The Evolution of LINE-1 in Vertebrates

Stéphane Boissinot[1,*] and Akash Sookdeo[2]

[1]NYU Abu Dhabi, Saadiyat Island campus, Abu Dhabi, United Arab Emirates

[2]Department of Biology, Queens College, CUNY

*Corresponding author: E-mail: sb5272@nyu.edu.

Accepted: October 3, 2016

Abstract

The abundance and diversity of the LINE-1 (L1) retrotransposon differ greatly among vertebrates. Mammalian genomes contain hundreds of thousands L1s that have accumulated since the origin of mammals. A single group of very similar elements is active at a time in mammals, thus a single lineage of active families has evolved in this group. In contrast, non-mammalian genomes (fish, amphibians, reptiles) harbor a large diversity of concurrently transposing families, which are all represented by very small number of recently inserted copies. Why the pattern of diversity and abundance of L1 is so different among vertebrates remains unknown. To address this issue, we performed a detailed analysis of the evolution of active L1 in 14 mammals and in 3 non-mammalian vertebrate model species. We examined the evolution of base composition and codon bias, the general structure, and the evolution of the different domains of L1 (5′UTR, ORF1, ORF2, 3′UTR). L1s differ substantially in length, base composition, and structure among vertebrates. The most variation is found in the 5′UTR, which is longer in amniotes, and in the ORF1, which tend to evolve faster in mammals. The highly divergent L1 families of lizard, frog, and fish share species-specific features suggesting that they are subjected to the same functional constraints imposed by their host. The relative conservation of the 5′UTR and ORF1 in non-mammalian vertebrates suggests that the repression of transposition by the host does not act in a sequence-specific manner and did not result in an arms race, as is observed in mammals.

Key words: LINE-1, L1, vertebrate, molecular evolution.

Introduction

The LINE-1 (or L1) non-LTR retrotransposon is one of the most widely distributed transposable elements in vertebrate genomes (Tollis and Boissinot 2012). The abundance and diversity of L1 differs considerably among vertebrates, and is probably one of the genomic features that show the most variation in this group. At one end of the spectrum, mammalian genomes host an extremely large number of L1 insertions that have accumulated since the origin of mammals and account for close to 20% of their mass (Lander et al. 2001; Mouse Genome Sequencing et al. 2002). In contrast, L1 in non-mammalian vertebrates are represented by much smaller copy numbers, from a few hundreds to several thousand elements, representing <0.5% of their genome size (Hellsten et al. 2010; Howe et al. 2013). This is likely due to a higher rate of DNA deletion in these genomes but could also reflect variations in the rate of fixation of novel insertions, or both (Duvernell et al. 2004; Furano et al. 2004; Novick et al. 2009; Blass et al. 2012; Tollis and Boissinot 2013).

Another difference between mammals and non-mammals reside in the mode of evolution of L1 (fig. 1). In mammals, only the most recently evolved group of elements is active at a given time so that a single family of progenitor is usually producing novel insertions. In the long-term, this mode of evolution results in a ladder-shaped phylogeny, demonstrating the replacement of one family by a younger one, and so forth (Smit et al. 1995; Furano 2000). This mode of evolution is consistent with an arms race between the host, which represses L1 transposition, and L1, which evolves to bypass repression by the host. Conversely, in reptiles and fish, several highly divergent families are concurrently active in the same genome. These active families have coexisted for extended period of time, since their divergence may pre-date the origin of vertebrates (Furano et al. 2004; Novick et al. 2009).

The differences in the evolutionary dynamics of L1 among vertebrates have far-reaching consequences because L1 activity has considerably influenced other genomic features and since L1 insertions can be both a source of deleterious alleles

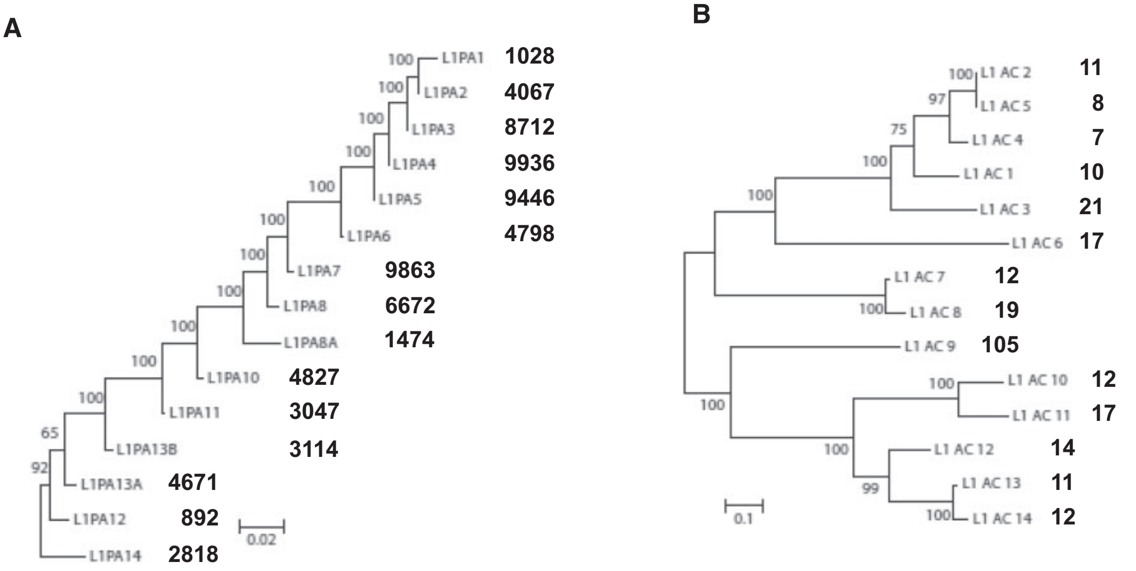

Fig. 1.—Pattern of evolution of L1 families in mammals and non-mammals. The phylogenies are ML trees based on the data of Khan et al. (2006) and Novick et al. (2009). The number of copy for each family is indicated in bold. (*A*) This phylogeny represents the evolution of L1 families in human and demonstrates the ladder-like mode of evolution typical of mammals. (*B*) This phylogeny is based on lizard L1 families (Novick et al. 2009) and is typical of non-mammalian vertebrates (reptiles, amphibians and fish).

(Boissinot et al. 2006) and evolutionary novelties (Warren et al. 2015). It is thus important to determine the mechanisms responsible for these differences. A number of studies have examined the population dynamics of L1 insertions in mammalian (Boissinot et al. 2006; Witherspoon et al. 2006; Rishishwar et al. 2015) and non-mammalian species (Duvernell, et al. 2004; Blass et al. 2012; Tollis and Boissinot 2013) but no studies have examined the evolution of the L1 sequence across vertebrates. In fact, almost everything we know about L1, from its structure to its mechanism of transposition, results from studies in mammals, with a focus on human and murine rodents (for a recent review, see Richardson et al. 2015). Considering the difference in the evolutionary dynamics of L1 between mammals and non-mammalian vertebrates, it is unlikely that everything we know from mammals applies to fish and reptiles. Analyzing L1 evolution in a phylogenetically broader comparative context could certainly improve our understanding of the biology of L1 across genomes but also give us powerful insights into mammalian L1 biology.

L1 transpose through a process called target-primed reverse transcription (TPRT) where reverse transcription of the L1 RNA into cDNA takes place at the site of insertion (Luan et al. 1993; Cost et al. 2002). A typical mammalian L1 element is 6–7 kb long and contains a 5′UTR, two open-reading frames (ORF1 and ORF2) and a 3′UTR (fig. 2*A*). L1 insertions typically end with an A-rich tail and are flanked by short (<10 bp) target site duplication. In modern human L1, the 5′UTR contains a CpG island and acts as an internal promoter, which drives transcription of the full-length L1 transcript (Swergold 1990; Severynse et al. 1992; DeBerardinis and Kazazian 1999). The 5′UTR of the mouse and rat L1 is bipartite and consist of tandem arrays of monomers (~200 bp for mouse, ~650 bp for rat), which contain CpG-island and transcriptional signals, connected to ORF1 by an ~250-bp region called the tether (fig. 2*A*) (Adey, Tollefsbol, et al. 1994; Furano 2000). The 5′UTR shows little or no homology among mammalian species or even among families within the same species (Adey, Schichman, et al. 1994; Khan et al. 2006; Sookdeo et al. 2013). Evolutionary analyses in primates and rodents have demonstrated that L1 lineages have repeatedly acquired novel 5′UTR, possibly in response to the host repression of L1 transcription (Jacobs et al. 2014). The human 5′UTR contains an anti-sense promoter on the negative strand (Speek 2001) and a small ORF, termed ORF0, which is transcribed and translated but has no known function (Denli et al. 2015).

ORF1 and ORF2 are both necessary for L1 transposition. ORF1 contains a coiled-coil domain (CCD), which promotes the formation of ORF1p trimers, a non-canonical RNA recognition motif (RRM) and a highly conserved C-terminus domain (Martin and Bushman 2001; Martin et al. 2003; Januszyk et al. 2007; Khazina and Weichenrieder 2009). The function of ORF1 remains obscure but it has been shown to have nucleic acid chaperone activity (Martin and Bushman 2001) and recent studies showed that the human ORF1p requires phosphorylation for retrotransposition in a cell culture-based assay (Cook et al. 2015). ORF1p participate in the formation of L1 ribonucleoprotein particles (RNP), which is a necessary step of

