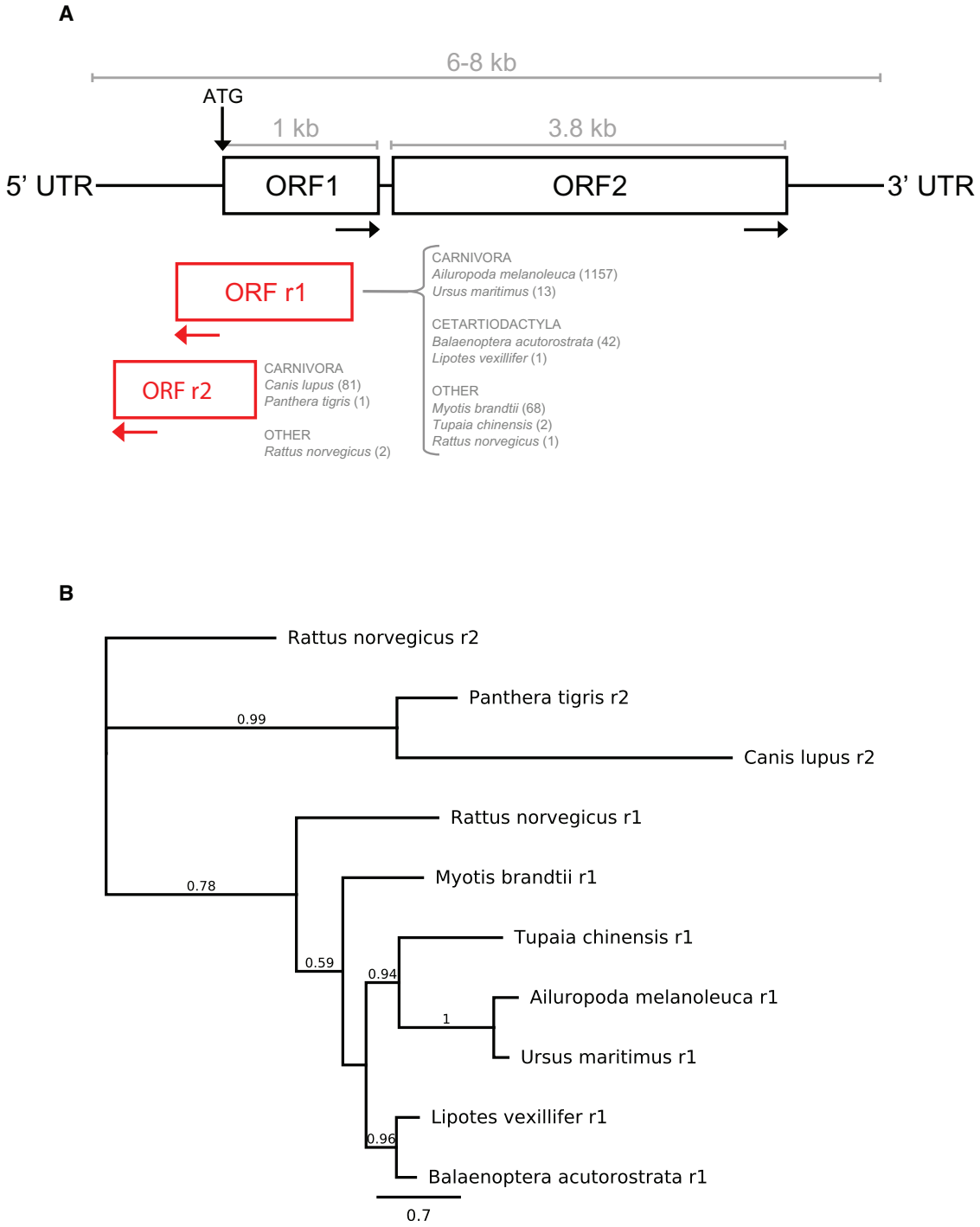


Fig. 10.—Novel antisense open reading frames found in some mammals. (*a*) Characteristics and distribution of the antisense ORFs. The position and approximate size of the novel antisense ORFs, as well as the order/species they are found in and the number of L1s that contain this ORF (in brackets). These ORFs have no known functional domains. (*b*) Antisense ORFp species consensus tree. Maximum likelihood phylogeny inferred using FastTree from extracted and aligned L1 reverse ORFp consensus sequences. Expected species relationships appear preserved within the r1 and r2 clades.

sheep) stand out because they have a high L1* proportion, unlike the other mammals. The variation between species illustrates those that are potentially active versus those that are truly active. However, we cannot establish a population variance because for almost all cases there is only one individual per species, due to the available data.

Addressing the first of these observations – non-mammalian species (plants and animals) all seem to have a relatively low L1 copy number. This is not unexpected in itself; many of these elements are divergent and have accumulated mutations, suggesting that they are older than their mammalian counterparts (as shown by the longer branch lengths in fig. 8). What is surprising is that, based on the identification of intact ORFs, a large proportion of L1s in these genomes seem putatively active. For instance, green alga (*Coccomyxa subellipsoidea*) only has 15 full-length L1s, yet all 15 of them are apparently active. But are these L1s really active? Such low copy number would suggest that there is high retrotransposition potential, but low effectiveness or a high turnover rate.

In contrast, we know that mammalian species typically have a high L1 copy number (Lander et al. 2001; Mouse Sequencing Consortium et al. 2002). We also know that L1 retrotransposition is extremely inefficient because the vast majority of new insertions are 5′ truncated and thus inactive (Sassaman et al. 1997; Boissinot et al. 2000). This seems to be the case for most of the mammals analyzed in this study: Although they have a high number of active L1s, the number of inactive L1s is much greater (~80%); hence they have a low level of observable activity within the genome.

However, there are a few mammals that have both a high L1 copy number and a high active percentage in the genome. Indeed, the most significantly 'hyperactive' species (minke whale) has never been mentioned before in the context of L1 activity, yet it contains 5006 active L1s that make up more than 62% of the total full-length L1 content in the genome – far surpassing the retrotranspositional activity of mouse. This directly contradicts the belief that most full-length L1s are inactive or truncated during replication. As such, it is a good indication that these species are truly active, not just potentially active. These L1s are dynamically replicating and expanding within the genome, resulting in a large copy number of elements that share high pairwise identity with each other. Therefore, out of the 206 putatively active species found in this analysis, these five genomes would be the best model organisms for studying genomic change due to L1 retrotransposition.

The Master Lineage Paradigm

The master lineage model is an evolutionary scenario where the active elements in a genome give rise to a single active lineage that dominates long-term retrotransposition (Clough et al. 1996). Phylogenetic analyses such as dendrogram constructions are often used to give an indication of existent lineages (Grahn et al. 2005; Adelson et al. 2009), under the rationale that longer branch lengths represent accumulated mutations (including insertions and deletions) due to age, whereas shorter branch lengths signify younger, closely related elements with little nucleotide divergence from the master template. If all of the active elements form polytomies with very short-branch lengths, as opposed to multiple divergent clusters, then this would be an example of a strict master lineage model.

It is hypothesized that there is selective pressure for the master LINE (and/or SINE) lineage to monopolise active retrotransposition in mammalian model organisms (Platt and Ray 2012). Our data supports this – all of the 'hyperactive' species and many of the potentially active ones contain a single active L1 family/cluster, as shown in figure 6 with the snub-nosed monkey example. This seems somewhat counterintuitive; given the vast number of active elements, it should be feasible for numerous independent lineages to amplify, over time. A possible explanation is that the single lineage we observe is due to a master element that was particularly effective at evading host suppression mechanisms, and thus initiated widespread retrotransposition throughout the genome.

In some species with relatively low active copy number, such as *Myotis lucifugus* (fig. 7), there appear to be multiple simultaneously active lineages. *Myotis lucifigus* also contains some L1 elements with a peculiar Type II ORF1p variant (fig. 9b), and some ORF1p with the traditional Transposase_22 domain, supporting the theory of different L1 lineages. A similar situation was observed in the (now extinct) megabat L1s (Yang et al. 2014) and two putatively active L1 lineages in rodent *Peromyscus californicus* (Casavant et al. 1998). There are various theories as to how multiple lineages may arise; for example, after a period of low activity, multiple 'stealth driver' (Cordaux and Batzer 2009) elements may be driven to retrotranspose at the same time; or horizontal acquisition of a retroelement from a different species can produce a foreign active lineage alongside the native lineage. Nonetheless, not much is known about how both lineages can be maintained, if there really is selective pressure to adhere to a master model. Yang et al. (2014) speculate that if the lineages are specialized in different tissue types (e.g. male germ line vs. female germ line), they can co-exist without competition – however, this is countered by the observation that in mouse, most L1 retrotransposition events seem to occur in the early embryo rather than in germ cells (Kano et al. 2009). Furthermore, the fact that we do not observe any high copy number species harboring more than one lineage suggests that multiple lineages are inhibitory to retrotransposition: Either through competition, or because it increases the chance that both lineages will be detected and suppressed by regulatory mechanisms, so neither lineage can effectively proliferate within the genome.

Discordance between ORF Nomenclature and Domain Classification

A predictable side effect of having access to more data and discovering new domains is that the existing nomenclature may need revision to reflect this new information. Based on the existing Type system for ORF1p elements (Khazina and Weichenrieder 2009), mammals typically have Type II; non-mammalian animals have both Types I and II; plants have variants of Type I; and the single remaining plant species (*Coccomyxa subellipsoidea*) belongs to Type V: Unclassified ORF proteins (fig. 9a and b). Such a categorization can be misleading because it implies that Type I sequences are alike and share high amino acid similarity – and even the HTH_1 domain in *C. subellipsoidea* cannot be that distantly related, by virtue of it being an 'ORF1p'. But at what point does a domain variant become too different to be an ORF1p? A phylogeny of ORF1p could not be reliably inferred because of the extreme variation found within these sequences, and the all-against-all clustering analysis showed that there are multiple independent ORF1p clusters within each species - despite using the default settings where two proteins in a pair are included in the same family if the homologous segment pairs have at least 35% similarity over 80% coverage (Penel et al. 2009). The protein domain network diagrams (e.g. fig. 9c) further show that the 'known' ORF1 domains are not always the key identifiers, and there are numerous strongly associated domains that are often overlooked.

Accordingly, we propose a more informative revision to the nomenclature to refer to ORF proteins by the dominant functional domain(s); for example, ORF2p = RVT_1-ORFp for mammals, or (RVT_1 + RVT_3)-ORFp for most plants (see supplementary table S10 and fig. S6*a–g*, Supplementary Material online). Likewise, ORF1p = HTH_1-ORFp for *C. subellipsoidea*. This allows us to forego predetermined Type or ORF# labels, especially for unusual cases. The discovery of additional ORF proteins such as the primate-specific ORF0 (Denli et al. 2015) or the reverse ORF proteins found in this study (fig. 10) makes a compelling argument for re-naming.

Confounding Bias Due to Genome Assembly Quality

Advances in technology mean that genomes are now being sequenced at alarmingly fast rates. However, once sequenced, many genomes tend to remain in their error riddled, scaffolded state. The majority of genomes used in this study are draft assemblies, so it is important to check that the quality of the assembly is not affecting the results (either by restricting the ability to detect repetitive 6kb elements, or by creating false positive hits from misread errors). Accordingly, we analysed independently-assembled closely related species (within the same genus or species) and used multiple searching strategies (e.g. LASTZ with genomic data versus TBLASTN with nucleotide databases). Consider the three horse genomes included in this study: *Equus przewalski* (submitted by IMAU, contig N50 of 57,610, SOAPdenovo assembly method used), *Equus caballus* Thoroughbred (submitted by GAT, contig N50 of 112,381, ARACHNE2.0 assembly method used) and *Equus caballus* Mongolian (submitted by IMAU, contig N50 of 40,738, SOAPdenovo assembly method used) (see supplementary tables S1–S3, Supplementary Material online). Based on the submitter, contig N50 and assembly method, *Equus przewalski* and the Mongolian *Equus caballus* would be expected to be the most similar. Based on species relationships, one would expect the two *Equus caballus* horses to be more similar. However, the actual findings show that while all three horses are marked L1*, only *Equus przewalski* and *Equus caballus* (Thoroughbred) have intact ORF2 in the genome. *Equus caballus* (Mongolian) was determined L1-active solely based on the TBLASTN results. This is a known problem with using draft assemblies – and it has been detailed previously with the Tasmanian Devil genome (Gallus et al. 2015), as well as the wallaby and cat genomes (Pontius et al. 2007; Renfree et al. 2011). It is likely that as genome assemblies improve, it will become possible to detect more ORF2-intact, active L1 (although the overall L1-status is unlikely to change).

As a contrasting example, the three *Arabidopsis* species that were submitted independently (*A. halleri:* TokyoTech, *A. lyrata:* JGI, *A. thaliana: Arabidopsis* Information Resource), have very different contig N50 values (*A. halleri:* 2864, *A. lyrata:* 227,391, *A. thaliana:* 11,194,537) and used different sequencing strategies (*A. halleri:* Illumina, *A. lyrata:* Sanger, *A. thaliana:* BAC physical map then Sanger sequencing of BACs) have very similar results in terms of L1 presence, activity and open reading frame structure. In fact, Illumina seems to be the most widely used sequencing technology across all the genomes (mammalian, non-mammalian, and plant) but it does not appear to introduce platform specific artifacts. This is encouraging because it demonstrates that draft genomes can be used to study repetitive sequences such as L1s, as long as suitable quality controls are taken into account.

The assembly level does not seem to hinder the ability to detect highly L1-active species (more so the ability to confirm L1 extinction). Out of the five so-called 'hyperactive' mammalian species labelled in figure 5, three (minke whale, snub-nosed monkey, antelope) are scaffold-level assemblies, whereas two (mouse and sheep) are chromosome-level with noticeably higher N50 values. One might argue that this just shows that draft assemblies are more likely to have duplication or misread errors, leading to greater L1 copy number. However, a de-duplication test of these genomes found very few identical hits (e.g. minke whale contains 13,681 L1s over 3 kb: The largest cluster of duplicates had 47 elements, and only two L1s shared the same 1 kb flanking region). This suggests that the majority of identical hits are likely to be true duplicates rather than assembly errors.

Implications for Our Perception of Genome Evolution

This study complements those of Kordis et al. (2006) (deuterostomes), Khan et al. (2006) (primates), Sookdeo et al. (2013) (mouse), Yang et al. (2014) (megabats), Metcalfe and Casane (2014) (Jockey non-LTR elements), and Heitkam et al. (2014) (plants) in demonstrating the diversity of TE evolutionary patterns across species. We have identified over 10 million L1 sequences from 503 different genomes, including ORF1 and ORF2 proteins with novel domain variations that strain the current L1 classification system. While most animals and plants still exhibit some form of L1 activity, the discovery of new extinction candidates leaves us better equipped to identify common factors in the genomic landscape that contribute to TE suppression (particularly in species with desirable characteristics, such as cancer resistance). Conversely, investigation into 'hyperactive' species such as minke whale and snub-nosed monkey, whose retrotranspositional activity seems to far surpass that of human, rat and mouse, could be used to study the extent to which L1s cause genomic change. Perhaps the presence of reverse ORFs helps the L1 in these species to attain hyperactivity. Multiple lines of evidence suggest that L1s can form an 'ORF-anage' by recruiting functional domains from the host, thus propagating their activity in the genome. As always, it is likely that our findings here are only the very tip of the iceberg. We present this data with the hope that it will provide a definitive reference for future studies, aiding our understanding of eukaryotic evolution.