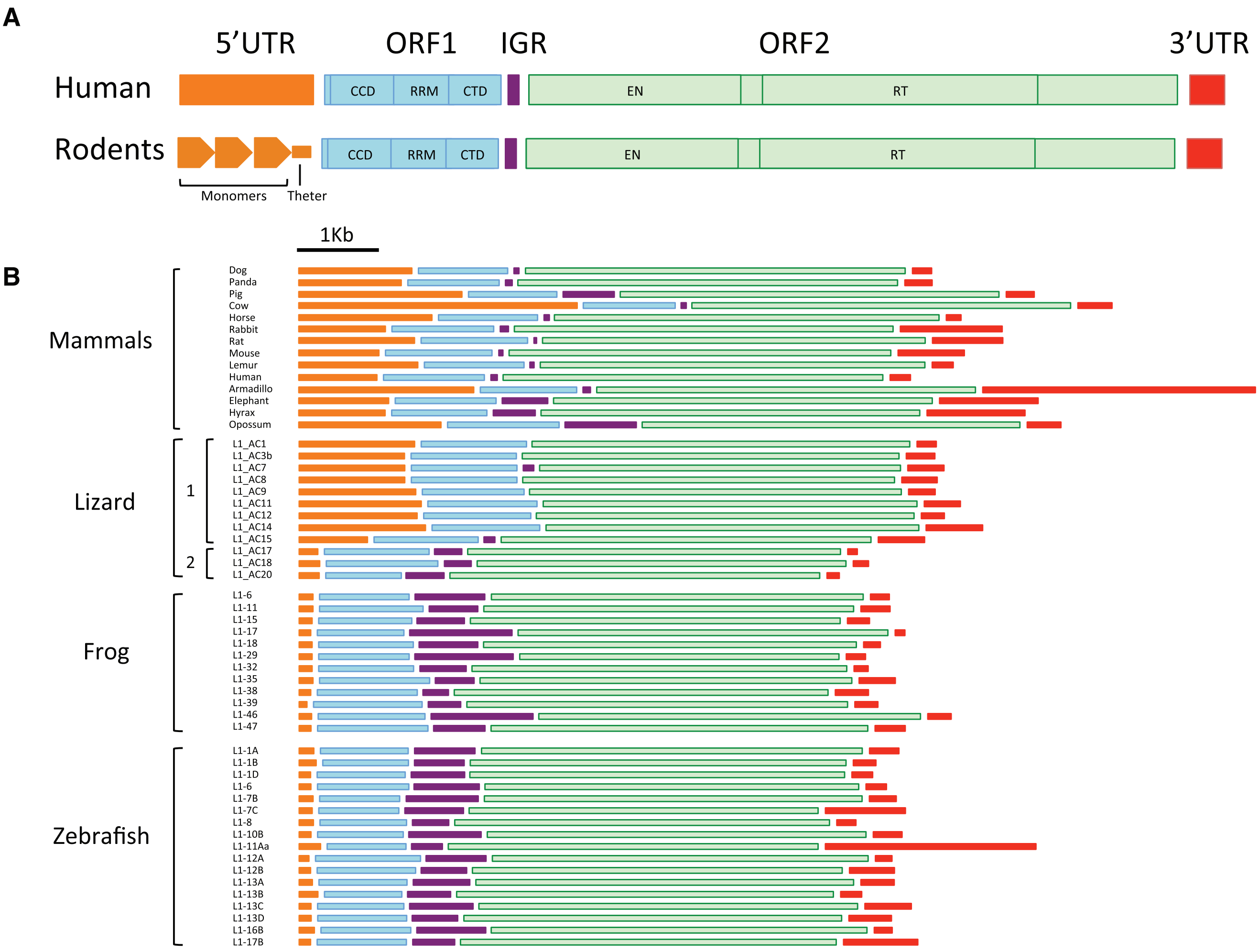

Fig. 2.—(*A*) Typical structure of human and murine rodents full-length L1 elements (CCD = Coiled-coil domain; RRM = RNA recognition motif; CTD = C-terminal domain; EN = Endonuclease domain; RT = Reverse transcriptase domain). (*B*) Schematic structure of full-length L1 families in mammals, lizard, frog and zebrafish.

the transposition process (Kolosha and Martin 1997; Kulpa and Moran 2005). Interestingly, recent studies showed that interactions of purified ORF1p with nucleic acids exemplified several of the predicted properties of the L1 RNP, including stabilization of the putative TPRT intermediate (Callahan et al. 2012). The latter paper is particularly informative as it demonstrated that rapid oligomerization between ORF1p trimers upon their binding to nucleic acid is essential for retrotransposition, a novel coiled-coil-dependent property which is conserved despite extensive remodeling of the coiled-coil during evolution. ORF2 is highly conserved among mammals and contains endonuclease and reverse transcriptase domains (Mathias et al. 1991; Feng et al. 1996). A short (~40 bp) inter-genic region (IGR) separates the two ORFs in human, whereas an IGR spanning several hundred base pairs was found in marsupials, megabats and afrotheria (Yang et al. 2014). Mouse L1 lacks an IGR and the 3′ end of ORF1 overlaps with the 5′ end of ORF2. The dicistronic structure of L1 is unusual in eukaryotes and it is still unclear how the ORFs are translated. Two possibilities have been offered. Either there are two ribosome entry sites (Li et al. 2006), one for each ORFs, or the ribosome that translated ORF1 scan through the IGR to ORF2 start codon and reinitiate translation (Alisch et al. 2006).

The ORFs of L1 are AT rich, with a strong A-bias on the positive strand, which could account for premature poly-adenylation signals and inefficient transcription, at least in cell culture based retrotransposition assays (Perepelitsa-Belancio and Deininger 2003; Han et al. 2004). The 3′UTR shows very little conservation among species, yet all mammalian 3′UTRs contain a poly-G tract of unknown function (Howell and Usdin 1997) and end with a functional but weak poly-adenylation signal, which is often by-passed during transcription, resulting in the transduction of 3′ flanking sequences (Pickeral et al. 2000).

Here, we performed a comparative analysis on the evolution of L1 active families across mammals and in three non-mammalian vertebrates species. We demonstrate that the length, structure, and base composition of L1 differs substantially among vertebrates but is remarkably conserved within species, exemplifying the finely tuned co-evolution between L1 and its host. We propose that these variations reveal fundamental differences in the nature of the interactions between L1 and its vertebrate hosts.

Materials and Methods

We collected full-length copies from the genome of 14 mammals (the opossum, armadillo, elephant, hyrax, rat, mouse, rabbit, mouse lemur, human, dog, giant panda, horse, cow, and pig), a non-avian reptile (the green anole *Anolis carolinensis*), an amphibian (the African-clawed frog *Xenopus tropicalis*), and a teleost fish (the zebrafish *Danio rerio*). The coordinates of the elements were retrieved from the repeatmasker tables available at the genome.ucsc.edu website. Each element was recovered with 2 kb of sequence upstream and downstream to accurately identify the start and the end of the full-length element. Only full-length and recently active elements were used in the analysis to limit the uncertainties inherent to the construction of consensus sequences in each species. A phylogenetic analysis using ORF2 was first performed to identify active or recently active families, which were recognized as monophyletic clusters of elements with branch length <2% divergence. To insure accuracy, we only used families for which we could collect at least eight full-length genomic copies. A consensus sequence was then derived for each active or recently active family. This approach was used for all organisms except for human and mouse, for which we used the consensi described in Khan et al. (2006) and Sookdeo et al. (2013).

Sequences were manipulated and consensi were generated using Geneious 8.1.5, created by Biomatters and available at www.geneious.com (last accessed October 20, 2016). The location of the ORFs was determined using the ORF finder tool implemented in Geneious 8.1.5 and the presence of functional motifs was determined with the search tool at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi (last accessed October 20, 2016). The level of identify among amino acid sequences was calculated using Geneious 8.1.5. Searches for similarity among regions that could not be reliably aligned (the UTRs) were performed using DOTMATCHER at http://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher (last accessed October 20, 2016). Repeated motifs were searched using the Tandem Repeats Finder website at https://tandem.bu.edu/trf/trf.html (last accessed October 20, 2016) (Benson 1999). The presence of known phosphorylation motifs was determined using the Eukaryotic Linear Motifs search engine at http://elm.eu.org/. The presence and structure of CCDs was assessed using the COILS server at http://www.ch.embnet.org/software/COILS_form.html (last accessed October 20, 2016). COILS calculate the probability that a given protein sequence forms a coiled-coil structure. Protein domains that can form coiled coils typically consist of seven residues repeats (or heptads) with non-polar or hydrophobic residues at the first (a) and fourth (d) positions of the heptads. The analysis was run with window width of 14, 21, and 28. Since the results were very similar among analyses, we present only the results obtained with the 28 residues window width. A conservative cut-off of 90% was used to define an amino acid as participating in a coiled coil structure. In our analysis, we differentiated canonical heptads (a–b–c–d–e–f–g), non-canonical coiled coils (regions with a high probability to participate in the formation of a coiled coil but which deviate from the canonical heptad structure, for example, a–b–c–b–c–d–e–f–g or a–b–c–g–a–f–g) and non-coiled coil regions (with low probability to participate in the formation of a coiled coil).

A phylogeny of all active families was built using the maximum likelihood method and a LG + G+I + F model of mutation, as determined by the model estimation tool, implemented in MEGA 5.0 (Tamura et al. 2011). The robustness of the nodes was determined using 1,000 bootstrap replicates.

Base composition and codon usage were determined using the CAIcal program at http://genomes.urv.es/CAIcal/ (last accessed October 20, 2016) (Puigbo et al. 2008). For each codon, the Relative Synonymous Codon Usage (RSCU) was estimated (Sharp et al. 1986). The RSCU is defined as the number of time a codon is used for a given amino acid divided by the number of synonymous codons for that amino acid. We also calculated two estimators of codon bias: Nc (Wright 1990) and CAI (Sharp and Li 1987). Nc (e.g., the effective number of codon used in a gene) quantifies how much the use of a specific codon deviates from equal use of all synonymous codons for a given amino acid. Nc ranges in value from 20 (when each amino acid is exclusively encoded by a single synonymous codon) to 61 (when all synonymous codons are equally represented). The parameter CAI (Codon Adaptation Index) estimates the codon bias given the codon usage of the organism and the GC content of the gene. It ranges from 0 to 1, 1 meaning that it is always the most common synonymous codon that is used and the codon bias is low. The codon usage of the organisms was obtained from the codon usage database at http://www.kazusa.or.jp/codon/ (last accessed October 20, 2016). For mammals we performed the analysis with the human and the mouse codon usage and we obtained identical results. Since the lizard codon usage was poorly represented in the database, we estimated the codon usage from the lizard cDNA entries available in GenBank. Statistical significance of CAI is estimated by comparing the observed CAI values with the expected CAI (or eCAI), which describes the random codon usage assuming the GC content of the gene studied.

RNA secondary structures were investigated using the RNAfold web server at the Vienna RNA web suite (Gruber et al. 2008). Putative Internal Ribosome Entry Sites (IRES) were identified using the IRESPred tool, which uses 35 features based on sequence and structural properties to predict the presence of cellular or viral IRES (Kolekar et al. 2016). We also used the Viral IRES Prediction System (VIPS), which uses the secondary structure of four groups of known viral IRES to predict putative viral IRES (Hong et al. 2013).

Results

We derived full-length consensus sequences for 14 mammalian species, the green anole, the African clawed frog and the zebrafish. A single consensus was derived for each mammalian species because their genome hosts a single active (or recently active) L1 family, with the exception of the house mouse. The mouse genome contains three active families but only one of them was included in this analysis (L1Md_A) since they are similar in sequence and have been analyzed in details elsewhere (Sookdeo et al. 2013). In the anole, frog, and zebrafish, we derived 12, 12, and 17 L1 consensi, respectively. These species are known to host a larger number of L1 families (Furano et al. 2004; Novick et al. 2009) but the stringency of the criteria we used to construct full-length consensi did not permit deriving consensi for all active families. Thus, the dataset analyzed here consists of 55 consensus sequences (available as fasta file in supplementary material S1, Supplementary Material online).

Phylogenetic Relationships and Divergences

We first performed a phylogenetic analysis using the most conserved region of L1, ORF2 (fig. 3). Mammalian L1 sequences form a monophyletic group with strong support. The lizard L1 elements also form a clade, composed of two divergent sub-clades (Lizard clade 1 and clade 2 on fig. 3). Two highly divergent clades are also found in zebrafish and frog but these clades do not form species-specific groups, suggesting that their divergence could have occurred before the split between teleostean fish and tetrapods.

The identity among mammalian ORF2 ranges from 48.4 to 76.1% (table 1). As expected the identity between the opossum L1 and the placental mammals is lower (48.4–53.6%) than among placentals (58.2–76.1%). The identity between the two most divergent clades in lizard, frog, and zebrafish is comparatively much lower, with average values of 26.5%, 27.5%, and 31.2%, respectively. The identity within each of the clades is also low with average values of 51.8%, 38.5%, and 37.9% for clade 1 in lizard, frog, and fish, respectively. The divergence between L1 families, as well as the phylogenetic analysis, clearly indicates that each non-mammalian genome contains a large diversity of L1 families, which is very ancient and has persisted since the origin of vertebrates.

Structural Evolution of LINE-1 in Vertebrates

Full-length L1 elements vary substantially in length among and within organisms (table 2, fig. 2 and supplementary material S2, Supplementary Material online). Mammalian L1s tend to be longer (7.1 kb on average) than the frog (5.7 kb), and fish (5.8 kb) L1s. There is no significant difference in the length of the elements belonging to the two main L1 clades in frog and fish. In lizard, elements of clade 1 are similar to mammalian L1 in length (6.4 kb on average) but clade 2 elements are similar to frog and fish with regard to length (5.4 kb).

Since there is very little variation in the length of ORF1 and ORF2, differences among vertebrate L1s are caused by variation in the length of the 5′UTR, 3′UTR, and IGR (table 2). Mammalian L1s and lizard clade 1 elements are characterized by 5′UTRs that are considerably longer (1.5 and 1.3 kb on average, respectively) than the fish, frog, and lizard clade 2 L1s (0.16, 0.14, and 0.23 kb, respectively). The length of the 3′UTR can also differ greatly among families, yet these variations do not follow a clear evolutionary pattern and the acquisition of long 3′UTRs seems to have occurred sporadically. For example, the 3′UTR of mammals shows a considerable range of length from 148 bp in horse to 994 bp in elephant, with an extreme value of 2,751 bp in armadillo. Similarly the 3′UTR of the zebrafish ranges from 167 to 807 bp, with the evolution of a very long 3′UTR of 2,124 bp in the L1-11A family.

The presence and length of an IGR also significantly affects the overall length of the elements (table 2, fig. 2). Fish, frog, and lizard clade 2 elements have a relatively long IGR that ranges from 257 to 1,032 bp. Conversely, six out of the nine lizard clade 1 elements are lacking an IGR, and for five of those, ORF1 and ORF2 overlap. Most mammalian L1 have a small IGR ranging from 26 to 82 bp. The exceptions are the opossum, elephant, hyrax, and pig, with IGR ranging from 423 to 719 bp (fig. 4). Assuming that Afrotheria (elephant and hyrax) is the sister group to all other placental mammals (Meredith et al. 2011), we can infer that the ancestral mammalian L1 had a long IGR that was lost after the split between Afrotheria and the other placentals and that the pig IGR was acquired independently.

Base Content and Codon Usage

To investigate the intrinsic constraints on LINE composition and how this is influenced by the context of their host genomes, we compared GC content and codon usage across species and among families within species. The overall base content of L1 in vertebrates tends to be AT-rich, with GC content ranging from 33.9 to 48.1%. There are considerable differences in base composition among regions of L1 and among vertebrates (table 2), although the average genomic GC content differs only moderately among vertebrates (~41% on average in mammals, ~40.3% in lizard, ~40.0% in frog, and ~38.6% in zebrafish). There is however very little variation in the GC content of L1 within species, even among the divergent lizard, frog, and fish families (table 2 and supplementary material S2, Supplementary Material online). The

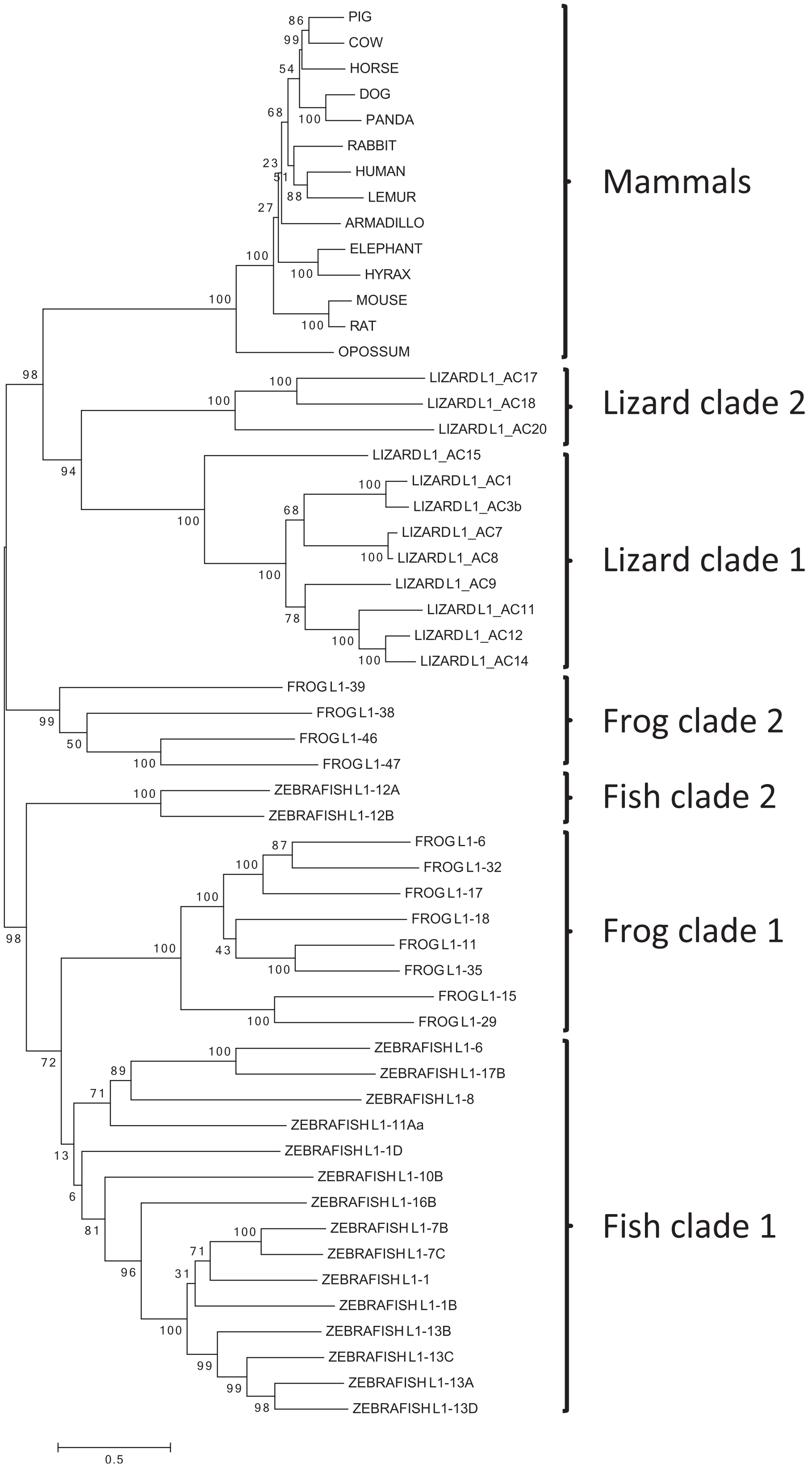

Fig. 3.—Maximum likelihood phylogeny of L1 families based on ORF2 amino acid sequences.

Table 1

Amino Acid Identity among L1 Families in Mammals, Lizard, Frog, and Zebrafish

	HUMAN	LEMUR	RAT	MOUSE	RABBIT	DOG	PANDA	HORSE	PIG	COW	ARMADILLO	HYRAX	ELEPHANT	OPOSSUM
HUMAN		58.8	46.2	43.4	50.0	52.2	50.5	55.5	53.8	55.5	50.0	52.7	50.0	33.9
LEMUR	67.2		45.6	44.0	51.6	45.1	47.3	54.9	50.0	53.8	45.1	45.6	47.8	29.5
RAT	63.0	59.9		86.3	46.2	45.1	46.7	46.2	51.1	48.4	49.5	43.4	45.6	32.8
MOUSE	62.6	60.7	84.0		46.7	40.1	44.0	42.9	47.8	46.2	45.1	40.7	43.4	30.1
RABBIT	67.4	65.7	63.9	63.4		51.4	47.0	55.8	54.7	54.1	54.1	49.7	51.4	30.8
DOG	64.2	64.1	61.5	61.2	66.9		61.9	55.8	57.5	53.6	56.9	50.8	51.9	33.0
PANDA	62.3	61.6	60.1	60.0	64.7	76.1		53.0	54.7	47.5	52.5	47.0	43.6	34.1
HORSE	67.0	64.7	64.2	63.2	67.0	67.4	67.0		64.6	61.3	61.3	56.4	52.5	33.1
PIG	67.4	66.3	64.3	62.7	68.3	68.2	67.8	72.0		63.0	63.0	52.5	55.8	34.1
COW	66.1	66.1	63.2	63.6	67.7	68.5	68.2	71.9	74.5		56.9	53.6	55.8	35.2
ARMADILLO	64.3	64.3	60.9	60.5	64.7	63.7	63.5	65.6	65.6	66.7		51.4	53.0	33.5
HYRAX	61.1	59.8	58.7	58.2	62.2	61.7	58.4	61.6	61.9	62.6	62.9		70.7	28.6
ELEPHANT	61.9	60.9	59.7	59.4	63.8	63.8	61.9	64.4	64.8	64.9	65.5	73.8		29.7
OPOSSUM	52.2	53.6	49.4	48.4	52.8	50.5	50.3	52.4	53.2	52.8	54.2	49.4	50.2	