Supplementary Material

Supplementary figures S1–S7 and Supplementary tables S1–S11 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank our collaborators (Broad Institute of MIT and Harvard for the chromosome-level elephant genome, and Frank Grutzner from the University of Adelaide for the echidna assembly) for making private genome assemblies available to us. We would also like to thank Iain Searle for taking the time to read the manuscript in full and offer helpful comments, and the anonymous reviewers for their clear, detailed feedback. Finally, this paper would not be possible without the invaluable insights of Lu Zeng, R language mastery of Reuben Buckley, HMMer knowledge of Zhipeng Qu, and extraordinary IT support from Matt Westlake.

Literature Cited

Abrusan G, Szilagyi A, Zhang Y, Papp B. 2013. Turning gold into 'junk': transposable elements utilize central proteins of cellular networks. Nucleic Acids Res. 41:3190–3200.

Adelson DL, Raison JM, Edgar RC. 2009. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci U S A. 106:12855–12860.

Adelson DL, Raison JM, Garber M, Edgar RC. 2010. Interspersed repeats in the horse (*Equus caballus*); spatial correlations highlight conserved chromosomal domains. Anim Genet. 41 (Suppl 2):91–99.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Beck CR, et al. 2010. LINE-1 retrotransposition activity in human genomes. Cell 141:1159–1170.

Beck RM, Bininda-Emonds OR, Cardillo M, Liu FG, Purvis A. 2006. A higher-level MRP supertree of placental mammals. BMC Evol Biol. 6:93.

Biedler J, Tu Z. 2003. Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity. Mol Biol Evol. 20:1811–1825.

Blass E, Bell M, Boissinot S. 2012. Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback. Genome Biol Evol. 4:687–702.

Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol. 17:915–928.

Boissinot S, Roos C, Furano AV. 2004. Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. J Mol Evol. 58:122–130.

Brouha B, et al. 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 100:5280–5285.

Cantrell MA, Scott L, Brown CJ, Martinez AR, Wichman HA. 2008. Loss of LINE-1 activity in the megabats. Genetics 178:393–404.

Casavant NC, Lee RN, Sherman AN, Wichman HA. 1998. Molecular evolution of two lineages of L1 (LINE-1) retrotransposons in the california mouse, *Peromyscus californicus*. Genetics 150:345–357.

Casavant NC, et al. 2000. The end of the LINE? Lack of recent L1 activity in a group of South American rodents. Genetics 154:1809–1817.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chen JM, Stenson PD, Cooper DN, Ferec C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet. 117:411–427.

Clough JE, Foster JA, Barnett M, Wichman HA. 1996. Computer simulation of transposable element evolution: random template and strict master models. J Mol Evol. 42:52–58.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 10:691–703.

Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. EMBO J. 21:5899–5910.

Deininger PL, Moran JV, Batzer MA, Kazazian HH. Jr. 2003. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 13:651–658.

Denli AM, et al. 2015. Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity. Cell 163:583–593.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 35:41–48.

Douzery EJ, et al. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol Biol Evol. 31:1923–1928.

Duvernell DD, Pryor SR, Adams SM. 2004. Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover. J Mol Evol. 59:298–308.

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

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 24:363–367.
Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29–W37.
Finn RD, et al. 2010. The Pfam protein families database. Nucleic Acids Res. 38:D211–D222. 2.
Furano AV. 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol. 64:255–294.
Furano AV, Duvernell DD, Boissinot S. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet. 20:9–14.
Gallus S, et al. 2015. Evolutionary histories of transposable elements in the genome of the largest living marsupial carnivore, the Tasmanian devil. Mol Biol Evol. 32:1268–1283.
Graham T, Boissinot S. 2006. The genomic distribution of L1 elements: the role of insertion bias and natural selection. J Biomed Biotechnol. 2006:75327.
Grahn RA, Rinehart TA, Cantrell MA, Wichman HA. 2005. Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet Genome Res. 110:407–415.
Gregory SG, et al. 2002. A physical map of the mouse genome. Nature 418:743–750.
Han JS, Boeke JD. 2005. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression?. Bioessays 27:775–784.
Harris RS. 2007. Improved Pairwise Alignment of Genomic DNA. Ph.D. Thesis, The Pennsylvania State University.
Heitkam T, et al. 2014. Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. Plant J. 79:385–397.
Heitkam T, Schmidt T. 2009. BNR - a LINE family from Beta vulgaris - contains a RRM domain in open reading frame 1 and defines a L1 sub-clade present in diverse plant genomes. Plant J. 59:872–882.
Heras SR, et al. 2006. L1Tc non-LTR retrotransposons from *Trypanosoma cruzi* contain a functional viral-like self-cleaving 2A sequence in frame with the active proteins they encode. Cell Mol Life Sci. 63:1449–1460.
Higashiyama T, Noutoshi Y, Fujie M, Yamada T. 1997. Zepp, a LINE-like retrotransposon accumulated in the Chlorella telomeric region. EMBO J. 16:3715–3723.
Janecka JE, et al. 2007. Molecular and genomic data identify the closest living relative of primates. Science 318:792–794.
Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.
Kaer K, Speek M. 2013. Retroelements in human disease. Gene 518:231–241.
Kano H, et al. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. 23:1303–1312.
Kapitonov VV, Jurka J. 2003. The esterase and PHD domains in CR1-like non-LTR retrotransposons. Mol Biol Evol. 20:38–46.
Kazazian HH. Jr. 1998. Mobile elements and disease. Curr Opin Genet Dev. 8:343–350.
Kazazian HH. Jr. 2000. Genetics. L1 retrotransposons shape the mammalian genome. Science 289:1152–1153.
Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.
Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16:78–87.
Khazina E, Weichenrieder O. 2009. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci U S A. 106:731–736.
Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.
Kojima KK, Fujiwara H. 2004. Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. Mol Biol Evol. 21:207–217.
Kondo-Iida E, et al. 1999. Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). Hum Mol Genet. 8:2303–2309.
Kordis D, Lovsin N, Gubensek F. 2006. Phylogenomic analysis of the L1 retrotransposons in Deuterostomia. Syst Biol. 55:886–901.
Kriventseva EV, et al. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 43:D250–D256.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Lee J, et al. 2007. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. Gene 390:18–27.
Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol. 23:450–468.
Maddison DR, Schulz KS. The Tree of Life Web Project. Available at: http://tolweb.org. (2007).
Mathews LM, Chi SY, Greenberg N, Ovchinnikov I, Swergold GD. 2003. Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. Am J Hum Genet. 72:739–748.
Metcalfe CJ, Casane D. 2014. Modular organization and reticulate evolution of the ORF1 of Jockey superfamily transposable elements. Mob DNA 5:19.
Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.
Moran JV, et al. 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87:917–927.
Mouse Genome Sequencing Consortium, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Murphy WJ, et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.
Nakamura M, Okada N, Kajikawa M. 2012. Self-interaction, nucleic acid binding, and nucleic acid chaperone activities are unexpectedly retained in the unique ORF1p of zebrafish LINE. Mol Cell Biol. 32:458–469.
Novick PA, Basta H, Floumanhaft M, McClure MA, Boissinot S. 2009. The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard Anolis carolinensis shows more similarity to fish than mammals. Mol Biol Evol. 26:1811–1822.
Odon V, et al. 2013. APE-type non-LTR retrotransposons of multicellular organisms encode virus-like 2A oligopeptide sequences, which mediate translational recoding during protein synthesis. Mol Biol Evol. 30:1955–1965.
Penel S, et al. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 (Suppl 6):S3.
Penzkofer T, Dandekar T, Zemojtel T. 2005. L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res. 33:D498–D500.
Platt RN, 2nd, Ray DA. 2012. A non-LTR retroelement extinction in Spermophilus tridecemlineatus. Gene 500:47–53.
Pontius JU, et al. 2007. Initial sequence and comparative analysis of the cat genome. Genome Res. 17:1675–1689.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Renfree MB, et al. 2011. Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. Genome Biol. 12:R81.
Sassaman DM, et al. 1997. Many human L1 elements are capable of retrotransposition. Nat Genet. 16:37–43.
Sayers EW, et al. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 40:D13–D25.
Scott AF, et al. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics 1:113–125.
Seleme MC, et al. 2006. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A. 103:6611–6616.
Smyshlyaev G, Voigt F, Blinov A, Barabas O, Novikova O. 2013. Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. Proc Natl Acad Sci U S A. 110:20140–20145.
Sookdeo A, Hepp CM, McClure MA, Boissinot S. 2013. Revisiting the evolution of mouse LINE-1 in the genomic era. Mob DNA 4:3.
Tollis M, Boissinot S. 2013. Lizards and LINEs: selection and demography affect the fate of L1 retrotransposons in the genome of the green anole (*Anolis carolinensis*). Genome Biol Evol. 5:1754–1768.
Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL. 2013. Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci U S A. 110:1012–1016.
Waters PD, Dobigny G, Waddell PJ, Robinson TJ. 2007. Evolutionary history of LINE-1 in the major clades of placental mammals. PLoS One 2:e158.
Yang L, Brunsfeld J, Scott L, Wichman H. 2014. Reviving the dead: history and reactivation of an extinct l1. PLoS Genet. 10:e1004395.
Zmasek C. 2015. Archaeopteryx: visualization, analysis, and editing of phylogenetic trees. Available at: https://sites.google.com/site/cmzmasek/home/software/archaeopteryx.

Associate editor: Esther Betran

Chapter 3

Horizontal Transfer of Retrotransposons has Shaped the Genomes of Modern Eukaryotes

"Problems worthy of attack prove their worth by fighting back."

— Piet Hein

Evidence of horizontal transfer (HT) is typically identified as a sporadic distribution of the retrotransposon across species and high sequence similarity (restricted solely to the retrotransposon) between divergent species. BovB retrotransposons, which were initially discovered because of their unexpected congruence between cattle and snake species, clearly satisfy these criteria. L1 elements, on the other hand, have a strong ancestral background in almost all branches of the eukaryotic tree of life. Distinguishing HT candidate elements from those that have been vertically inherited poses a significant challenge, and requires a re-analysis of the current techniques used to determine transfer. The following manuscript uses BovB to identify the distinguishing characteristics of horizontal transfer events, and thus find similar events involving L1 elements. It has been prepared for submission to *Science* in the form of a Scientific Report: abstract, brief introduction and major findings. Materials and Methods are included as supplementary material.

Statement of Authorship

Title of Paper	Horizontal transfer of retrotransposons has shaped the genomes of modern eukaryotes
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Atma M. Ivancevic, R. Daniel Kortschak, Terry Bertozzi, David L. Adelson (2016). Horizontal transfer of retrotransposons has shaped the genomes of modern eukaryotes. Prepared for submission as a Scientific Report to Science.

Principal Author

Name of Principal Author (Candidate)	Atma M. Ivancevic		
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	6/12/16

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. Daniel Kortschak		
Contribution to the Paper	Supervised the development of work and assisted in analysing the results and writing the manuscript.		
Signature		Date	6/12/16

Name of Co-Author	Terry Bertozzi		
Contribution to the Paper	Supervised the development of work, provided access to DNA samples and performed wet lab experiments, assisted in analysing the results and writing the manuscript.		
Signature		Date	14.xii.16

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	7 Dec 2016

Horizontal transfer of retrotransposons has shaped the genomes of modern eukaryotes

Atma Ivancevic,[1] Daniel Kortschak,[1] Terry Bertozzi,[1,2] David Adelson[1*]

[1]School of Biological Sciences, The University of Adelaide, SA 5005, Australia
[2]Evolutionary Biology Unit, South Australian Museum, SA 5000, Australia

[*]To whom correspondence should be addressed; E-mail: david.adelson@adelaide.edu.au

Accumulating reports of horizontal transfer (HT) in eukaryotes are changing our perception of the role that retrotransposons play in genome evolution. In this study, we compared BovB and L1 retrotransposons across the eukaryotic tree of life to identify characteristics indicative of horizontal transfer. We extended the analysis of BovBs to include: a wider range of species (e.g. bats, frog), new vectors of transfer in the form of blood-sucking parasites, and more than twice the number of estimated transfer events compared to previous studies. Contrary to popular belief, we found evidence to support multiple ancient L1 HT events in plants and further support that L1s were introduced to mammalian genomes after the monotreme-therian split. We conclude that both BovB and L1 elements are capable of HT, although the rate of transfer differs significantly. With regard to L1s, while the transfer is not recent or frequent, the extensive colonisation of therian mammals illustrates the drastic and long-term impact of introducing foreign DNA into new host genomes.

Transposable elements (TEs) are mobile segments of DNA which occupy large portions of eukaryotic genomes, including more than half of the human genome (*1*). Retrotransposons

are TEs which move from site to site using a "copy and paste" mechanism, facilitating their amplification throughout the genome (*2, 3*). The insertion of retrotransposons can interrupt existing genetic structures, resulting in gene disruptions, chromosomal breaks and rearrangements, and numerous diseases such as cancer (*4–6*). Two of the most abundant retrotransposon families in eukaryotes are LINE-1 (L1) and Bovine-B (BovB) (*7, 8*).

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 (*2, 9*). Studies investigating the possibility of HT in retrotransposons are limited, mainly including CR1s and RTEs (*10–13*). Given the limited evidence to date, we tested the hypothesis that horizontal transfer is a ubiquitous process not restricted to certain species or retrotransposons. We used L1 and BovB elements as exemplars because of their contrasting dynamics and predominance in mammalian genomes. BovB retrotransposons provide an excellent example of horizontal transfer: divergent species contain highly similar BovB sequences and the analysis of various tick species reveals a plausible vector of transfer (*10*). In contrast, L1 elements are believed to be only vertically inherited, based on knowledge gained primarily on mammalian organisms (*14*). We hypothesise that the presence of L1s in therian mammals, and absence in monotremes, is due to an ancient HT event. In this study, we use BovBs as a comparison to identify common characteristics of horizontally transferred elements in contemporary eukaryotic species.

Three criteria are typically used to detect HT candidates: 1) a patchy distribution of the TE across species; 2) unusually high sequence similarity between divergent taxa; and 3) phylogenetic inconsistencies between the TE tree topology and species relationships (*15*). To

comprehensively test these criteria, we performed large-scale phylogenomic analyses of over 500 eukaryotic genomes (plants and animals) using iterative, query-driven searches of BovB and L1 sequences. Where possible, independently assembled closely related species were used in conjunction with multiple searching strategies to control for differences in genome assembly quality.

Our findings suggest that the horizontal transfer process has two parts: effective insertion of the TE, then expansion throughout the genome. Addressing the former, Fig. 1 shows that both BovB and L1 elements have a patchy distribution across eukaryotes. Both are absent from most arthropod genomes yet appear in relatively primitive species such as sea urchins and sea squirts. Furthermore, both TEs are present in a diverse array of species including mammals, reptiles, fish and amphibians. The main difference lies in the number of colonised species. BovBs are only present in 60 of the 503 species analysed, so it is easy to trace their horizontal transfer between the distinct clades (e.g. squamates, ruminants). In contrast, L1s encompass a total of 407 species, including both plants and animals, and they are ubiquitous across the well-studied therian mammals. Nonetheless, there are several species that exclusively contain BovB elements, with no L1s. The platypus and echidna genomes are a good example of this - especially since monotremes are known to contain other ancestral vertebrate retrotransposons, such as L2s and CR1s (*16*). There are only two possible explanations for the lack of L1s here: either L1s were expunged shortly after the monotreme-therian split but before they had a chance to accumulate; or monotremes never had L1s.