	LIZARD L1_AC1	LIZARD L1_AC3b	LIZARD L1_AC7	LIZARD L1_AC8	LIZARD L1_AC9	LIZARD L1_AC11	LIZARD L1_AC_12	LIZARD L1_AC14	LIZARD L1_AC15	LIZARD L1_AC17	LIZARD L1_AC18	LIZARD L1_AC20
LIZARD L1_AC1		90.4	53.4	52.8	54.5	50.0	50.0	48.9	26.7	21.5	16.6	22.1
LIZARD L1_AC3b	82.7		55.1	54.5	59.0	50.6	49.4	47.8	25.0	21.0	16.6	21.5
LIZARD L1_AC7	52.4	53.2		96.6	52.8	56.2	56.2	53.4	25.6	17.1	16.0	19.3
LIZARD L1_AC8	52.7	53.1	93.8		52.2	56.2	55.6	52.8	25.0	17.1	16.0	19.3
LIZARD L1_AC9	50.0	49.8	51.1	52.0		50.6	52.8	50.0	26.7	21.0	18.8	20.4
LIZARD L1_AC11	48.7	49.6	48.3	48.6	51.8		67.4	67.4	26.1	19.3	17.7	23.2
LIZARD L1_AC_12	49.8	50.3	50.2	51.0	54.9	66.2		76.4	26.7	19.9	17.7	21.5
LIZARD L1_AC14	50.2	50.5	50.7	51.3	53.8	65.0	79.2		27.8	17.7	16.6	19.3
LIZARD L1_AC15	38.2	37.4	36.7	37.5	38.4	38.6	38.1	37.5		23.1	22.5	17.6
LIZARD L1_AC17	28.0	28.1	25.3	25.4	27.6	26.9	27.2	26.8	27.1		63.6	34.1
LIZARD L1_AC18	27.5	26.8	25.1	24.8	25.7	24.6	25.8	25.4	25.5	43.5		31.8
LIZARD L1_AC20	28.2	28.3	26.2	26.5	26.5	26.3	26.5	26.0	26.5	34.1	34.5	

	FROG L1-6	FROG L1-11	FROG L1-15	FROG L1-17	FROG L1-18	FROG L1-29	FROG L1-32	FROG L1-35	FROG L1-38	FROG L1-39	FROG L1-46	FROG L1-47
FROG L1-6		53.3	44.6	51.7	56.1	46.7	51.1	55.6	24.9	24.3	24.3	21.3
FROG L1-11	41.4		48.0	50.3	48.9	55.5	54.4	65.0	27.1	26.5	28.7	23.0
FROG L1-15	33.7	32.9		44.6	48.0	52.5	50.3	52.5	24.7	28.1	25.3	25.0
FROG L1-17	45.2	40.7	33.5		55.0	44.5	51.1	53.1	28.2	19.9	22.7	26.2
FROG L1-18	41.3	42.5	31.8	42.1		42.9	50.0	52.2	26.0	23.2	24.3	23.5
FROG L1-29	31.6	35.0	39.9	34.3	32.9		44.0	56.0	24.6	25.1	23.5	20.2
FROG L1-32	47.6	40.2	32.0	45.7	41.1	32.7		54.4	27.3	24.0	26.2	25.7
FROG L1-35	40.3	50.4	34.3	38.2	41.0	34.7	40.6		27.1	24.3	26.5	21.9
FROG L1-38	27.4	29.7	26.1	27.5	27.2	28.0	27.3	28.0		31.3	33.0	34.8
FROG L1-39	27.5	27.2	26.0	27.8	28.9	25.0	27.9	27.8	32.1		35.8	29.8
FROG L1-46	28.3	27.1	27.1	28.2	27.6	27.0	26.6	29.3	34.0	32.8		55.8
FROG L1-47	28.1	28.7	25.2	28.5	29.5	24.8	27.8	28.0	33.4	34.5	42.5	

(continued)

Table 1
Continued

	FISH L1-12A	FISH L1-12B	FISH L1-1	FISH L1-1B	FISH L1-1D	FISH L1-6	FISH L1-7B	FISH L1-7C	FISH L1-8	FISH L1-10B	FISH L1-11Aa	FISH L1-13A	FISH L1-13B	FISH L1-13C	FISH L1-13D	FISH L1-16B	FISH L1-17B
FISH L1-12A		47.5	25.9	29.8	33.5	29.3	27.3	29.2	26.4	29.5	20.7	26.5	26.3	32.0	31.1	30.6	26.7
FISH L1-12B	48.3		33.1	27.9	33.5	27.1	30.7	32.0	27.5	32.4	16.6	32.6	27.9	37.0	32.8	33.7	29.5
FISH L1-1	33.4	33.6		42.0	25.0	28.8	33.9	32.8	25.1	29.9	19.5	33.3	30.9	36.2	30.5	39.7	35.2
FISH L1-1B	30.7	31.5	46.4		29.1	23.8	34.1	36.1	29.3	26.1	20.2	35.0	33.0	37.8	35.0	41.2	27.2
FISH L1-1D	34.4	34.2	37.2	34.6		33.7	33.9	30.9	30.2	40.7	17.1	34.5	26.3	36.2	33.3	29.9	29.9
FISH L1-6	27.6	29.1	31.2	28.8	31.7		29.4	27.4	21.8	34.5	20.2	29.1	30.0	33.0	30.9	31.7	33.1
FISH L1-7B	31.8	31.7	49.6	46.0	37.7	33.3		68.2	30.7	29.4	21.2	37.3	36.0	40.1	40.1	33.7	35.2
FISH L1-7C	32.6	30.8	48.9	46.0	36.5	31.6	60.1		31.8	30.5	20.0	37.0	36.1	42.5	41.6	35.0	33.9
FISH L1-8	29.6	30.0	34.6	31.3	33.1	32.8	32.4	32.3		26.8	16.8	31.8	28.9	31.8	26.3	27.8	26.1
FISH L1-10B	31.4	30.2	37.3	35.0	34.7	29.9	36.1	35.3	32.2		19.8	32.8	33.5	32.2	32.8	28.1	35.4
FISH L1-11Aa	30.4	30.6	36.1	33.1	35.3	33.8	35.5	35.0	33.5	34.2		17.8	20.6	19.5	18.3	17.5	19.8
FISH L1-13A	31.3	30.7	45.8	44.0	35.4	27.8	46.3	44.4	31.4	34.7	35.6		37.0	61.1	64.2	34.6	31.1
FISH L1-13B	32.6	30.5	48.3	43.5	36.6	30.9	48.1	46.9	32.0	36.3	35.1	49.0		35.8	35.4	33.7	31.5
FISH L1-13C	31.5	30.4	46.0	45.4	36.5	30.0	47.8	46.0	32.2	35.7	34.1	55.3	51.0		56.8	35.2	32.8
FISH L1-13D	31.3	30.9	46.7	43.4	36.6	30.6	45.4	45.7	31.9	34.7	34.2	58.6	49.4	54.2		33.7	31.6
FISH L1-16B	32.5	31.9	41.3	38.8	36.6	32.0	40.6	40.7	34.6	35.7	36.3	38.8	39.4	38.8	38.7		33.1
FISH L1-17B	28.8	28.8	32.8	30.5	33.4	42.6	32.3	31.3	32.2	31.7	32.7	30.4	31.9	29.9	31.6	31.4	

Note.—ORF1 above diagonal and ORF2 below.

Table 2
Length and GC Composition of L1 Families in Mammals, Lizard, Frog, and Fish

			Total	5′UTR	ORF1	IGR	ORF2	3′UTR
Mammals		Length	7,144 ± 894	1,471 ± 581	1,011 ± 82	172 ± 222	3,836 ± 22	654 ± 638
			[6,020–9,646]	[906–3,229]	[891–1,158]	[26–719]	[3,807–3,882]	[148–2,751]
		% GC	42.6 ± 1.6	57.2 ± 5.8	39.1 ± 2.2	36.9 ± 4.1	37.9 ± 1.3	46.3 ± 2.5
			[39.2–45.3]	[43.3–63.1]	[35.9–44.1]	[29.7–42.3]	[35.5–39.2]	[40.7–49.5]
Lizard	Clade 1	Length	6,435 ± 165	1,268 ± 198	1,066 ± 21	24 ± 48	3,760 ± 52	345 ± 118
			[6,151–6,703]	[792–1,465]	[762–1,125]	[0–110]	[3,645–3,813]	[194–570]
		% GC	36.4 ± 1.0	45.2 ± 2.9	37.3 ± 2.8	34.5 ± 5.7	33.5 ± 1.1	33.1 ± 5.8
			[34.8–37.8]	[39.5–49.7]	[34.7–43.8]	[30.5–38.5]	[31.7–35.0]	[22.5–40.2]
	Clade 2	Length	5,381 ± 139	229 ± 11	981 ± 193	310 ± 64	3,737 ± 17	125 ± 29
			[5,234–5,510]	[216–238]	[762–1,125]	[270–384]	[3,723–3,756]	[96–154]
		% GC	35.1 ± 1.2	44.5 ± 1.2	39.7 ± 2.1	46.1 ± 5.7	32.5 ± 0.7	32.6 ± 6.7
			[33.6–35.9]	[43.1–45.4]	[37.3–40.9]	[39.6–50.0]	[31.7–33.0]	[25.3–38.5]
Frog		Length	5,712 ± 234	142 ± 18	999 ± 83	598 ± 263	3,735 ± 159	239 ± 81
			[5,470–6,340]	[91–162]	[873–1,113]	[257–1,032]	[3,225–3,855]	[98–365]
		% GC	44.5 ± 3.3	55.2 ± 4.4	50.5 ± 4.0	43.7 ± 4.6	43.3 ± 3.2	35.0 ± 4.9
			[38.3–48.1]	[44.0–59.5]	[42.7–54.2]	[33.7–50.1]	[37.5–47.0]	[25.9–42.3]
Fish		Length	5,794 ± 404	165 ± 42	877 ± 70	536 ± 132	3,773 ± 71	448 ± 471
			[5,380–7,293]	[113–267]	[780–1,059]	[311–730]	[3,516–3,837]	[167–2,157]
		% GC	36.9 ± 1.3	41.3 ± 3.6	47.1 ± 2.9	33.7 ± 3.2	35.9 ± 1.5	24.7 ± 4.6
			[33.9–39.2]	[36.0–47.5]	[41.0–52.8]	[27.7–40.4]	[33.3–38.6]	[16.7–34.0]

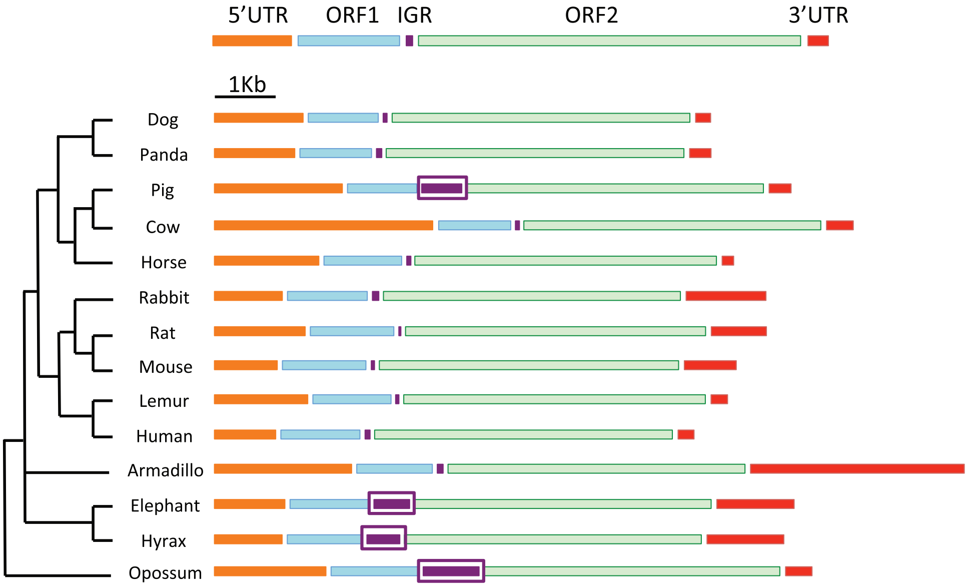

Fig. 4.—Evolution of the mammalian IGR. The figure suggests that the ancestor of mammals had an IGR that was lost after the split between afrotheria (elephant and hyrax) and other mammals and that an IGR was regained in pig. The branch-lengths on the phylogeny are not up to scale.

5′UTR tend to be enriched in GC, relative to other regions of L1. This is particularly true in mammals, with an average GC content of 57.2% in the 5′UTR. The mammalian 3′UTR is also GC-rich (46.3%) when compared with other vertebrates, which have remarkably low GC content in this region of L1 (from 24.7% in fish to 35.0% in frog). Since mammals have on average longer UTRs than other vertebrates, the high GC-content at the extremities of the elements contributes significantly to the higher GC content of mammalian L1 relative to lizard and fish.

There are also remarkable differences in the GC content of the ORFs (table 2). The GC-content of ORF1 and ORF2 are significantly different among vertebrates (ANOVA; $F_{3,\,51} = 70.61$; $P < 0.00001$ for ORF1; $F_{3,\,51} = 79.31$; $P < 0.00001$ for ORF2). In mammals and lizard, both ORFs show a considerable enrichment in adenine (42.2% on average in ORF2 and 43.4% in ORF1), which is observed at the three codon positions (fig. 5). In frog and zebrafish, adenine also tends to be more frequent than the other three bases (33.7% in ORF2 and 32.1% in ORF1), yet the difference is not as pronounced as in mammals and lizard, resulting in an overall higher GC content of the ORFs (table 2). Zebrafish ORF2 is unique because it is enriched for both adenine and thymine. It can be noted that within each species the base composition, and in particular the frequency of adenine, is strikingly similar at all codon positions. In all vertebrates, however, the GC content of ORF1 is significantly higher than ORF2 ($P < 0.05$ for all species using *t*-test; table 2), the largest difference being found in zebrafish (ORF1 = 47.1%; ORF2 = 35.9%).

Considering the differences in base composition in the ORFs, we decided to examine how this relates to codon usage. Table 3 shows the codon usage for ORF1 and ORF2 in all taxa as well as an estimator of the bias for each codon (RSCU). With very few exceptions, when a codon with an adenine at the third position is available, it will be the codon most frequently used. This is true for both ORFs and in all taxa. This was further confirmed by calculating two estimators of codon usage bias, CAI and Nc. CAI compares the codon usage of a gene of interest (ORF1 and ORF2 in our case) with the codon usage of the host's genome whereas Nc estimates how the codon usage differs from equal usage of synonymous codons (table 4). The elevated values of Nc are consistent with a substantial codon usage bias, yet none of the values of CAI were significantly different from expectation given the codon usage of the host and the GC content of the genes. This is not really surprising since the enrichment in adenine is found at all three codon positions.

We examined how differences in base composition affect the amino acid composition of the ORFs (fig. 6). Two observations can be made. First, mammalian and lizard L1 are enriched in lysine and glutamic acid, two amino acids encoded by A-rich codons (AAA and AAG for lysine and GAA and GAG for glutamic acid), and for both amino acids it is the A-rich codon that is strongly preferred (AAA and GAA; table 3). Second, the frog and fish L1s contain a higher proportion of three amino acids encoded by codons that are less likely to contain an A: alanine (GCA, GCC, GCG, GCU), proline (CCA, CCC, CCG, CCU), and Serine (AGC, AGU, UCA, UCC, UCG, UCU). These differences in amino acid composition are observed for both ORFs and are thus unlikely to result from selection on the function of the proteins.

One of the consequences of an enrichment in A-rich codon is the potential formation of premature polyadenylation signal and thus inefficient transcription (Perepelitsa-Belancio and Deininger 2003). We estimated the average number of canonical and non-canonical (AATAAA, ATTAAA) poly-adenylation signals in the ORFs. As expected, the number of potential premature poly-adenylation signals in the ORFs is larger in mammals (16.3 on average) and lizard (25.0) than it is in fish (12.2) and frog (7.5), suggesting that the transcription of L1 might be more efficient in frog and fish than in amniotes (supplementary material S2, Supplementary Material online).

Evolution of the 5′UTR

Previous work in mammals has shown that L1 has the ability to recruit novel 5′UTR, possibly to bypass host repression of transcription (Adey, Schichman, et al. 1994; Khan et al. 2006; Sookdeo et al. 2013). We thus decided to examine how common the replacement of 5′UTR across vertebrates is. As reported above, 5′UTRs fall in two categories: the long 5′UTR of mammals and lizard clade 1 and the short 5′UTR of fish, frog and lizard clade 2. These differences reflect an L1-specific evolutionary trend, since there are no substantial differences in the length of 5′UTRs among eukaryotes (Mignone et al. 2002) Based on the phylogeny of L1 (fig. 3), we can infer that the ancestral state is most likely a short 5′UTR and that a long 5′UTR evolved independently twice, in the ancestor of all mammals and in the anole lineage.

Using DOTMATCHER we compared the 5′UTRs of mammals but we failed to find any significant similarity among them, which is consistent with the rapid 5′UTR turnover described in primates and rodents (Adey, Schichman, et al. 1994; Khan et al. 2006; Sookdeo et al. 2013). Despite the absence of homology among 5′UTRs, almost all mammalian L1s begin with a sequence of consensus $G_{2\text{-}6}$(A/C)G_2AGNCA AGATGGCGGA, the motif CAAGATGGC corresponding to a YY1 transcription factor binding site which is critical for transcription initiation (Athanikar et al. 2004). The only exceptions are the mouse and rat elements which do not start with the YY1 binding site but it was shown that the monomers constitutive of their 5′UTR contain signals for transcription initiation (Adey, Tollefsbol, et al. 1994). Mammalian 5′UTRs have very high GC content (from 51.9% to 61.6%), with two notable exceptions: the cow (43.3%) and the armadillo (46.4%), which also happen to have the longest 5′UTRs in mammals (3,229 and 2,029 bp, respectively). All mammalian 5′UTRs are enriched in CpG dinucleotides, which are forming CpG islands. The average number of CpG dinucleotides is 62.1 and varies from 31 in opossum to 94 in horse. The region of the 5′UTR that fits the definition of a CpG island always reside at the 5′ extremity of the UTR, with the exception of the cow (whose CpG island begins ~250 bp from the 5′ extremity). A number of 5′UTRs contain motifs (~70–100 bp) that are tandemly duplicated two (rat, hyrax), three (horse, elephant, opossum), or many (mouse) times (fig. 7). Other elements are dramatically enriched in G-rich (pig) or T-rich (dog, cow, armadillo) low-complexity repeats, the cow presenting the most extreme examples since it contains a ~840-bp region composed exclusively of T-rich short repeats (fig. 7).

Though similar in length, the long 5′UTRs of the lizard clade 1 differ from the mammalian 5′UTRs in several respects. The

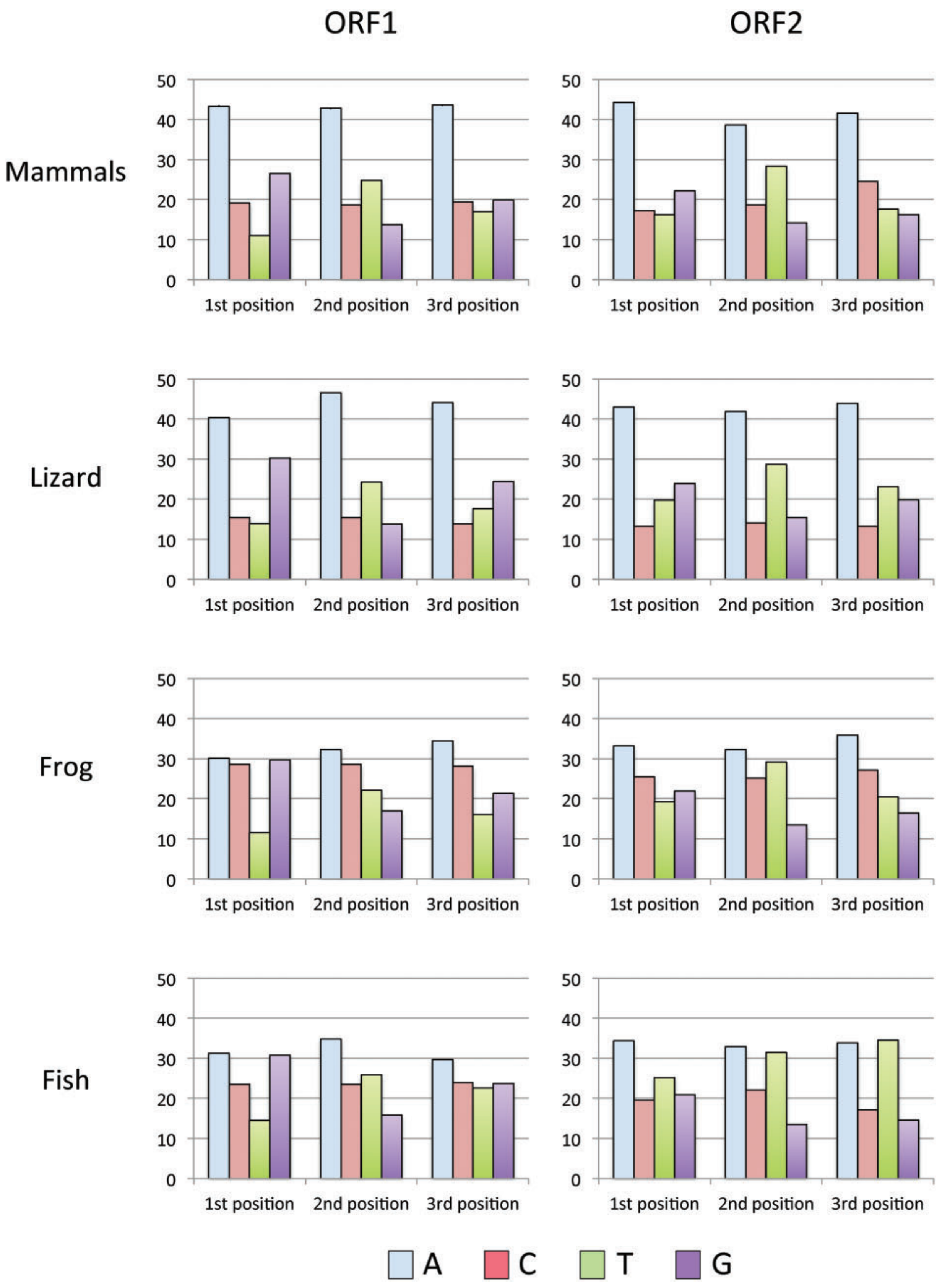

Fig. 5.—Base composition at the three codon positions for ORF1 and ORF2.