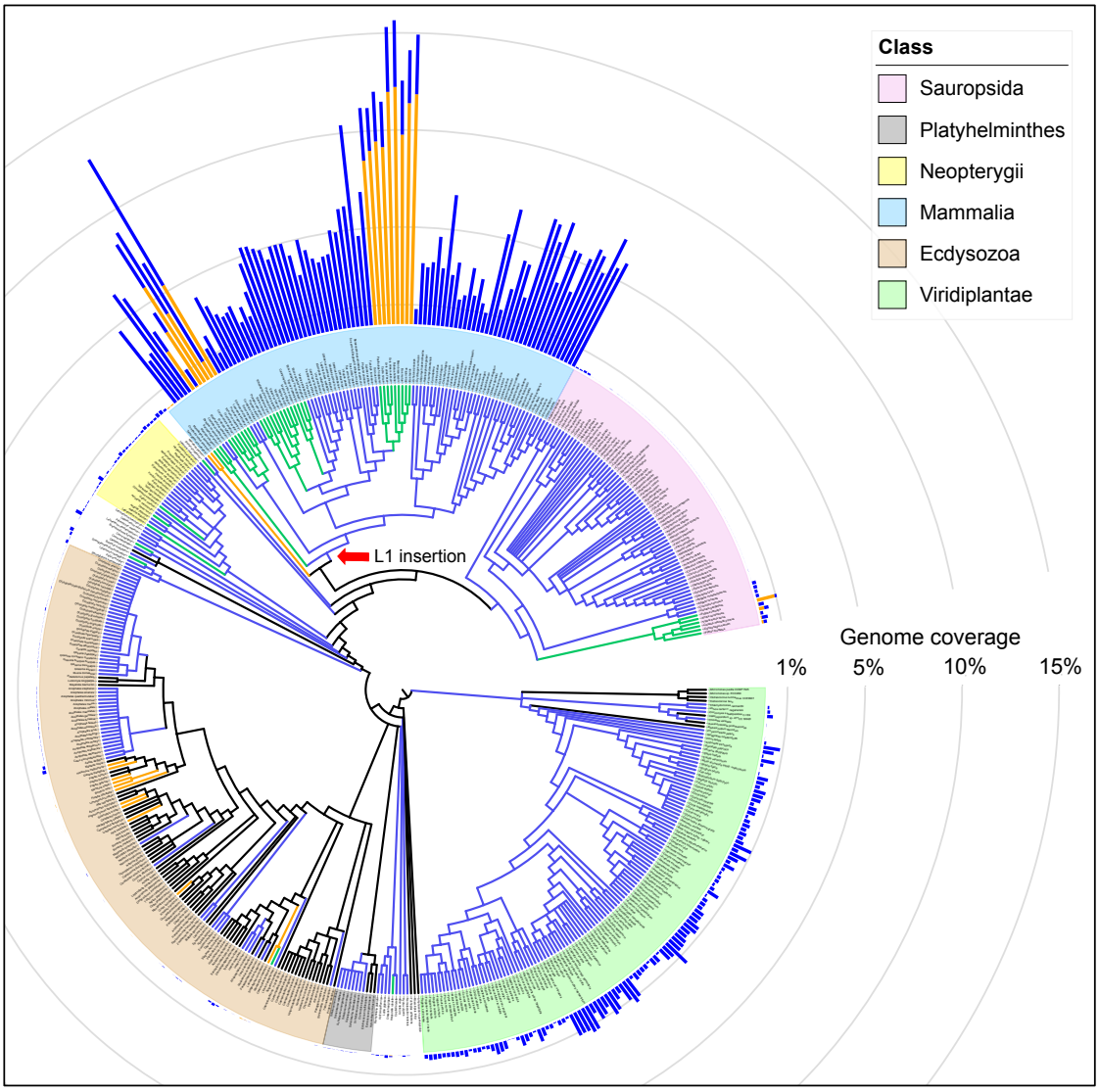


Figure 1: **Presence and coverage of L1 and BovB elements across eukaryotes.** The Tree of Life (*17*) was used to infer a tree of the 503 species used in this study; iTOL (*18*) was used to generate the bar graph and final graphic. The red arrow marks the putative L1 horizontal transfer event into therian mammals. Branches are coloured to indicate which species have BovB and L1 (green), only BovB (orange), only L1 (blue), or neither (black). Bar graph colours correspond to BovB (orange) and L1 (blue).

The abundance of TEs differs greatly between species. As shown in Fig. 1, mammalian genomes are incredibly susceptible to BovB and L1 expansion. More than 15% of the cow

genome is covered in these TEs (12% BovB, 3% L1) - and this is without considering the contribution of fragments (*19*). Even within mammals there are noticeable differences in copy number: for example, bats and equids have a very low number of full-length BovBs (<50 per genome), compared to the thousands found in ruminants and Afrotherian mammals. The low copy number here is TE-specific rather than species-specific: there are plenty of L1s in bats and equids. Hence, the rate of TE propagation is determined both by the host species (e.g. mammal versus non-mammal) and the type of retrotransposon (e.g. BovB versus L1).

To develop a method for identifying horizontal transfer events, we used BovB, a TE known to undergo HT. We clustered and aligned BovB sequences (both full-length nucleotide sequences and amino acid reverse-transcriptase domains) to generate a representative consensus for each species, and infer a phylogeny (Fig 2a shows the nucleotide-based tree). The phylogeny supports previous results (*10*), with the BovB tree topology noticeably different from the tree of life (Fig. 1). We were further able to expand single species into consistent clades, refining our estimates for the times of insertion. For example, the cluster of equids includes the white rhino, *Ceratotherium simum*. This suggests that BovBs were introduced into the most recent common ancestor before these species diverged. The low copy number in equids and rhino, observed in Fig. 1, is not because of a recent insertion event. The most likely explanation is that the donor BovB inserted into an ancestral genome, was briefly active, then lost its ability to retrotranspose and was inherited to descendent species.

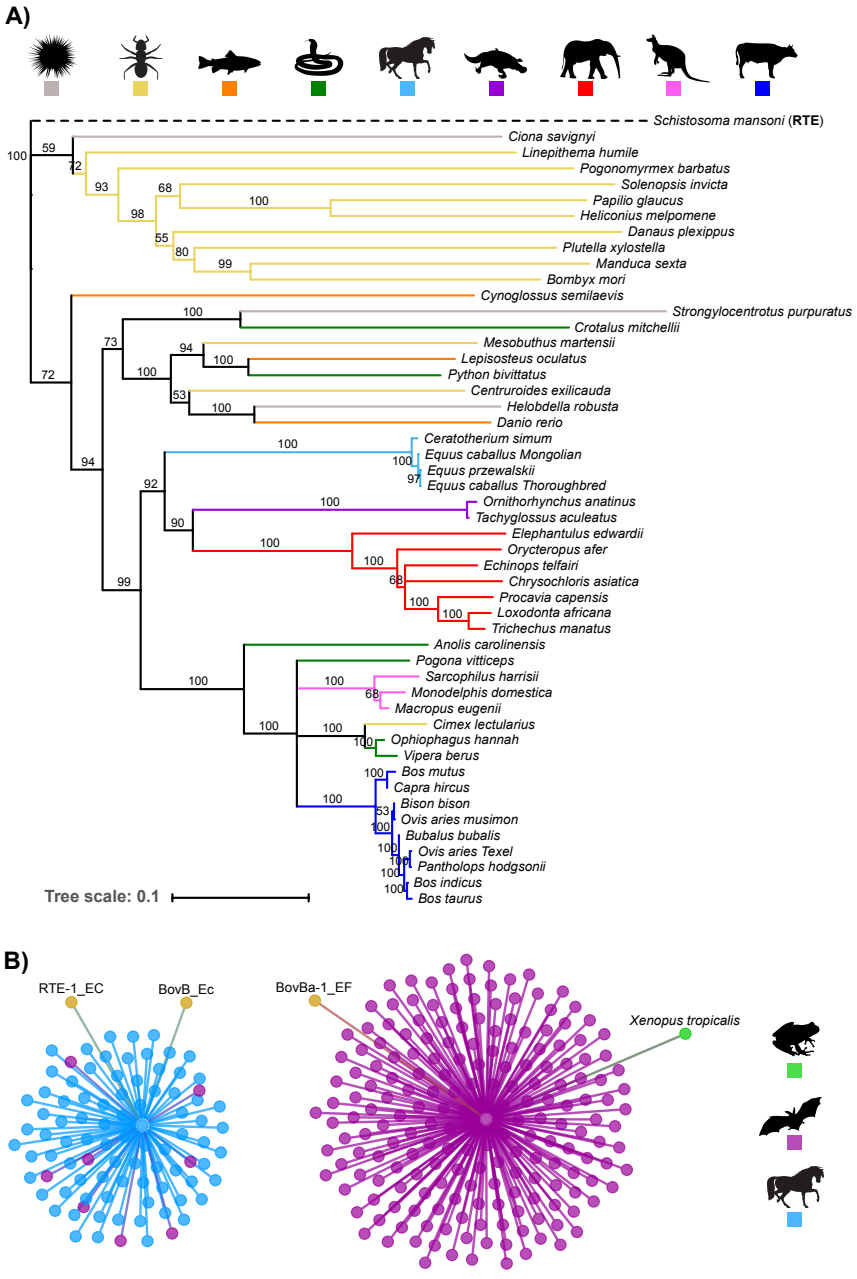


Figure 2: **HT of BovB retrotransposons.** (2a) Neighbour-joining tree (1000 bootstrap replicates) inferred using full-length nucleotide BovB consensus sequences, representing the dominant BovB family in each species. Nodes with confidence values over 50% are labelled and branches are coloured taxonomically. RTE sequence from *Schistosoma mansoni* was used as the outgroup. (2b) Network diagram representing the two distinct BovB clades in bats. Nodes are coloured taxonomically apart from the RepBase (*20*) sequences (light brown). RTE-1_EC and BovB_Ec are shown to belong to a single family, while BovBa-1_EF-like bat sequences form a separate family containing a single full-length BovB from the frog *Xenopus*.

The placement of arthropods is intriguing. Most of the arthropods (butterflies, moths and ants) appear as a monophyletic group at the top of the tree, sister to sea squirt *Ciona savignyi*. The presence of BovB in all these species, including *Ciona savignyi*, suggests that BovB TEs may have originated as a subclass of ancient RTEs. The next grouping consists of two scorpion species (*Mesobuthus martensii* and *Centruroides exilicauda*) nestled among the snakes, fish, sea urchin and leech - a possible vector. But the most interesting arthropod species is *Cimex lectularius*: the common bed bug, known to feed on animal blood. The full-length BovB sequence from *Cimex* shares over 80% identity to viper and cobra BovBs; their reverse transcriptase domains share over 90% identity at the amino acid level. Together, the bed bug and the leech provide support for the theory that blood-sucking parasites can transfer retrotransposons between the animals they feed on.

We were able to extend the BovB paradigm to include 10 species of bats and one frog (*Xenopus tropicalis*). The bats were not included on the phylogeny because their BovB sequences were too divergent to construct an accurate consensus. Instead, we clustered all of the individual BovB sequences to identify two distinct subfamilies: one containing all of the horse and rhino BovBs as well as eight bat sequences, and the other containing the remaining bat BovBs as well as the single BovB from *Xenopus*. We also included three annotated sequences from a public database (*20*) to resolve an apparent discrepancy between the naming of BovB/RTE elements. Our results (illustrated in Fig 2b) have several implications: first, bat BovBs can be separated into two completely distinct clades, suggesting the BovBs arose from independent insertion events; second, the BovBa-1-EF bat clade may have arisen from an amphibian species, or vice versa; and third, the naming conventions used in RepBase (*20*) need updating to better distinguish BovB versus RTE sequences. This third point is discussed in the Supplementary (e.g Supp. Fig. 1).

As a final test with BovB sequences, we replicated the all-against-all BLAST (*21*) approach used in El Baidouri *et al.* (*22*) to detect HT candidates across all species. Briefly, this compares all sequences in a multi-fasta database to generate BovB clusters or families. We identified 215 HT candidate families which contained BovBs belonging to at least two different eukaryotic species. Many of these were closely related species; so to find the HT families most likely to be true events, we restricted the analysis to families that crossed between different eukaryotic Orders (e.g. Afrotheria and Monotremata). We performed *in silico* validation for each candidate family: pairwise alignments of the flanking regions to rule out possible contamination or orthologous regions, and phylogenetic reconstructions to confirm discordant relationships. A total of 22 families passed all of the tests, suggesting at least 22 HT events. In many cases, the family groupings included one or two reptile BovBs, and numerous mammalian BovBs (see Supp. Table 6). This is important for determining the direction of transfer. BovBs are thought to have entered ruminants after squamates (*13*). The single reptile element in a family is therefore likely to be the donor sequence that instigated the transfer, supporting the theory that retrotransposons undergo HT to escape host suppression or elimination (*22*). Altogether, our results demonstrate that the horizontal transfer of BovB elements is even more widespread than previously reported, providing one of the most compelling examples of eukaryotic horizontal transfer to date.

L1 retrotransposons present a challenge because they are older; they have had more time to diverge, accumulate, and build a strong vertical background. Producing a consensus for each species was impractical because most species contained a divergent mixture of old, degraded L1s and young, intact L1s. Instead, we used the all-against-all clustering strategy on the collated dataset of L1 nucleotide sequences over 3kb in length (>1 million sequences total).

2815 clusters contained L1s from at least two different species: these were our HT candidates. As with BovBs, to reduce the dataset, we looked for families displaying cross-Order transfer. Most non-mammalian L1s (insects, reptiles, amphibians) had already been excluded because they resolutely grouped into species-specific clusters, even at low (50%) clustering identity. The majority of remaining families were from plants and mammals. After the validation tests, it was found that all of the mammalian candidate families were very small (e.g. one L1 element per species), and located in repeat-dense, orthologous regions in the genome typical of a vertical inheritance scenario (see Supp. Fig. 3). Thus, we found no evidence for continued transfer of L1s since their insertion into the therian mammal lineage.

Nevertheless, four plant families presented a strong case for L1 horizontal transfer (Fig 3a). The high sequence identity was restricted to the elements themselves, there were more than two L1 copies in each family, the sequences encoded open reading frames or had intact reverse-transcriptase domains, and the phylogenetic reconstructions showed evolutionary incongruence. The number of elements in each family mimicked the patterns seen with BovBs: very few elements from the 'donor species', and a noticeable expansion of L1s in the 'host species'. This demonstrates that transferred L1s have the ability to retain activity and expand within their new host. Moreover, it contradicts the belief that L1s are exclusively vertically inherited, and allows us to postulate that a similar event introduced L1s to mammals. At this stage, we do not know the vector of transfer since none of the analysed arthropods showed similarity to plant L1 sequences.