Table 3
Codon frequency and Relative Synonymous Codon Usage (RSCU) for ORF1 and ORF2 in Mammals, Lizard, Frog, and Fish

AA	Codon	ORF1								ORF2							
		Mammals		Lizard		Frog		Fish		Mammals		Lizard		Frog		Fish	
		FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU
Ala	**GCA**	22.6	2.0	21.2	2.1	33.1	1.5	20.0	1.2	25.1	2.3	19.5	2.1	23.5	1.5	17.8	1.6
Ala	GCC	11.8	1.0	7.5	0.8	26.4	1.2	18.1	0.9	11.5	1.0	7.2	0.8	21.5	1.3	8.6	0.8
Ala	GCG	2.2	0.2	4.6	0.5	15.6	0.7	17.5	0.8	1.0	0.1	4.5	0.5	5.3	0.3	2.1	0.2
Ala	GCU	8.8	0.8	7.7	0.7	12.6	0.6	21.3	1.1	7.1	0.6	6.3	0.7	14.0	0.9	15.3	1.4
Arg	**AGA**	43.1	3.3	50.4	3.9	31.9	2.0	21.7	1.5	35.3	4.0	35.1	3.8	19.3	2.4	20.0	2.7
Arg	AGG	22.6	1.7	14.6	1.1	16.6	1.0	10.5	0.7	11.9	1.3	15.8	1.7	11.9	1.5	8.4	1.1
Arg	**CGA**	5.2	0.4	4.0	0.3	9.5	0.6	13.9	1.0	2.9	0.3	1.4	0.2	5.0	0.6	4.2	0.6
Arg	CGC	2.4	0.2	1.8	0.1	19.3	1.2	15.6	1.1	1.3	0.1	0.6	0.1	5.6	0.7	3.9	0.6
Arg	CGG	3.6	0.3	4.6	0.4	11.0	0.7	10.5	0.7	1.1	0.1	2.3	0.2	3.6	0.4	2.1	0.3
Arg	CGU	2.8	0.2	3.3	0.2	6.5	0.4	14.5	1.1	0.6	0.1	1.1	0.1	2.8	0.3	6.0	0.8
Asn	AAC	33.4	1.0	19.0	0.9	26.4	1.2	22.9	1.1	35.6	1.1	21.2	0.6	29.5	1.2	19.9	0.6
Asn	AAU	34.0	1.0	23.7	1.1	17.6	0.8	18.7	0.9	30.1	0.9	47.5	1.4	19.7	0.8	41.1	1.4
Asp	GAC	24.2	1.0	30.1	1.0	33.4	1.2	32.0	1.1	26.4	1.2	18.3	0.8	24.9	1.2	17.9	0.8
Asp	GAU	22.2	1.0	29.2	1.0	20.8	0.8	27.1	0.9	19.0	0.8	29.1	1.2	16.3	0.8	29.1	1.2
Cys	UGC	2.8	1.1	1.5	0.7	3.5	0.9	3.4	0.9	8.4	1.3	3.5	0.7	7.4	1.3	5.9	0.7
Cys	UGU	1.0	0.5	2.2	0.7	1.3	0.2	3.0	0.9	4.6	0.7	7.1	1.3	4.5	0.7	10.1	1.3
Gln	**CAA**	35.6	1.1	40.9	1.4	39.7	1.3	26.1	1.1	25.3	1.4	30.5	1.4	35.0	1.4	24.1	1.3
Gln	CAG	26.6	0.9	19.9	0.6	22.3	0.7	20.0	0.9	10.5	0.6	11.9	0.6	14.3	0.6	13.8	0.7
Glu	**GAA**	78.3	1.4	86.0	1.3	41.9	1.2	53.0	1.3	50.6	1.5	57.1	1.5	29.4	1.4	29.8	1.4
Glu	GAG	32.4	0.6	49.1	0.7	27.1	0.8	31.3	0.7	17.1	0.5	19.7	0.5	12.8	0.6	12.8	0.6
Gly	**GGA**	13.0	1.7	13.7	1.6	13.1	1.1	13.5	1.4	18.1	2.0	17.9	1.6	11.5	1.2	12.3	1.4
Gly	GGC	4.8	0.6	4.9	0.6	16.3	1.4	11.4	1.2	7.7	0.8	6.7	0.6	10.7	1.1	7.4	0.8
Gly	GGG	8.0	1.1	10.0	1.2	10.3	0.9	7.1	0.7	6.1	0.7	10.9	1.0	9.1	0.9	5.6	0.6
Gly	GGU	5.0	0.6	5.5	0.6	6.8	0.6	7.6	0.8	5.1	0.5	8.9	0.8	7.9	0.8	11.6	1.2
His	CAC	7.4	1.0	4.6	0.9	15.3	1.4	12.6	1.3	14.0	1.3	5.3	0.7	18.6	1.2	9.5	0.7
His	CAU	6.4	1.0	5.8	1.1	7.0	0.6	6.9	0.7	8.4	0.7	8.8	1.3	11.5	0.8	16.0	1.3
Ile	**AUA**	42.7	1.5	31.4	1.5	20.6	1.3	13.5	0.7	47.3	1.3	51.2	1.6	39.3	1.5	32.2	1.1
Ile	AUC	20.0	0.7	11.7	0.5	16.8	1.0	22.5	1.2	35.9	1.0	14.1	0.4	19.3	0.7	15.6	0.5
Ile	AUU	20.6	0.8	20.8	1.0	13.1	0.8	20.6	1.1	27.2	0.7	33.3	1.0	18.9	0.7	39.4	1.4
Leu	**CUA**	22.8	1.7	13.7	1.0	28.6	1.9	12.8	0.8	23.9	1.6	19.6	1.2	36.0	1.7	17.0	0.9
Leu	CUC	12.4	0.9	7.1	0.5	16.8	1.2	16.4	1.1	18.6	1.3	5.8	0.3	23.0	1.1	11.2	0.6
Leu	CUG	14.2	1.1	17.9	1.2	20.1	1.4	22.9	1.4	16.0	1.1	10.1	0.6	22.1	1.0	12.6	0.7
Leu	CUU	11.4	0.9	9.7	0.6	9.5	0.6	16.8	1.1	9.1	0.6	8.3	0.5	16.5	0.8	24.9	1.3
Leu	**UUA**	13.6	1.0	22.3	1.6	8.8	0.6	10.7	0.7	13.8	0.9	38.3	2.3	20.6	1.0	34.6	1.8
Leu	UUG	5.8	0.4	15.3	1.0	6.5	0.4	13.9	0.9	7.3	0.5	15.9	1.0	8.9	0.4	15.2	0.8
Lys	**AAA**	75.7	1.3	92.0	1.4	35.7	1.3	54.7	1.4	86.0	1.5	98.1	1.5	53.6	1.5	59.9	1.5
Lys	AAG	35.6	0.7	40.5	0.6	17.1	0.7	24.2	0.6	31.5	0.5	32.3	0.5	17.1	0.5	20.8	0.5
Met	AUG	25.6		31.6		17.1		22.9		21.0		23.3		17.5		16.5	
Phe	UUC	16.6	1.2	11.3	0.9	15.3	1.2	17.3	1.0	21.7	1.2	10.1	0.6	15.6	0.9	14.4	0.5
Phe	UUU	9.8	0.8	14.8	1.1	9.5	0.8	19.1	1.0	14.7	0.8	23.6	1.4	17.9	1.1	42.0	1.5
Pro	**CCA**	22.8	2.2	10.2	1.9	31.9	1.6	15.8	1.4	22.4	2.2	14.2	2.1	27.4	1.7	21.1	1.7
Pro	CCC	9.6	0.9	3.8	0.7	18.8	0.9	8.4	0.7	8.8	0.9	5.9	0.8	20.9	1.3	10.4	0.8
Pro	CCG	0.6	0.1	2.0	0.4	13.3	0.7	10.1	0.9	1.7	0.2	3.4	0.5	4.0	0.2	2.1	0.2
Pro	CCU	7.8	0.8	4.6	0.9	15.8	0.8	11.2	1.0	8.6	0.8	4.3	0.6	12.4	0.8	16.8	1.3
Ser	AGC	9.8	1.1	3.5	0.4	11.8	1.3	13.7	1.2	7.9	1.0	3.9	0.6	8.6	0.8	6.1	0.4
Ser	AGU	7.6	0.9	8.8	1.1	3.8	0.4	6.9	0.6	6.9	0.8	7.8	1.2	5.2	0.5	11.3	0.8
Ser	**UCA**	19.8	2.2	16.8	2.1	11.8	1.2	14.3	1.2	19.5	2.3	14.1	2.1	16.2	1.5	30.9	2.0
Ser	UCC	8.2	0.8	4.4	0.6	15.3	1.6	10.1	0.8	8.2	1.0	5.5	0.8	19.7	1.8	12.8	0.8
Ser	UCG	2.0	0.2	6.0	0.8	4.8	0.5	12.8	1.1	1.7	0.2	3.7	0.6	3.7	0.4	3.2	0.2
Ser	UCU	6.8	0.7	8.0	1.0	10.5	1.0	14.5	1.1	6.0	0.7	5.5	0.8	10.9	1.0	27.3	1.8

(continued)

Table 3 Continued

AA	Codon	ORF1								ORF2							
		Mammals		Lizard		Frog		Fish		Mammals		Lizard		Frog		Fish	
		FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU	FREQ	RSCU
Thr	**ACA**	29.8	1.9	27.6	2.0	26.9	1.3	16.4	1.1	36.1	2.2	25.7	2.2	27.7	1.6	23.2	1.7
Thr	ACC	17.0	1.1	10.6	0.8	22.1	1.2	13.0	0.8	15.0	0.9	6.5	0.6	21.5	1.2	9.6	0.7
Thr	ACG	5.0	0.3	7.1	0.5	12.8	0.7	15.2	1.0	4.1	0.2	5.5	0.5	6.2	0.3	2.8	0.2
Thr	ACU	12.2	0.8	11.3	0.8	15.1	0.8	16.8	1.1	10.8	0.7	8.7	0.7	15.6	0.9	17.0	1.3
Trp	UGG	5.8		9.7		7.0		4.0		23.7		30.7		24.2		19.9	
Tyr	UAC	7.8	1.0	11.3	0.8	12.3	1.3	11.6	1.1	18.3	1.1	13.5	0.7	20.3	1.0	11.6	0.7
Tyr	UAU	7.2	1.0	13.3	1.2	6.3	0.7	7.1	0.9	14.0	0.9	25.5	1.3	18.6	1.0	23.3	1.3
Val	**GUA**	10.2	1.2	9.1	1.1	11.5	1.2	9.1	0.7	9.7	1.4	15.7	1.9	13.8	1.6	10.8	1.1
Val	GUC	7.4	0.9	5.8	0.6	11.5	1.2	10.1	0.9	5.6	0.8	4.0	0.5	9.1	1.0	6.0	0.7
Val	GUG	8.4	1.0	12.8	1.4	10.8	1.1	16.4	1.3	7.8	1.1	8.3	1.0	7.2	0.8	8.2	0.8
Val	GUU	6.2	0.8	7.1	0.8	5.0	0.5	13.9	1.1	4.3	0.6	5.7	0.6	5.3	0.6	13.9	1.4

Note.—The most frequent codon and the highest RSCU for each amino acid are highlighted in grey. Codons with an A at the third position are in bold.