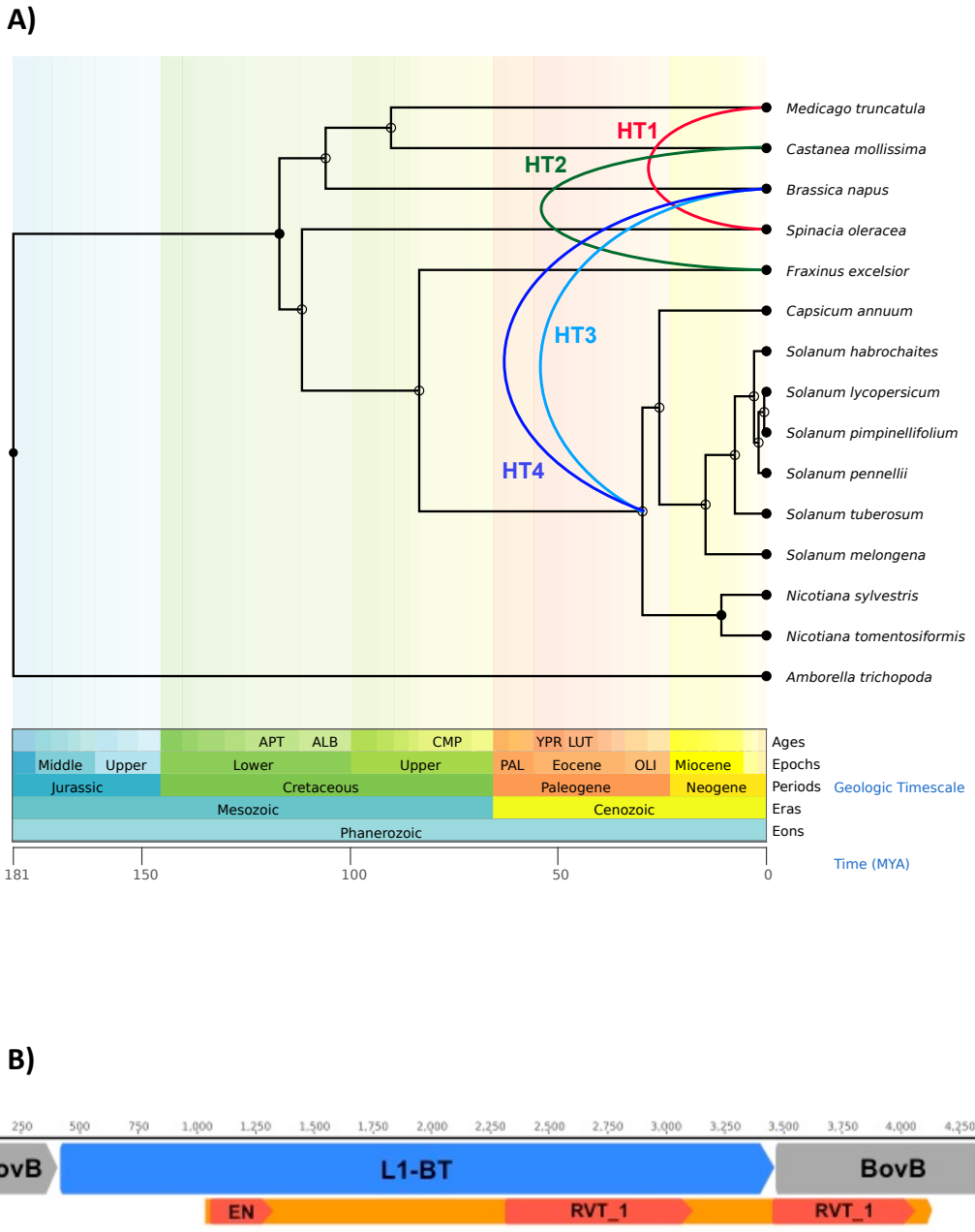


Figure 3: **HT of L1 in plants and newfound chimeric L1-BovB element.** (3a) TimeTree (*23*) illustrating the putative L1 horizontal transfer events between plant species. Shows only the species involved in HTs, and *Amborella trichopoda* as the outgroup. Background is coloured to match the ages in the geological timescale. (3b) Chimeric L1-BovB retrotransposon found in cattle genomes (*Bos taurus* and *Bos indicus*). L1-BT and BovB correspond to RepBase names (*20*), representing repeats which are known to have been recently active. RVT_1 = reverse-transcriptase, EN = endonuclease domain. The orange bar is the length of the entire open reading frame.

During our mining of candidate L1 HT families, we inadvertently discovered a chimeric L1-BovB element present in cattle genomes (*Bos taurus* and *Bos indicus*), shown in Fig. 3b. This particular element most likely arose from a recently active L1 element (98% identical to the canonical *Bos* L1-BT) inserting into an active BovB (97% identical to *Bos* BovB). In fact, L1s and BovBs have accumulated to such extents in these two genomes that they have created the ideal environment for chimeric repetitive elements. With two reverse-transcriptase domains and high similarity to currently active L1/BovB elements, this chimeric element has the potential to still be functional - presenting the possibility for L1 elements to be horizontally transferred throughout mammals by being encapsulated within BovBs.

In summary, the studies reported here suggest that all retrotransposons have the ability to undergo horizontal transfer, albeit at different rates. We extracted millions of retrotransposon sequences from a 503-genome dataset, demonstrating that both BovBs and L1s are sporadically distributed across eukaryotes. We further extended the analysis of BovBs to include blood-sucking arthropod vectors capable of infecting mammals and squamates, as well as two distinct bat clades and the first report of BovB in an amphibian. Contrary to the belief of exclusive vertical inheritance, our results with L1s reveal multiple ancient HT events in plants. However, we were unable to find evidence of recent L1 horizontal transfer, and our analyses indicate that the occurrence of such events is rare (i.e. using over 1 million L1 sequences from 407 eukaryotes, we identified only 4 families showing evidence of HT). Finally, our discovery of a potentially active L1-BovB fusion element in cattle presents the possibility of increasing this rate of transfer, particularly in mammalian species.

References

1. E. T. L. Prak, H. H. Kazazian, *Nature Reviews Genetics* **1**, 134 (2000).

2. O. Piskurek, D. J. Jackson, *Genes* **3**, 409 (2012).

3. C. Feschotte, *Nat Rev Genet* **9**, 397 (2008).

4. J. M. Chen, P. D. Stenson, D. N. Cooper, C. Ferec, *Hum Genet* **117**, 411 (2005).

5. K. Kaer, M. Speek, *Gene* **518**, 231 (2013).

6. J. Kazazian, H. H., *Curr Opin Genet Dev* **8**, 343 (1998).

7. J. Kazazian, H. H., *Science* **289**, 1152 (2000).

8. D. Kordis, F. Gubensek, *Gene* **238**, 171 (1999).

9. A. M. Ivancevic, A. M. Walsh, R. D. Kortschak, D. L. Adelson, *Bioessays* **35**, 1071 (2013).

10. A. M. Walsh, R. D. Kortschak, M. G. Gardner, T. Bertozzi, D. L. Adelson, *Proc Natl Acad Sci U S A* **110**, 1012 (2013).

11. I. Sormacheva, *et al.*, *Molecular Biology and Evolution* **29**, 3685 (2012).

12. A. Suh, *et al.*, *Nature Communications* **7** (2016).

13. D. Kordis, F. Gubensek, *Genetica* **107**, 121 (1999).

14. P. D. Waters, G. Dobigny, P. J. Waddell, T. J. Robinson, *PLoS One* **2**, e158 (2007).

15. S. Schaack, C. Gilbert, C. Feschotte, *Trends Ecol Evol* **25**, 537 (2010).

16. W. C. Warren, *et al.*, *Nature* **455**, 256 (2008).

17. D. Maddison, K. Schulz, The tree of life web project (2007).

18. I. Letunic, P. Bork, *Nucleic Acids Research* **44**, W242 (2016).

19. D. L. Adelson, J. M. Raison, R. C. Edgar, *Proc Natl Acad Sci U S A* **106**, 12855 (2009).

20. J. Jurka, *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).

21. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403 (1990).

22. M. El Baidouri, *et al.*, *Genome Res* **24**, 831 (2014).

23. S. Kumar, S. B. Hedges, *Bioinformatics* **27**, 2023 (2011).

Chapter 4

Using Repetitive Elements to Infer Species Relationships from Ancient DNA

"I do not fear computers. I fear the lack of them."
— Isaac Asimov

A significant part of this project has been the development and optimisation of bioinformatics methods for interpreting large datasets. So far, we have explored the dynamics of transposable elements within genomes and their movement across species. This research has primarily focused on young, recently active elements - since these are the most likely to undergo horizontal transfer and cause genomic changes. However, the vast majority of repetitive sequences in eukaryotic genomes are ancient remnants that have long been inactive. Such repeats are informative because they contain information about the species ancestry and evolution. The following manuscript details a novel approach for inferring phylogenetic relationships by using repetitive intervals as binary genetic markers. The approach is tested on a dataset of 21 elephants, including both modern and ancient specimens. The manuscript has been prepared for submission to *Genome Biology and Evolution*, formatted according to the guidelines of a Genome Resources article.

Statement of Authorship

Title of Paper	Using repetitive elements to infer species relationships from ancient DNA
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Atma M. Ivancevic, Thu-Hien To, R. Daniel Kortschak, Joy M. Raison, Zhipeng Qu, Tat-Jun Chin, David L. Adelson (2016). Using repetitive elements to infer species relationships from ancient DNA. Prepared for submission as a Genome Resources article to Genome Biology and Evolution.

Principal Author

Name of Principal Author (Candidate)	Atma M. Ivancevic
Contribution to the Paper	Performed the analysis together with Thu-Hien To, interpreted the results and wrote the manuscript.
Overall percentage (%)	60
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 5/12/16

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	Thu-Hien To
Contribution to the Paper	Performed analysis, interpreted the results and assisted in writing the manuscript.
Signature	Date 14 Dec 2016

Name of Co-Author	R. Daniel Kortschak
Contribution to the Paper	Designed the de novo repeat identification method, supervised development of the work and assisted in writing the manuscript.
Signature	Date 6/12/16

Name of Co-Author	Joy M. Raison		
Contribution to the Paper	Performed repeat screening and annotation and assisted in interpreting the results.		
Signature		Date	6/12/16

Name of Co-Author	Zhipeng Qu		
Contribution to the Paper	Assisted in performing analyses and interpreting the results.		
Signature		Date	7/12/16

Name of Co-Author	Tat-Jun Chin		
Contribution to the Paper	Helped design algortithms for repeat clustering.		
Signature		Date	5/12/2016

Name of Co-Author	David L. Adelson		
Contribution to the Paper	Supervised development of the work and assisted in analysing the results and writing the manuscript.		
Signature		Date	7/Dec/2016

Using repetitive elements to infer species relationships from ancient DNA

Atma M. Ivancevic,[1] Thu-Hien To,[1] R. Daniel Kortschak,[1]
Joy M. Raison,[1] Zhipeng Qu,[1] Tat-Jun Chin,[2] David L. Adelson[1*]

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

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

Single nucleotide polymorphisms (SNPs) are often used as genetic markers to identify differences between species and individuals. In this study, we present an alternative approach utilising the 'dark matter' of the genome: repetitive elements. Huge portions of eukaryotic genomes are made up of non-coding, ancient repeats which can be used to describe ancestral species relationships. Our model is simple: use the repetitive intervals in each genome to identify binary variance (presence versus absence) and thus infer phylogenetic relationships. Using a test dataset of 21 modern and ancient elephants, we compare our results to the previously established SNP phylogeny and discuss the advantages and limitations of a repeat-based approach.

Introduction

Before the advent of genome sequencing technologies, scientists devoted their attention to protein-coding genes and proteins. Since then, over a hundred mammalian genomes have been sequenced - revealing the prevalence of non-coding, repeat-derived sequences. The vast majority of repeats are remnants of insertion events that occurred millions of years ago.

This provides a genetic footprint of evolutionary relationships between different species and individuals.

The most frequently used method for resolving genome differences is by looking at single nucleotide polymorphisms (SNPs) (*1*). However, SNP-based approaches are not always robust because they rely on single alleles. Similarly, human genome projects sometimes use transposable element polymorphisms (polyTEs) from recent insertions to infer ancestry (*2*). This method was primarily designed for use between individuals of the same species - it does not take into account ancient repeats. We propose an alternative approach: using a reference genome, identify *all* of the repetitive intervals and compare them between genomes to find binary variance (presence versus absence). Variant intervals are then used to infer a phylogeny of the species relationships. Due to its binary nature, our method executes quickly and can be used on any dataset of genomic sequences, including those with no known SNP variants.

To test our approach, we used a set of 21 ancient and modern elephants from diverse geographic backgrounds (provided by collaboration with David Reich's group). An evolutionary phylogeny of this dataset has already been inferred using SNP data (see Supp. Fig. 1). Using the publicly available chromosome-level assembly of *Loxodonta africana* as a reference, we characterised the interspersed repeats in the genome, mapped them onto the genomic data of the 20 other elephants and compared our results to the SNP-based phylogeny. Various criteria were tested to determine the optimal parameters, particularly when dealing with ancient DNA.

Materials and Methods

Reference genome: repeat identification, annotation and analysis

Ab initio repeat identification and annotation

Loxodonta africana (KB13542, chromosome-level assembly LA4v2) was used as the reference genome. Entire LA4v2 chromosome sequences were locally aligned with the krishna program (http://godoc.org/code.google.com/p/biogo.examples/krishna) (*3, 4*) using the default parameters. Alignments identified by krishna were clustered by the igor program (http://godoc.org/code.google.com/p/biogo.examples/igor) (*3, 4*) using default parameters, except for the "-overlap-strictness=1" parameter to prevent inclusion of overlapping features. The minimum accepted cluster size was two members. Sequences corresponding to members of alignment clusters were extracted from the LA4v2 sequence and aligned using MUSCLE (*5, 6*) with default parameters; only members within 95% of the length of the longest member were aligned and when clusters contained more than 100 members, 100 randomly chosen sequences satisfying the length constraint were included in the alignment. A consensus for each cluster was constructed from its MUSCLE alignment and subsequently used in the repeat annotation process.

CENSOR (*7*) was used to annotate identifiable repeats. WU-BLAST (*8*) was further used with a comprehensive retroviral and retrotransposon protein database assembled from the National Center for Biotechnology Information (*9*) to further annotate repeats, and with 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.

LTR class identification

Initially, the LA4v2 genome assembly was analysed using CENSOR (*7*) with the long terminal repeat (LTR) records from RepBase version 20140131 (*10*). The output from CENSOR was then run through censormerge; briefly, this program merges adjacent repeat features with matching annotation allowing for limited insertions by different elements or backtracking within the annotating repeat sequence, or until a significant fraction of the annotating repeat sequence has been covered. Merged sequences identified by censormerge were tested for the presence of GAG- or POL-encoding sequence by BLAST (*11*) alignment against databases containing representatives of these protein sequences, using an e-value threshold of 1e-12.

Repeat analysis

The genome was divided into 1.5 Mb contiguous bins and the number of each of the features within each bin was counted. Genomic features analysed include: interspersed repeat groups, obtained from running CENSOR (*7*) with the mammal library and our elephant repeat library and grouping based on the repeat sequence classification, genes, CpG islands and G4s. The count data was transformed by first dividing by the number of known base pairs (A,C,T,G) in the bin and then taking the square root. For each bin, the CpG coverage was calculated and an arcsine transformation taken, and the GC content was calculated. Bins that had less than 500,000 known base pairs (bp) were excluded from the analysis. For each feature, outlier bins within the genome were identified using a 2-tail t-test at the 5% significance level.

To identify ancient and recent regions of the genome, a principal components analysis was performed on the transformed bin data. The principal component with high weights for the ancient repeats MIR, L2 and CR1 was selected as the indicator for 'ancientness'. For each bin in the analysis, the average value of the principal component (in a 5-bin window centered on the bin) was calculated. If this value was significantly greater than zero the bin was classified

as ancient; if it was significantly less than zero the bin was classified as recent. Significance tests were based on 2-tail t-tests at the 5% significance level. Window variance was used unless there was only one bin in the window; in that case, the variance of the principal component was used.

Extraction of repetitive intervals from elephant genomic data

All interspersed repeats identified and annotated in the reference assembly were used to produce BED intervals (>50 bp) for each major repeat class (DNA, ERV, LINE, SINE; total of 4,353,898 interspersed repeat intervals). The BED intervals were used to extract BAM slices from the provided whole genome data of 21 elephants (2 *Loxodonta cyclotis*, 2 *Loxodonta africana*, 6 *Elephas maximus*, 5 *Mammuthus primigenius*, 2 *Mammuthus columbi*, 2 *Mammut americanum* and 2 *Elephas antiquus*). Extraction was performed using BEDTools (*12*) with default parameters for intersecting a BAM alignment with features in BED format, irrespective of strand.