Table 4

CAI, the Effective Number of Codons, and the Average GC Content at the Three Codon Positions

	ORF1					ORF2				
	CAI	Nc	GC1	GC2	GC3	CAI	Nc	GC1	GC2	GC3
Mammals	0.728 ± 0.017	48.67 ± 4.33	45.6 ± 2.5	32.4 ± 3.2	39.4 ± 3.1	0.734 ± 0.009	45.30 ± 1.68	39.5 ± 1.2	32.9 ± 1.0	40.7 ± 2.2
Lizard	0.756 ± 0.014	47.28 ± 2.72	45.7 ± 3.7	29.2 ± 3.5	38.3 ± 2.9	0.719 ± 0.013	44.79 ± 3.07	37.3 ± 1.4	29.4 ± 0.8	33.0 ± 2.4
Frog	0.783 ± 0.020	55.08 ± 2.22	58.3 ± 4.4	45.5 ± 3.0	49.6 ± 4.8	0.789 ± 0.001	51.63 ± 2.11	47.5 ± 3.6	38.6 ± 2.5	43.7 ± 4.0
Fish	0.745 ± 0.020	56.88 ± 3.85	54.3 ± 3.5	39.3 ± 3.3	47.7 ± 4.3	0.733 ± 0.008	49.08 ± 1.59	40.5 ± 2.2	35.6 ± 2.0	31.7 ± 1.9

lizard 5′UTRs are more conserved across families than mammalian 5′UTRs (with the exception of family L1_AC9), although lizard families are more divergent than mammalian families in their ORFs. The first ~120 bp of the elements are the most conserved, with a number of motifs that are found across all families (fig. 8). Similar to mammals, the 5′ extremity is remarkably conserved among elements with a consensus sequence GACTTCCGGTGN$_8$ATGGCG. Lizards 5′UTRs have a significantly lower GC content (45.2% vs. 57.2% in mammals; $t = 4.957$, $P < 0.001$) and the presence of two CpG islands separated by ~300–400 bp, instead of a single one in mammals. The number of CpG is however similar to mammals with an average of 56 CpGs in lizard. None of the lizard 5′UTRs shows sign of tandem duplication, nor do they contain regions enriched in low-complexity repeats. As mentioned above, the 5′UTR of L1_AC9 shows no similarity with other lizard 5′UTRs, and probably results from the acquisition of a novel promoter, as occurs frequently in mammals. It should be noted though that the L1_AC9 5′UTR is remarkable among L1 since it has the lowest GC content (39.5%) of all elements analyzed here, it does not have a CpG island and it contains an extremely small number of CpG dinucleotides (13), given its length (1,352 bp).

The small 5′UTRs of the lizard clade 2, frog and fish do not have much in common. Although similar in length, these 5′UTRs differ substantially in GC content, the frog 5′UTR being more GC-rich (55.2%) than the fish (41.3%) and lizard clade 2 (44.5%) 5′UTRs, but the number of CpG dinucleotides is similar among species with ~7 CpG on average. In all three species, the 5′ extremity of the 5′UTR is extremely conserved across families, with consensus sequences GGGNGCTGCGCATGC, GGGGGCG TGGCC and GGACTTCCGGTT in lizard, frog, and zebrafish, respectively (fig. 8). A search for transcription factor binding sites revealed that the start of the zebrafish L1s corresponds to the canonical target sequence of the XrpFI transcription factor whereas the 5′ end of the lizard and frog elements show similarity to the Sp1 transcription factor binding site. In lizard, the first 100 bp is relatively conserved among families and all frog families share two conserved motifs (A/G)GACGC(G/A) and GA GCTCCG, located about 30 and 40 bp from the start of the element, respectively. In zebrafish, we failed to find any similarity among 5′UTRs past the very beginning of the elements.

Evolution of ORF1

ORF1 has recently attracted the attention of researcher in the field of L1 biology because the function of ORF1p remains

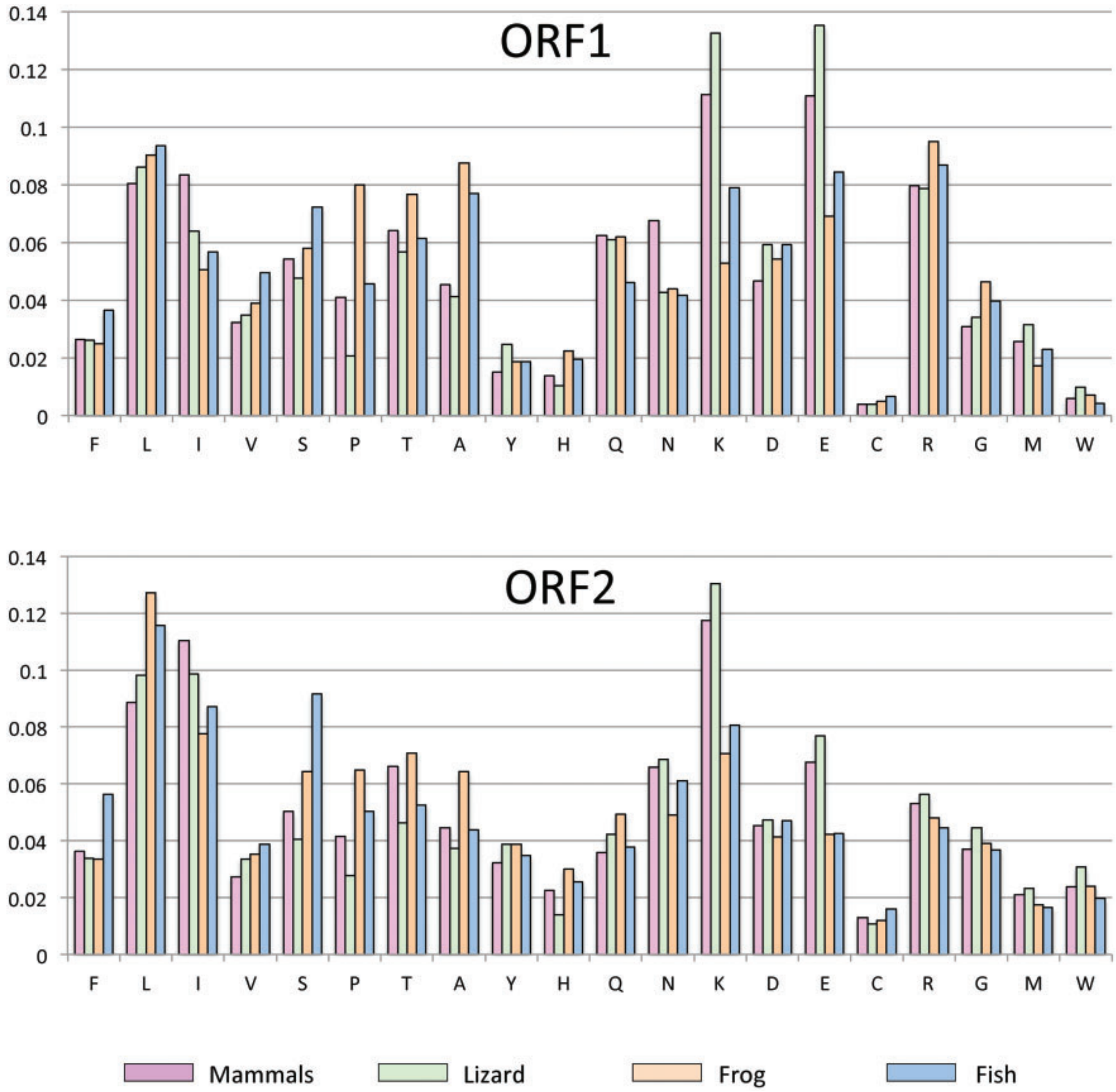

Fig. 6.—Frequency of amino acids in ORF1 and ORF2 for mammals, lizard, frog, and zebrafish.

incompletely understood and because a region of ORF1, the CCD, is particularly unstable at the sequence level (Furano 2000; Boissinot and Furano 2001; Sookdeo et al. 2013). It was suggested that this instability might reflect an antagonistic arms race between L1 and its host. In all species examined here, the general structure of ORF1 is conserved and includes a CCD, a non-canonical RRM, and a C terminal domain (CTD). The CC domain is located 2 to ~123 amino acid from the N terminus and is highly variable in sequence, length and structure. Conversely the RRM and the CTD are very conserved among L1 families as attested by high identity value, similar to the identity reported for ORF2 (table 1). A number of residues and motifs are conserved across all elements, including two non-canonical RNA-recognition motifs (the blue boxes on fig. 9), the three amino acids forming the conserved salt bridge that stabilize the structure of ORF1p (orange arrows) and the residues providing RNA-binding side chains (green arrows) (Khazina and Weichenrieder 2009). In addition, several motifs involved in the phosphorylation of ORF1p (Cook et al. 2015) are conserved, although never across all families (fig. 9). The putative PDPK docking motif at the start of the RRM and the PP1 docking motif in the center of the CTD are conserved in mammals, lizard clade 2, frog and fish but not in lizard clade 1. Lizard L1s appear to have an additional PDPK docking motif that is absent from all other species. Other PDPK and PKA docking motifs shown to be important in human (Cook et al. 2015) are not predicted to act as phosphorylation docking sites in other vertebrates or even other mammals.