Phylogenetic inference of species relationships using variant sites

The intersected BAM slices were converted to BED format, sorted with respect to chromosome name and start/end position, and merged to form a set of unique, non-redundant repeat intervals for each genome. Intervals were then transformed into a binary system for each individual, where '1' indicated presence of that interval and '0' indicated absence (see Supp. Fig. 2). Each elephant had a distinct signature of 1's and 0's to compare to the reference. The binary signatures were compared across all elephants to find persistent sites (i.e. present in all taxa) and variant sites (absent from some taxa). Only the variant sites were used to infer a phylogeny.

RAxML (*13*) and PAUP (*14*) were used to infer maximum likelihood (ML) and maximum parsimony (MP) phylogenies from the binary sequences. The RAxML model used was

BINCAT: a memory and time efficient approximation for the standard GAMMA model of rate heterogeneity from binary data. Bootstrapping was used to estimate confidence values.

Parameter testing for quality control

The aim of this project was to use a repeat-based approach to infer phylogenetic relationships. To determine the optimal parameters for an accurate phylogeny, we needed to take into account external factors that differ between genomes (e.g. level of coverage, or ancient DNA versus modern DNA). Several criteria were tested for defining presence/absence of an interval: (1) presence is indicated by at least 1 bp in the expected interval; (2) setting a minimum length of 20 bp for each interval (such that any intervals containing $<$20 bp are considered absent, not present); (3) as per (1) but only including taxa with $>$5x coverage; (4) as per (2) but only including taxa with $>$5x coverage.

Other tests included a triplets analysis for incomplete lineage sorting (*15*): in brief, this test performs groupings by counting the number of intervals present in two species and absent in the third. Every possible combination of three elephants was inspected to find the most likely grouping, and determine whether this grouping was due to repeat content or data quality.

We also considered setting other genomes as the reference (e.g. *Mammuthus columbi*_U or *Mammut americanum*_I), although this was limited by the fact that *Loxodonta africana*_C was the original reference, so sites specific to ancient elephants could not be observed.

Similarly, we tried to minimise bias due to coverage by imputing common intervals from high coverage elephants to low coverage elephants.

Finally, the absent intervals in each elephant were categorised by repeat class (e.g. DNA, ERV, LINE, SINE), to see if there was an under-representation of some repeats in certain species compared to others.

Results

Repeat coverage and ancient regions in *Loxodonta africana*

The total repeat coverage of the reference elephant genome was found to be about 50% (Supp. Table 1), which is comparable to other mammalian genomes. However, the non-LTR retrotransposon fraction of the genome is significantly higher in the elephant compared to other placental mammals. Non-LTR over-representation may be attributable to the presence of LINE retrotransposons, which are horizontally transferred in higher organisms (*16, 17*).

Our identification of Ancient Genomic Regions (AGR) through principal component analysis indicates that AGR exist in the elephant genome as they do in the bovine genome (*18*). AGR seldom contain recent, clade-specific repeats (Fig. 1, Supp. Fig. 4). In contrast, regions of low Ancient Repeat density tend to contain many recent, clade-specific repeats.

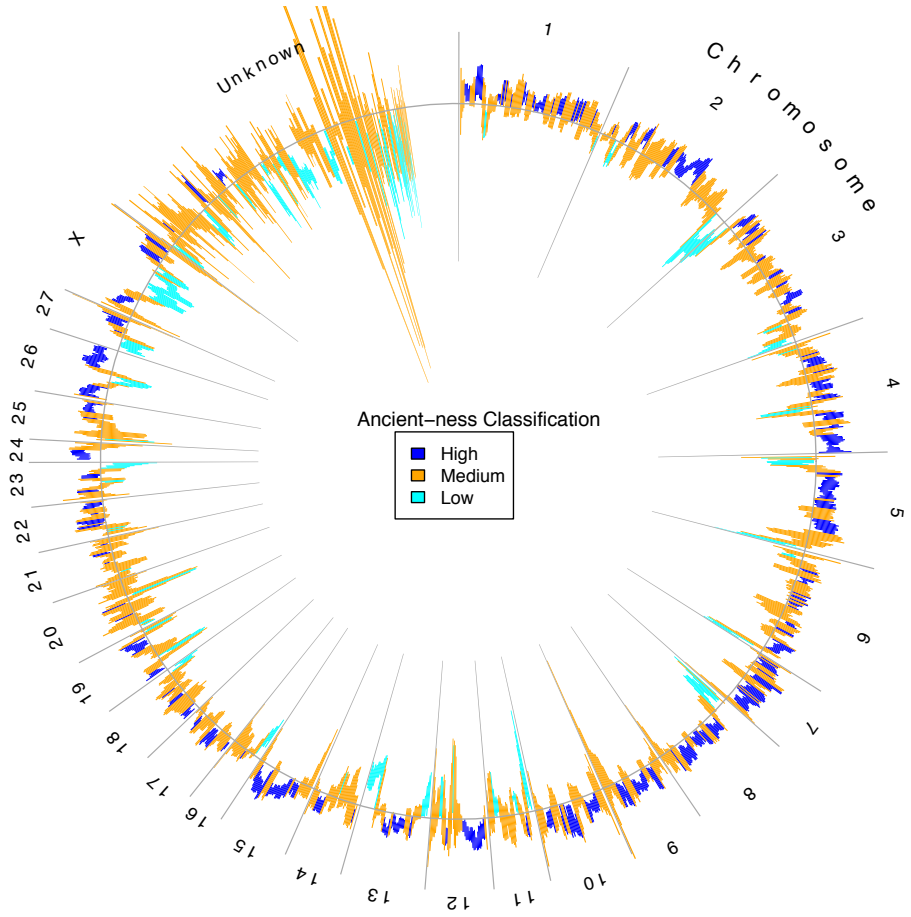


Figure 1: **Ancient-ness classification for *Loxodonta africana*:** PCA analysis of ancient repeat regions in the reference genome (LA4v2).

Initial subset of full-length LINEs as genomic markers

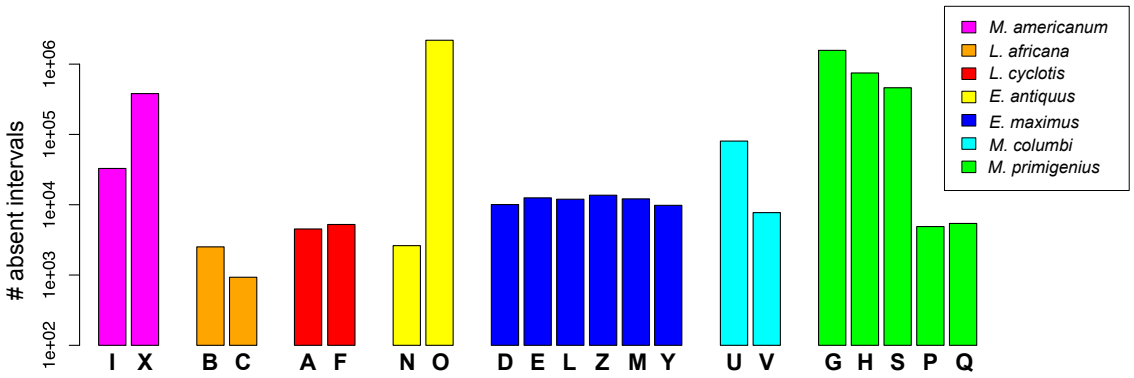
As a preliminary test, we used a subset of full-length BovB and LINE-1 repeat intervals (4929 BovB, 10697 L1) to identify variant sites between the elephants. The subset was too small for reliable phylogenetic inference because the binary sequence analysis reduced the dataset to a mere 18 variant sites (Supp. Fig. 5). This suggests that full-length retrotransposons, particularly active ones, tend to persist in elephants.

Full dataset of interspersed repeats

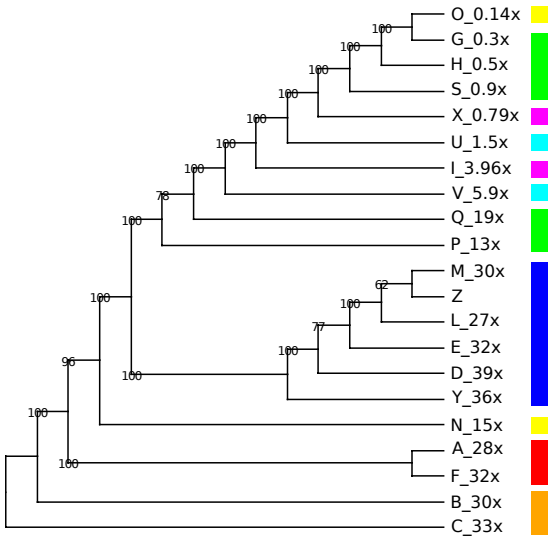
Using the entire interspersed repeats collection was more successful, variant sites increased from 18 to over 3 million. Many possible phylogenies were produced depending on the filtering parameters. To simplify discussion, we will use specific test cases to demonstrate differences due to coverage, minimum interval length, and repeat class.

Trivial case: all 21 elephants, no filtering

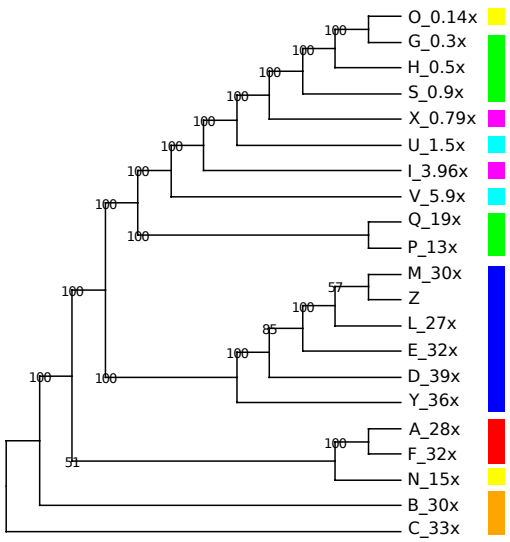
The trivial case involved no filtering: all 21 elephants were used, despite some having very low coverage and only 1 bp was needed to classify an interval as 'present' (as shown in Supp. Fig. 2). This produced 3,037,698 variant sites. Elephants with low coverage (e.g. *M. americanum*_X, *E. antiquus*_O, *M. columbi*_U, *M. primigenius*_G,_H,_S) stand out as having a huge number of absent intervals (Fig. 2a), and are thus grouped together in the inferred phylogenies (Fig. 2b, 2c). Ignoring these, most of the other elephants uphold previous species relationships. The main difference between the ML (Fig. 2b) and MP (Fig. 2c) phylogenies is the placement of *E. antiquus*_N. The MP tree agrees with the original phylogeny (Supp. Fig. 1), with *E. antiquus*_N sister to the *L. cyclotis* group, while the ML tree places it closer to *E. maximus* elephants. Both ML and MP trees show *E. maximus*_Y as an outgroup to the other *E. maximus* elephants; a placement not seen in the original tree.



(a) Absent intervals



(b) Maximum likelihood tree



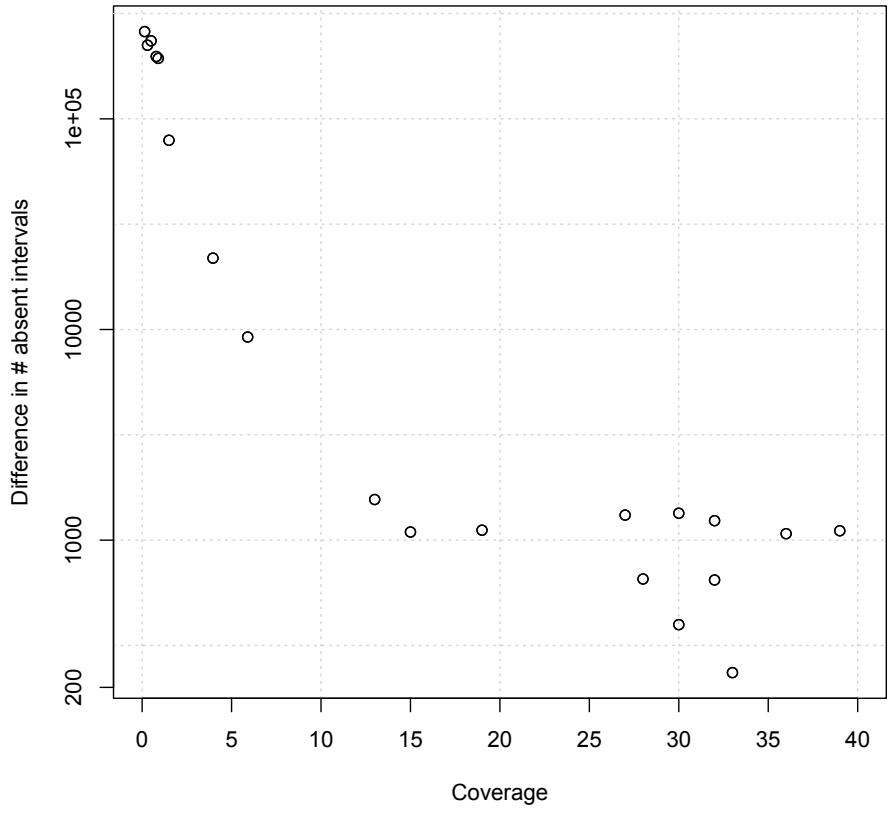
(c) Maximum parsimony tree

Figure 2: **All elephants, no filtering:** (a) shows the number of absent intervals in each elephant, coloured by species and labelled with the appropriate initial. Note that the y-axis is a log scale. (b) and (c) show the inferred phylogenies from these absent intervals. The trees are rooted with *L. africana*_C (reference genome). Coloured bars represent different species, using the legend from (a). Bootstrap support values are shown; branch lengths are not shown because the low coverage species were too long.

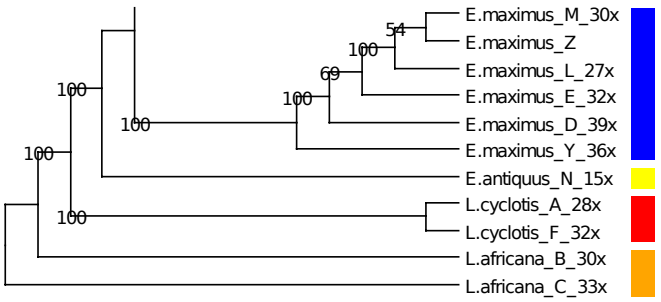
All elephants, minimum length of 20 bp per interval

Setting a minimum threshold of 20 bp for each interval increased the total number of variant sites to 3,235,889. The difference was largely attributed to the low coverage species, which showed a huge increase in the number of absent intervals (Fig. 3a). Consequently, the topology did not change at all for the low coverage elephants, they still grouped together in order of lowest coverage.

However, there was one difference in the high coverage elephants. Previously (Fig. 2), the ML and MP trees differed in their placement of *E. antiquus*_N. With a 20 bp minimum, the ML and MP trees now agree that *E. antiquus*_N should be distinct from the *Loxodonta* elephants (Fig. 3b). This is the only test case where the ML and MP trees concur.



(a) Change in absent intervals is due to low coverage



(b) Inferred tree topology

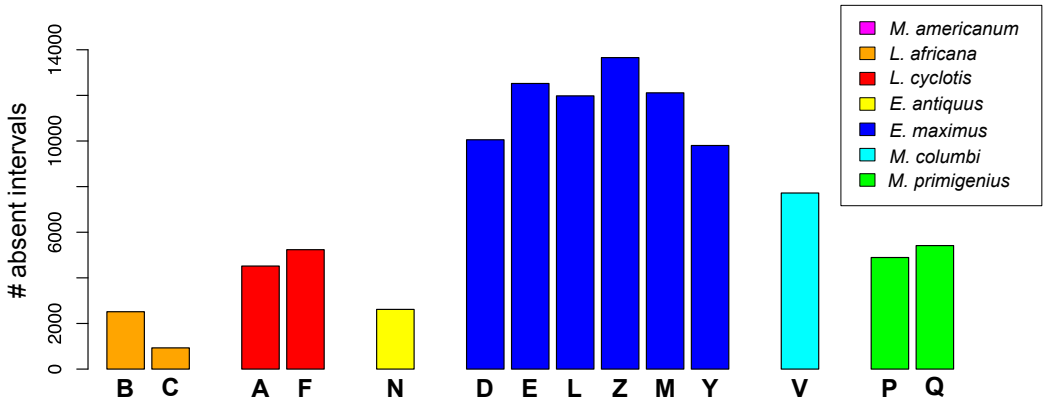
Figure 3: **All elephants, minimum length of 20 bp per interval:** (a) shows that increasing the minimum interval length from 1 bp to 20 bp drastically affects the low coverage species. More intervals in these elephants are classified as 'absent'. (b) shows the inferred topology seen using both a maximum likelihood and maximum parsimony approach. Low coverage species are not shown because they grouped by lowest coverage instead of repeat content, as seen previously in Fig. 2.

All elephants, triplets test

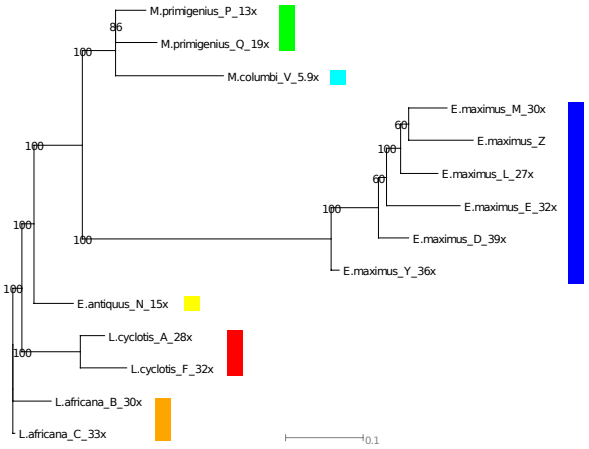
The triplets test was inconclusive. It confirmed the obvious topologies seen in the high coverage, modern elephants (e.g *L. africana* is more closely related to *L. cyclotis* than *E. maximus*). It was also useful for resolving the placement of *E. antiquus*_N as sister to *L. cyclotis* elephants, supporting the SNP-based phylogeny (Supp. Fig. 1). However, it could not sensibly resolve the low coverage elephants. This, along with the previous tests, suggested that the only way to infer a high-confidence phylogeny would be to exclude low coverage species.

High coverage elephants only

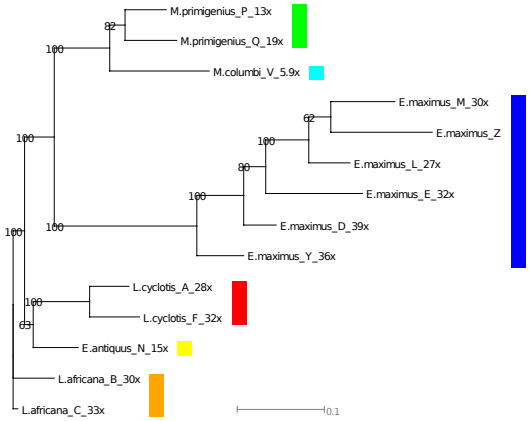
If we remove genomes with <5x coverage and do not set a minimum interval length, 34,175 variant sites remain. It is clear that *E. maximus* elephants are missing the largest number of repeat intervals (Fig. 4a). *E. maximus* are a modern species, with abundant genomic data available, so the absence of repeat intervals is not due to poor coverage or mapping quality. The trees (Fig. 4b, 4c) mirror that seen previously in Fig. 2b, 2c. The *E. maximus* subgroup is markedly distinct from the other elephants, with *E. maximus*_Y acting as the species outgroup. *E. antiquus*_N is separate from *L. africana* and *L. cyclotis* in the ML tree, but clustered with *L. cyclotis* in the MP tree. *M.columbi*_V is always nestled amongst the woolly mammoths. Using only high coverage elephants increased the support values and produced more reliable phylogenies.



(a) Absent intervals



(b) Maximum likelihood tree



(c) Maximum parsimony tree

Figure 4: **High coverage elephants only:** (a) shows the number of absent intervals in each elephant, coloured by species and labelled by initial. (b) and (c) show the inferred phylogenies from these absent intervals. The trees are rooted with *L. africana*_C (reference genome). Coloured bars represent different species, using the legend from (a). Bootstrap support values and branch lengths are shown.

High coverage elephants only, separated by repeat class

Next, we separated the absent intervals by repeat class to look for under-represented repeats between species. We wanted to know if the missing intervals in *E. maximus* elephants belonged to a certain repeat group. Fig. 5 shows the breakdown of 4 major repeat classes: DNA (e.g.

DNA transposons such as mariner elements); ERV (including LTRs); LINE (BovBs, LINE-1s, etc); and SINE (7SL, tRNA, 5S, etc), as categorised by CENSOR.

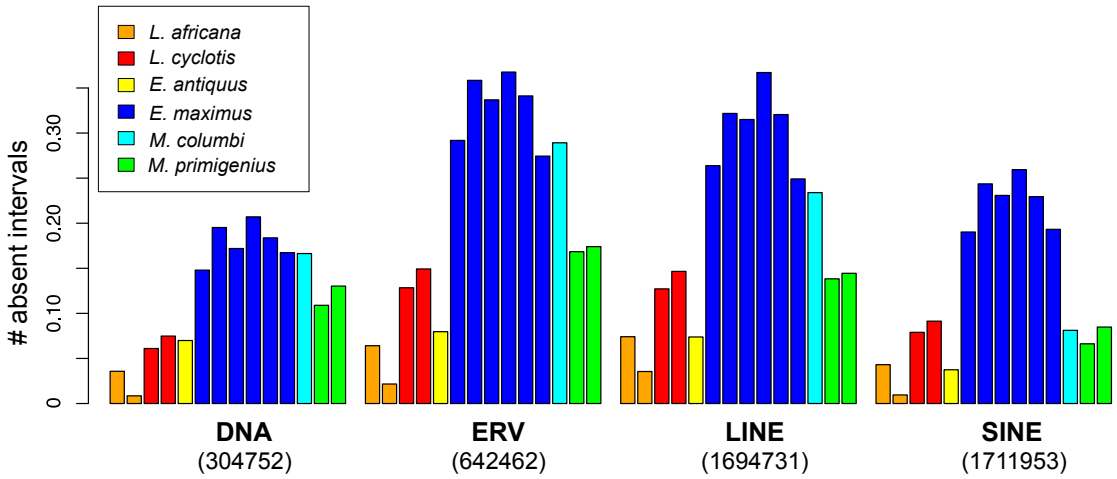


Figure 5: **Variance between high coverage elephants, categorised by repeat class:** The x-axis shows the repeat class (e.g. ERV) and total number of intervals belonging to that repeat class (e.g. 642462). The y-axis shows the percentage of absent intervals (i.e. number of absent ERV intervals/total number of ERV intervals x 100). Elephants are coloured by species and appear in the same order as seen previously (Fig. 4a).

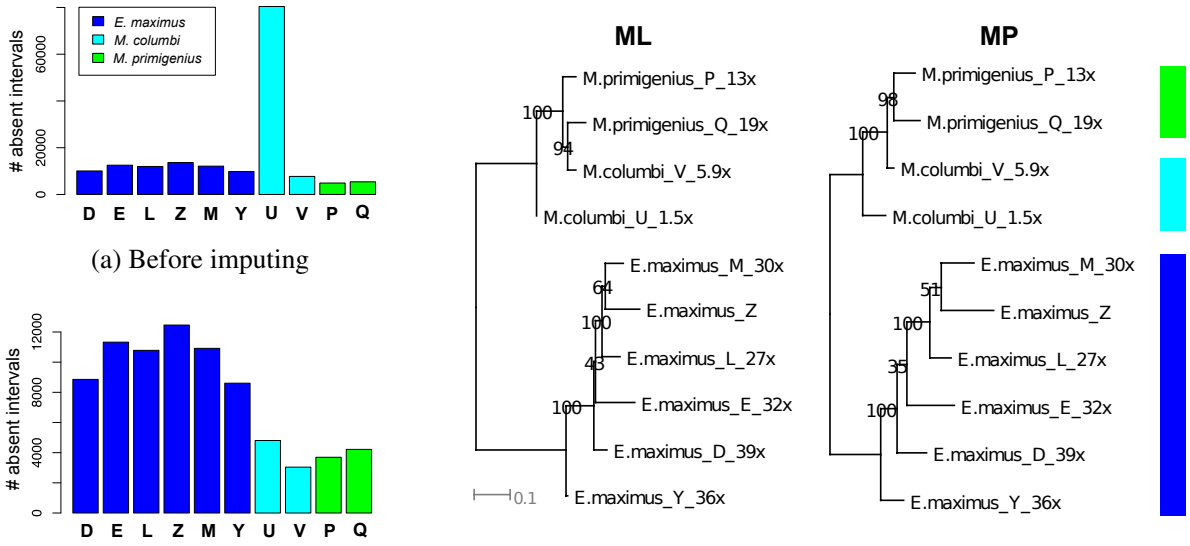
There does not appear to be a specific bias towards any category of repeats. *E. maximus* elephants consistently have the largest proportion of absent intervals. In particular, SINE repeats are very under-represented. Each member of the *E. maximus* group is missing more than twice as many SINE sites as any other elephant.

Imputing intervals from high coverage elephants to low coverage elephants

Repeatedly, we observed bias due to low coverage. High-confidence phylogenies could only be produced by excluding low coverage elephants. This does not help us resolve the topology of the mammoths.

As a final test, we tried imputing common intervals from the high coverage elephants to the low coverage elephants. Consider the *E. maximus*/*Mammuthus* clade. *M. columbi*_U has by far the most absent intervals in this clade, due to the 1.5x coverage (Fig. 6a). In contrast, the

high coverage (>5x) elephants share 4,328,440 common intervals. If we assign these common intervals to the low coverage elephants (*M. columbi*_U and *M. columbi*_V), we can obtain a more realistic barplot (Fig. 6b), with the total number of variant sites being 24,261. The corresponding trees differ slightly in their placement of *M. columbi*_V, but largely support the SNP-based phylogeny (Supp. Fig. 1).



(a) Before imputing

(b) After imputing

(c) Inferred trees

Figure 6: **Resolving the mammoths:** (a) shows the number of absent intervals in each elephant, coloured by species as before. (b) shows the number of absent intervals after imputing common intervals on the two low coverage elephants. (c) shows the maximum likelihood (left) and maximum parsimony (right) trees generated with RAxML and PAUP.

Recurring topology

The tests detailed above indicate that the ideal parameters on this data are to exclude low coverage species. Changing the minimum interval cutoff from 1 bp to 20 bp only resulted in minor differences. Likewise, there is no particular bias to any repeat class, but using all interspersed repeat intervals is the most effective.

Fig. 7 shows the recurring patterns seen throughout all of the tests. The low coverage

elephant species could not be confidently resolved because they constantly grouped by lowest coverage instead of repeat content (as shown by the initial phylogenies and triplets test). With so many absent intervals, their binary signature became too different to be compared to the other elephants.

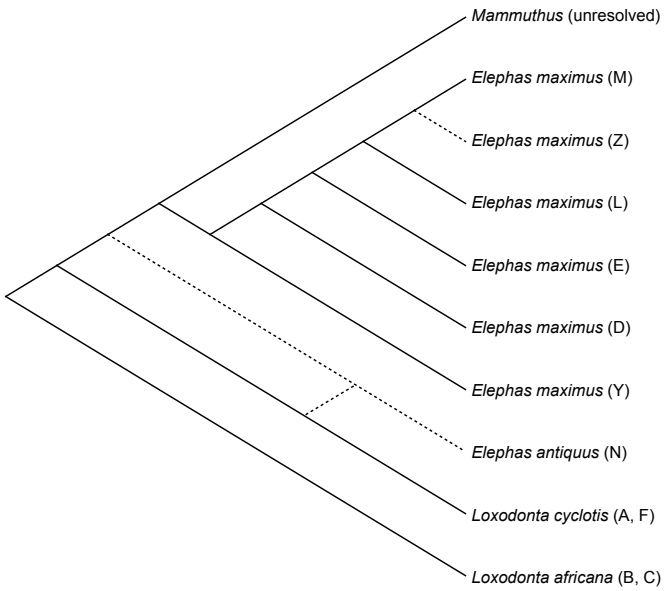


Figure 7: **Recurring tree topology:** *L. africana*_C was used as the reference genome, so differences between B and C are likely due to individual polymorphism. *L. cyclotis*_A and F are always seen grouped together. The position of *E. antiquus*_N changes depending on the approach used (within the *Loxodonta* elephants with a maximum parsimony approach; closer to the *E. maximus* elephants with a maximum likelihood approach). *E. maximus*_Y is always outgroup to the other *E. maximus* elephants (this was even true for the initial subset of full-length LINEs, which only contained 18 variant sites). *E. maximus*_Z is most often seen as a sibling to M, but with low support. The *Mammuthus* genus is unresolved in terms of inter-species relationships due to low coverage.

Discussion

Do elephant genomes only carry dead LINEs

The initial test using full-length BovB and LINE-1 retrotransposons failed to produce a reliable phylogeny. However, the lack of differences is an interesting finding. It means that even extinct elephants have all of the full-length repeats. Considering we used a modern elephant

(*Loxodonta africana*) as the reference, the most likely explanation is that there has not been any recent retrotransposition in the elephant lineage (if there had been, we would expect the modern elephants to have more full-length intervals than the ancient ones). There are a number of L1s in the *L. africana* genome that appear active, based on their structural characteristics. If they are not truly active, then they have persisted as conserved artefacts throughout the entire elephant lineage.

Distinct differences between Asian and African elephants

Nuclear and mitochondrial DNA analyses have shown that Asian elephants are the closest living relatives of mammoths (*19*). The repeat-based phylogenies generated above provide supporting evidence. The most striking differences in repeat content occur in the modern Asian elephants (*E. maximus*), which stand out as having a huge number of absent intervals compared to the other elephants. The SINE repeat class is particularly lacking (Fig. 5): the *E. maximus* subgroup are missing more than twice as many intervals as any other elephant (including mammoths).

Does this mean that Asian elephants are less repeat-dense? Or are their repeats found in different locations to the African elephants (and are thus undetectable with an African elephant reference)? In order to resolve this, we would need to use a modern Asian elephant as the reference and map those repeat intervals back against the other elephants.

Changes from the original phylogeny revolve around *Elephas* elephants (*maximus* and *antiquus*)

Our repeat-based model (results summarised in Fig. 7) consistently produced two deviations from the SNP phylogeny (Supp. Fig. 1). Firstly, the Indian elephant *Elephas maximus*_Y always appears as an outgroup to the rest of the *E. maximus* clade. The original tree has a support value of 63/60 at this node, suggesting that it may be misplaced. In an attempt to resolve this, we added a second Indian elephant *Elephas maximus*_Z, which groups with M and L. However,

*Elephas maximus*_Y remained resolutely distinct, indicating that there may be repeat-specific differences in this genome which distinguish it from the other Asian elephants.

Secondly, our maximum likelihood trees place *Elephas antiquus*_N outside of the *Loxodonta* elephants. *E. antiquus*_O could not be used to confirm this due to it's low coverage. Placement of *Elephas antiquus*_N changed according to the method used (maximum likelihood versus maximum parsimony). Due to lack of further evidence, the SNP phylogeny must be accepted.

The limitations of a reference-based binary system

Using a reference genome is never ideal. In this case, it restricts the dataset to intervals found on the *L. africana* genome. We cannot determine if other species have additional repetitive elements at different positions. Given a modern elephant as the reference, we cannot detect any repeats that were present in ancient elephants and lost over time, or any new insertions in other, significantly diverged modern elephants (case in point: *Elephas maximus*).

Using ancient DNA raises other problems due to low coverage. By creating a system of presence/absence of intervals at given sites, and defining variance based on the absent intervals, we are making an inherent assumption that each genome is represented at equivalent coverage and quality to the reference. In reality, most genomes will be far worse than *L. africana*_C (33x). Unfortunately, this problem is not easily resolved due to the low availability and high degradation of ancient DNA data. Hence we cannot make a confident prediction about the *Mammuthus* and *Mammut* elephants.

Conclusions

Based on our results, we believe it is necessary to re-assess the positions of *E. maximus*_Y and *E. antiquus*_N in the evolutionary tree. All of the other relationships are supported, or unresolved with our method. This experiment has shown that repeats can be used as variant site

markers for determining species relationships; but, the results should be interpreted carefully to assess potential bias due to low coverage. We recommend this method be used together with SNP-based approaches as a way of confirming or resolving branches with low support. In cases where there are no known SNP variants, this method can be used to quickly surmise evolutionary relationships and pinpoint species which require further testing.

References

1. U. Landegren, M. Nilsson, P. Y. Kwok, *Genome Res* **8**, 769 (1998).

2. L. Rishishwar, C. E. Tellez Villa, I. K. Jordan, *Mob DNA* **6**, 21 (2015).

3. R. D. Kortschak, D. L. Adelson, *bioRxiv* (2014).

4. R. C. Edgar, E. W. Myers, *Bioinformatics* **21 Suppl 1**, i152 (2005).

5. R. Edgar, *Bmc Bioinformatics* **5**, 113 (2004).

6. R. C. Edgar, *Nucleic Acids Research* **32**, 1792 (2004).

7. O. Kohany, A. Gentles, L. Hankus, J. Jurka, *Bmc Bioinformatics* **7**, 474 (2006).

8. W. Gish, WU BLAST (1996-2004).

9. D. L. Wheeler, *et al.*, *Nucleic Acids Research* **35**, D5 (2007).

10. J. Jurka, *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).

11. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403 (1990).

12. A. R. Quinlan, I. M. Hall, *Bioinformatics* **26**, 841 (2010).

13. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).

14. D. L. Swofford, Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0.b5 (2001).

15. A. Kuritzin, T. Kischka, J. Schmitz, G. Churakov, *PLoS Comput Biol* **12**, e1004812 (2016).

16. A. M. Ivancevic, A. M. Walsh, R. D. Kortschak, D. L. Adelson, *Bioessays* **35**, 1071 (2013).

17. A. M. Walsh, R. D. Kortschak, M. G. Gardner, T. Bertozzi, D. L. Adelson, *Proc Natl Acad Sci U S A* **110**, 1012 (2013).

18. D. L. Adelson, J. M. Raison, R. C. Edgar, *Proceedings Of The National Academy Of Sciences Of The United States Of America* **106**, 12855 (2009).

19. N. Rohland, *et al.*, *PLoS Biol* **8**, e1000564 (2010).

Chapter 5

Retrotransposons: Genomic and Trans-Genomic Agents of Change

"Scientific progress goes 'boink'?"
— Hobbes

The general view of retrotransposons is that they are parasitic; selfishly replicating within the genome until they are silenced by host suppression mechanisms, at which point they transfer to new species in an attempt to escape extinction. It is certainly true that retrotransposon insertions into genes can cause numerous genetic diseases, including but not limited to cancer, autoimmunity or neuropsychiatric disorders. However, much is still unknown about their role in host genomes. In mammalian genomes, for instance, there is growing evidence linking retrotransposon exaptation to regulation of the innate immune system. Evolution works in mysterious ways and at this stage, we cannot dismiss retrotransposons as solely mutagenic agents. 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 w ork w ritten 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

Co-Author

Name of Author (Candidate)	Atma M. Ivancevic		
Contribution to the Paper	Designed two figures, provided suggestions and proof-read the book chapter.		
Overall percentage (%)	15%		
Certification:	This paper reports on original research that was 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 third author of this paper.		
Signature		Date	6/12/16

Other 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 to 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	7 Dec 2016

Name of Co-Author	Reuben M. Buckley		
Contribution to the Paper	Designed figures, provided suggestions and proof-read the book chapter.		
Signature		Date	7/12/2016

Name of Co-Author	Zhipeng Qu		
Contribution to the Paper	Designed figures, provided suggestions and proof-read the book chapter.		
Signature		Date	7/12/16

Name of Co-Author	Lu Zeng		
Contribution to the Paper	Designed figures, provided suggestions and proof-read the book chapter.		
Signature		Date	07/12/16

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

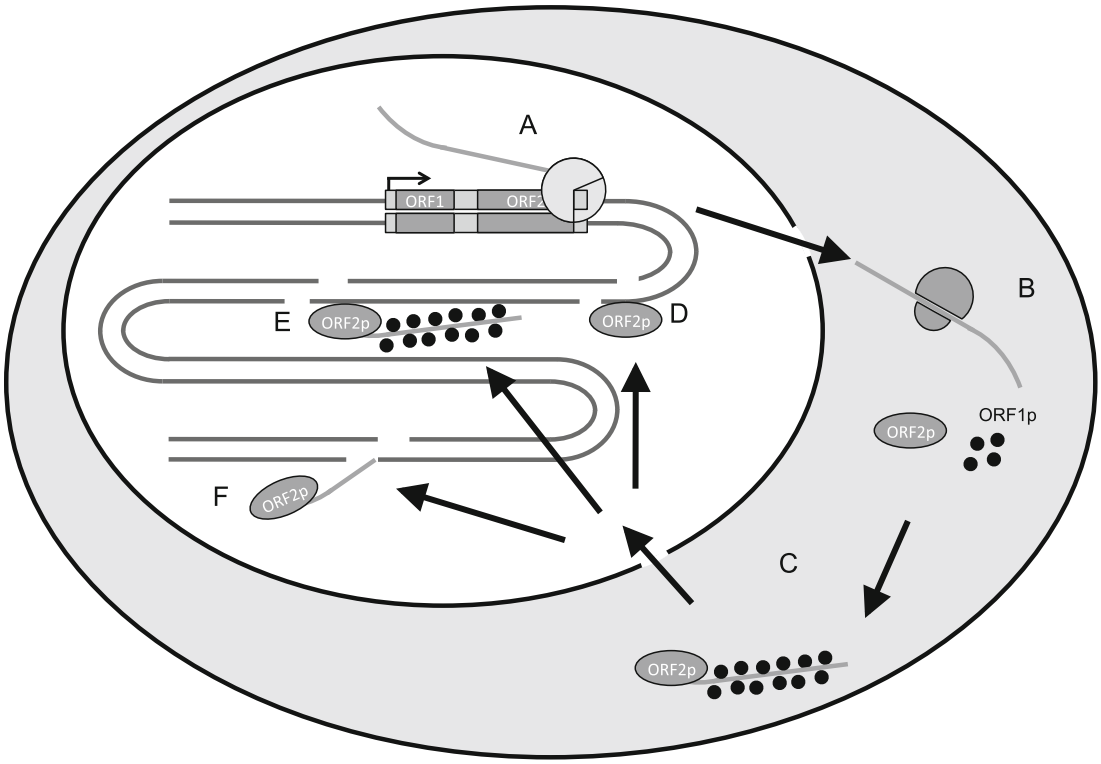


Fig. 4.1 Retrotransposon life cycle: *A* TEs are transcribed by RNA Pol II and exported to the cytoplasm (Swergold 1990). *B* In the cytoplasm, ORF1 and ORF2 are both translated. The ORF1 protein (*ORF1p*) is an RNA-binding protein believed to aid the entry of LINE L1 RNA into the nucleus (Martin 2006). 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

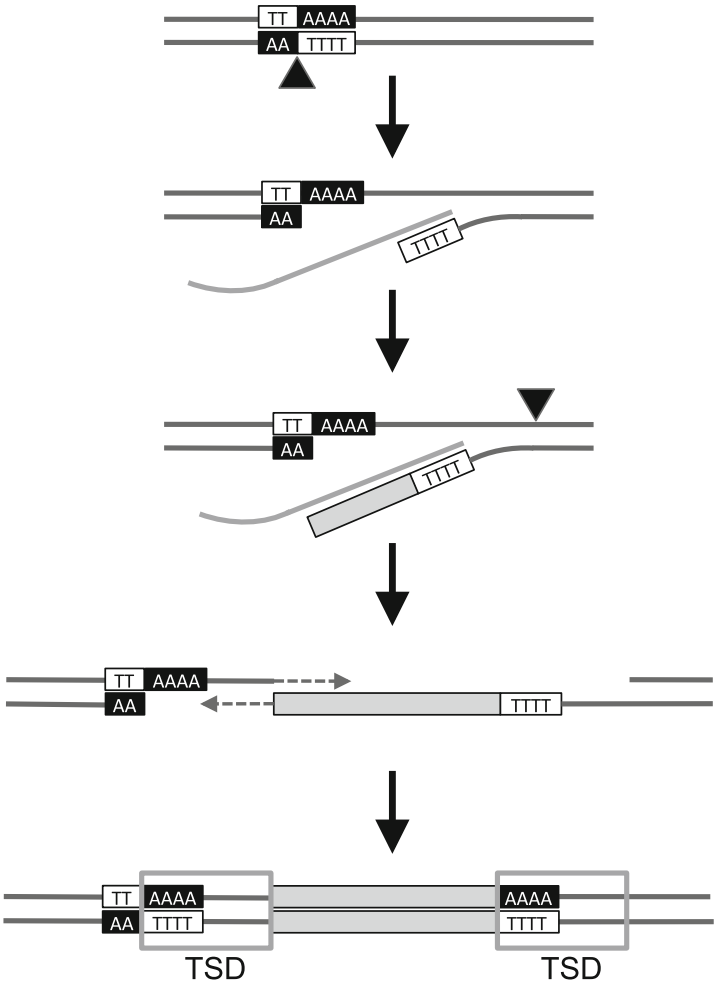


Fig. 4.2 Target-primed Reverse Transcription (*TPRT*) is how retrotransposons are inserted into the genome. ORF2p endonuclease activity creates a nick in the DNA at the AA/TTTT target site (Cost and Boeke, 1998). ORF2p reverse transcriptase activity then uses the cDNA copy as a template for DNA synthesis. 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).

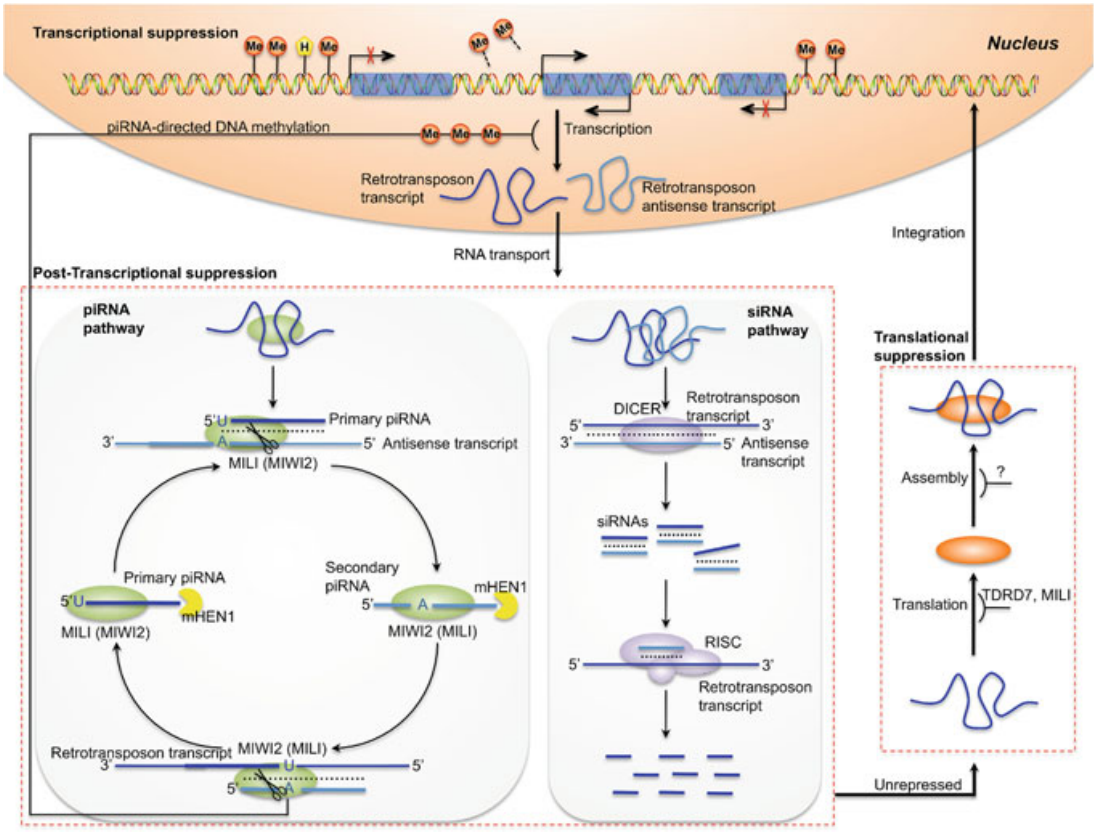


Fig. 4.3 A schematic overview of retrotransposon suppression. Retrotransposons can be suppressed by different mechanisms throughout their life cycle (Crichton et al. 2014). Transcriptional suppression: In most cell types, retrotransposons are in a repressed state due to high levels of DNA methylation 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

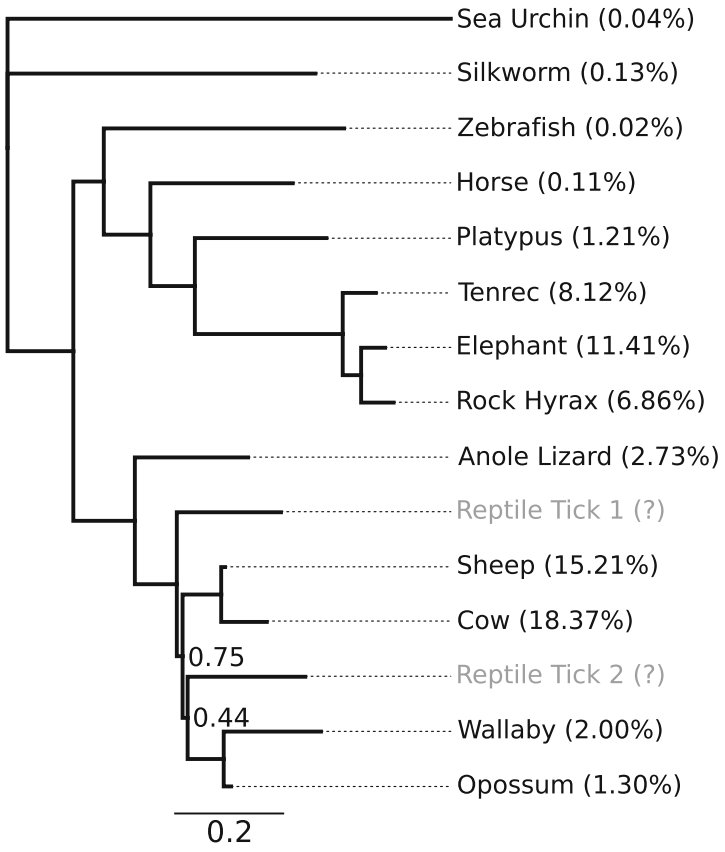


Fig. 4.4 BovB phylogeny Maximum likelihood tree of aligned BovB sequences based on Walsh et al. (2013), showing the sporadic distribution, sequence similarity and abundance of BovB elements across taxa. Local support values are only shown if <0.9. The labels at each branch tip give the species common 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.

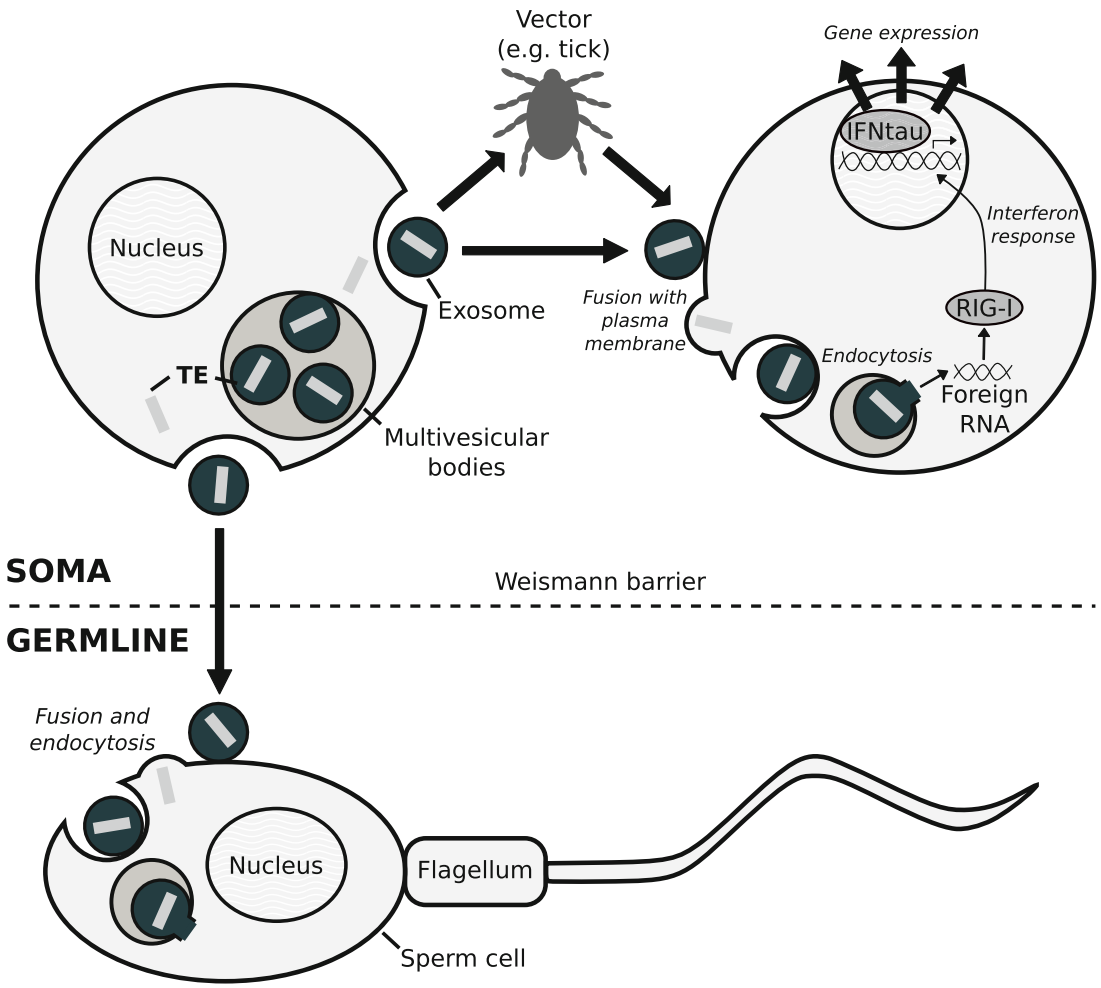


Fig. 4.5 Possible scenarios of intercellular transfer of transposable elements via exosomes. TEs packaged in exosomes can be transferred between both somatic and germline cells. Within an organism, a TE can travel from a somatic, exosome-generating cell directly (e.g. through the blood) into a somatic, 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

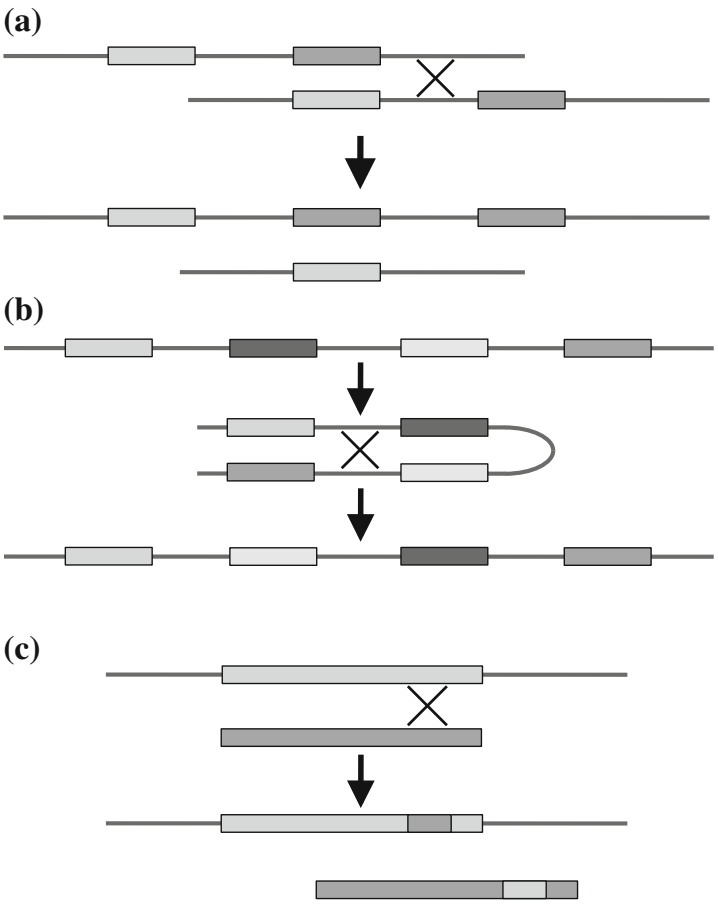


Fig. 4.6 Retrotransposons can lead to changes in genome structure. **a** Changes in CNVs result from non-allelic homologous recombination (*NAHR*) caused by the insertion of many TEs from the same family (Stankiewicz and Lupski 2002; Startek et al. 2015). **b** Chromosomal inversion is also the result 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,

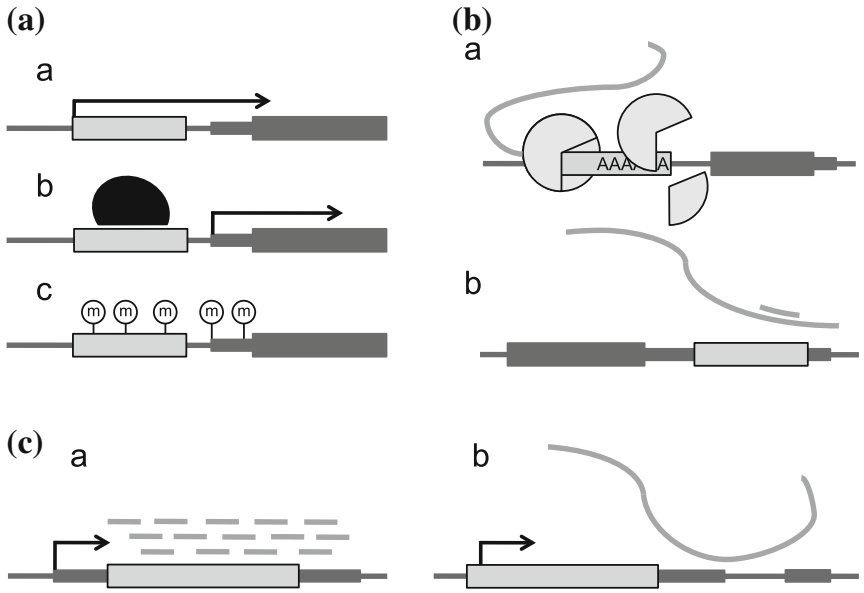


Fig. 4.7 Retrotransposons can alter gene expression. **a** 5' insertion of a retrotransposon with respect to a gene. *a* TEs are able to act as alternative promoters to adjacent genes (Faulkner et al. 2009; Speek 2001). *b* TEs are able to act as transcription factor binding sites (*TFBS*) and are thereby 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

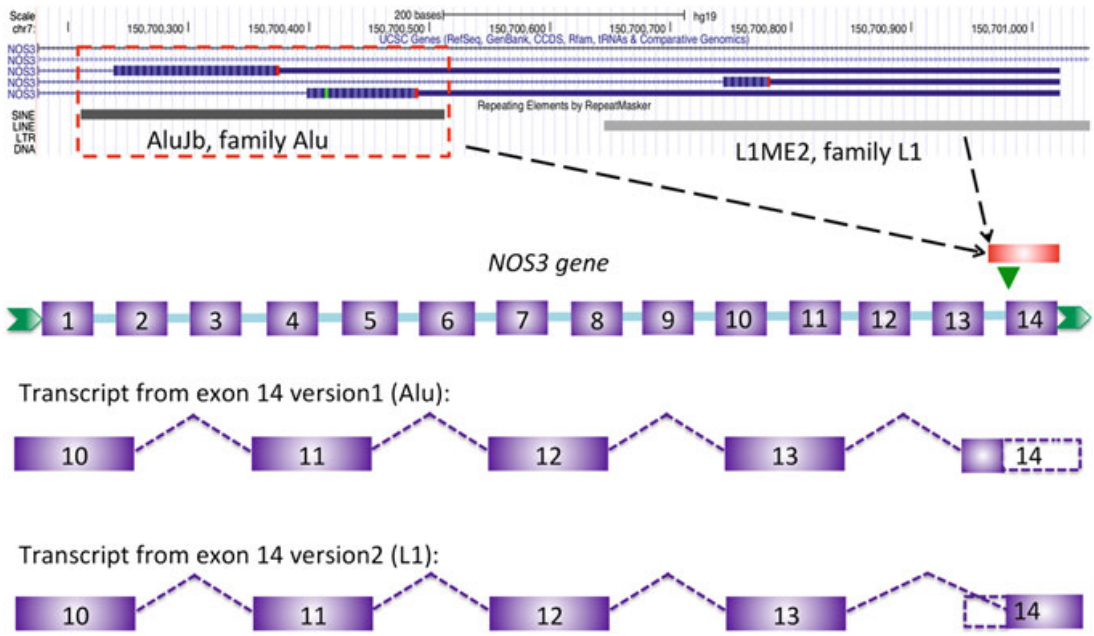


Fig. 4.8 Retrotransposon exaptation influences mRNA processing and can cause multiple splice variants. At the top, the UCSC browser (Kent et al. 2002) track for the human NOS3 gene is shown, including repeat element annotation. *Below*, a schematic of the 3' end of the human NOS3 gene illustrating an 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

> *"Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world."*
>
> — Albert Einstein

Transposable elements are largely still considered 'junk' DNA because they provide no clear adaptive advantage and their repetitive nature makes them difficult to analyse. However, they occupy huge portions of eukaryotic genomes and it is becoming clear that they are a continual source of genomic change. This thesis has contributed to the understanding of how retrotransposons have shaped the genomes we see today, and their potential to cause further changes. It has also implemented novel bioinformatics approaches for handling large datasets, including the use of repeats as informative markers to resolve genome differences.

One of the biggest challenges of working with repeats is that many bioinformatics tools are written for unique sequences, and discard repeats. For example, programs such as CENSOR and RepeatMasker were used in this thesis to annotate different types of repeats, but in the literature they are primarily used to generate a masked, repeat-free version of the genome. Other programs will by default ignore repetitive portions of the query sequence. Alignment programs also struggle to accurately resolve repetitive sequences (especially when there are thousands of them). Protein prediction models such as HMMer rely on similarity to known motifs from gene-dominant databases, resulting in many unknown or uncharacterised domains being found in L1s. Every step of this project required careful parameter testing and evaluation to optimise the analysis for retrotransposons.

Another challenge was using publicly available draft genome assemblies as species representatives. Advances in technology have resulted in a significant increase in the

number of sequenced genomes, but most of these are left at the scaffold or contig level rather than fully-assembled chromosomes. Repetitive regions are poorly represented because next-generation sequencing technologies often produce short reads, making it difficult to reliably align the reads which map to multiple locations. To this end, it was often necessary to use multiple searching strategies and independent assemblies of closely related species, where available.

There are many areas of this thesis which can be explored in more depth. For example, we were able to identify many novel domains in L1s from different species - but we do not know their importance in terms of functional capability. Future work should focus on elucidating these cryptic regions of L1s with laboratory experiments, e.g. mouse models can be used to generate L1 mutants which lack certain domains and determine whether this has an impact on L1 activity.

In primates, we found that the snub-nosed monkey contained 'hyperactive' levels of (structurally) active L1s compared to human. A key difference distinguishing this monkey from other primates is over-representation of DUF4417 - an uncharacterised protein domain which appears directly after the reverse transcriptase in Old World monkeys and apes. Human has 60 instances of L1s containing DUF4417; the snub-nosed monkey has over 3000. To test this correlation, subsequent analyses could introduce mutations in the conserved amino acids identified within this domain in human and test their effect on L1 retrotransposition, or test the behaviour of snub-nosed monkey L1s in retro-assays.

In terms of horizontal transfer, the mechanism and frequency of transfer are largely still unknown. Future research should focus on determining factors which influence the ability of retrotransposons to jump into new hosts and then expand throughout the genome. For example, is it possible to introduce BovB retrotransposons to humans or mice? And if so, can BovBs infiltrate the germline and become permanent fixtures in these species? Wet lab experiments to this effect would greatly increase our knowledge of retrotransposon capabilities.

Finally, there are two main shortcomings to the approach described in chapter 4: 1) there are significant limitations to using a reference-based system, and 2) the approach was only tested on one set of species. Further testing should include a wider range of species, both modern and ancient, and multiple possible reference species. This would allow us to identify the best parameters for resolving TE variant sites, especially with low coverage data.

Appendix A

Supplementary Material

The attached CD-ROM contains supplementary material for chapters 1, 2, 3 and 4.

For chapter 1, there is one supplementary table (Table A1) describing known eukaryotic horizontal transfer cases and proposed vectors.

For chapter 2, the supplementary material includes Materials and Methods (Table B1 to Table B3, Figure B1 to B4) and Results (Table B4 to B12, Figure B5 to B7).

For chapter 2, the supplementary material includes a detailed description of the Materials and Methods used (as well as Table C1) and additional Results (Table C2 to C8, Figure C1 to C3).

For chapter 3, the supplementary material includes additional figures and tables (Figure D1 to D5, Table D1).