The N-terminal region of ORF1 and the CCD are extremely variable in length, so that it was not possible to obtain a

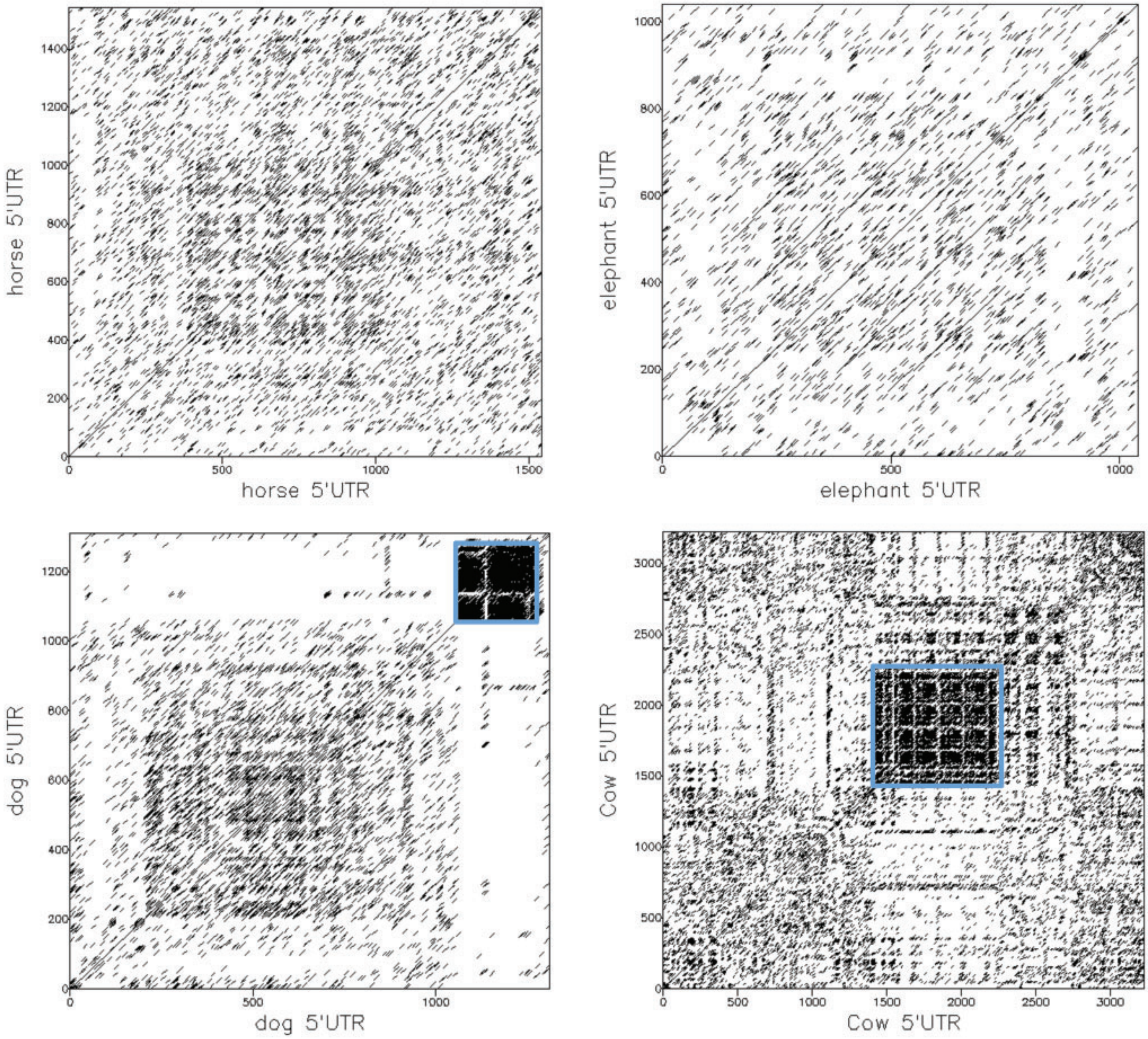

Fig. 7.—Dotmatcher analysis of the horse, elephant, dog, and cow 5′UTR against themselves. Note the long tandem duplication in horse and elephant and the repeats rich region of the dog and cow 5′UTRs (framed with blue boxes).

reliable alignment for this region. Yet, several observations regarding the general structure of the CCD can be made. First, the structure of the CCD is more complex and more variable in mammals and lizard than in fish and frog (fig. 10). All frogs and most fish elements have CCDs that consist of uninterrupted series of canonical heptads. In lizard and in mammals, the structure of the CCD is imperfect and consists of series of heptads separated by non-canonical coiled-coil forming sequences or non-coiled-coil forming group of amino acids. Second the sequence conservation of the CCD differs considerably among organisms. This region is so divergent among mammals that it is not possible to obtain a reliable alignment in this group of vertebrate. Similarly, this region is highly variable in frog and fish and an alignment could not be obtained within these two species. In contrast, the CCD of lizard's L1 is remarkably conserved and a reliable alignment can readily be generated for all elements belonging to clade 1. This strongly suggests that the selective pressure acting on this region is different among vertebrates.

We also examined the presence in the CCD of the RhxxhE motif which is often associated with parallel trimeric coiled coil (R occupying the g position of an heptad and E in position e of the following heptad; R = Arg, E = Glu, h stands for any hydrophobic residue and x for any residue) (Kammerer et al.

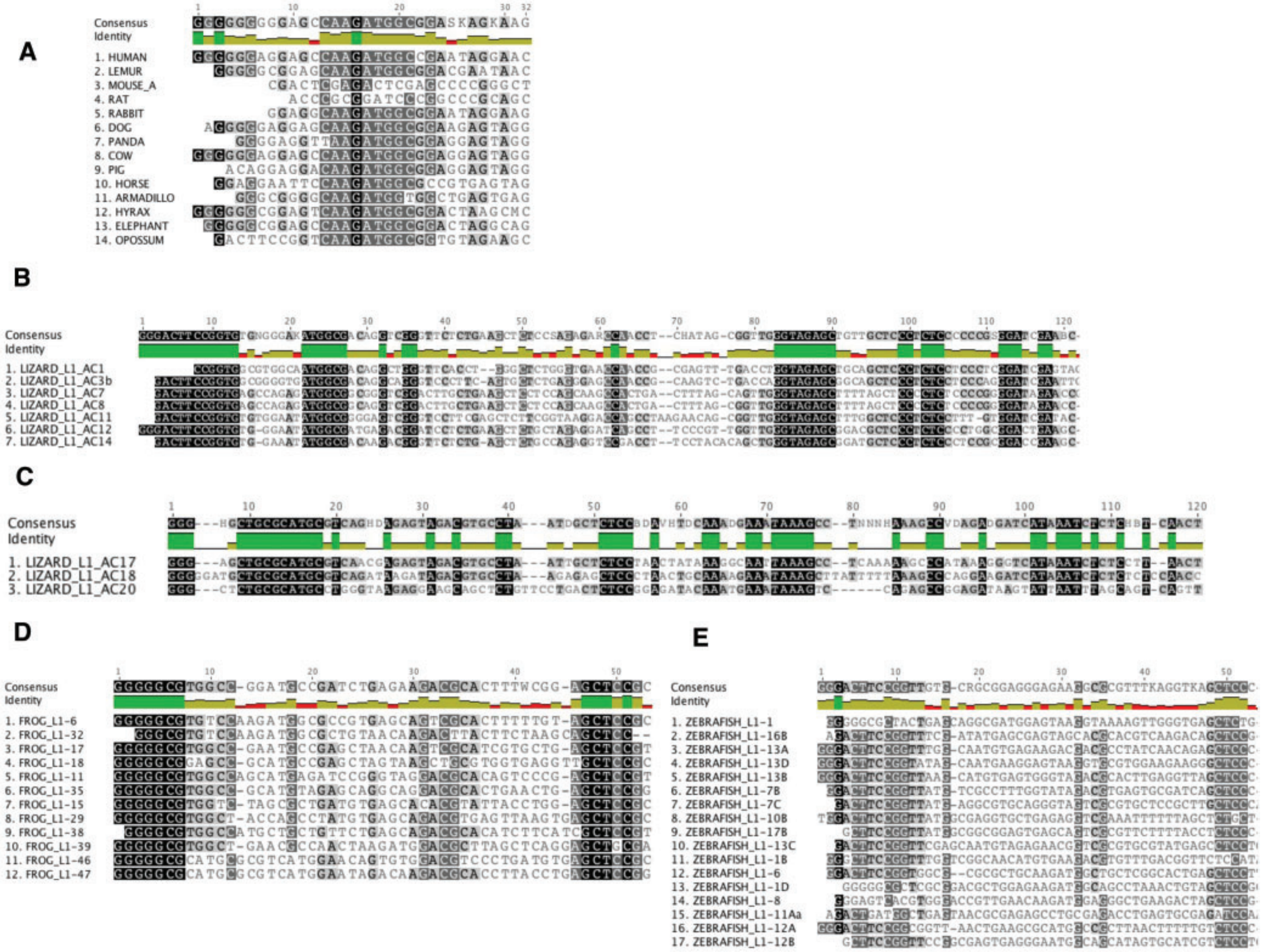

Fig. 8.—Alignment of the 5′ termini of L1 in mammals (*A*), lizard clade 1 (*B*), lizard clade 2 (*C*), frog (*D*), and zebrafish (*E*). The length of the alignments varies among groups since the length of the 5′ termini that could be aligned differed.

2005), although it can be found associated with dimeric and even tetrameric coiled coils (Xu and Minor 2009). We found that all placental L1 possess two such motifs arranged in tandem, with a third downstream motif present in some species (fig. 10). The RhxxhE motif (either isolated or in tandem) is also found in all but two frog L1 families and all but four fish families. This motif is however conspicuously rare in lizard and is found in only 2 families out of 12 families. We examined the composition of the different position of the heptads in lizard (supplementary material S3, Supplementary Material online) and we found that Arg at position g is indeed rare (with a frequency of 0.1) while Glu and Lys are the most common amino acid at this position (0.34 and 0.22, respectively). In all other species, Arg is the most common amino acid in g (0.3 in mammals to 0.4 in frog) whereas Glu and Lys are found at low frequency at this position (~0.1). Position e of the lizard L1 is occupied principally by Lys or Gln (~0.25 for each) while Glu is the most frequent amino acid in all other species (~0.3). Interestingly, mutations of Arg (to Ala or Lys) or Glu (to Ala or Leu) residues in known trimeric coiled coils were shown to produce dimeric or tetrameric structures (Kammerer et al. 2005). The near complete absence of the RhxxhE motif and the differences in the amino acid composition of the heptads in lizard suggests either that the trimeric structure of ORF1p is achieved by different means in lizard or that the lizard ORF1p does not form trimers, but tetramers or dimers.

Evolution of the ORF1-ORF2 Inter-Genic Spacer

The presence or absence of an IGR is one of the most variable structural features in L1. Most mammals and lizard clade 1 elements have no or very short IGR but the phylogenetic analysis on figure 3 suggests that the presence of a long IGR is probably the ancestral state, with independent losses in lizard clade 1 and in mammals, following the split between

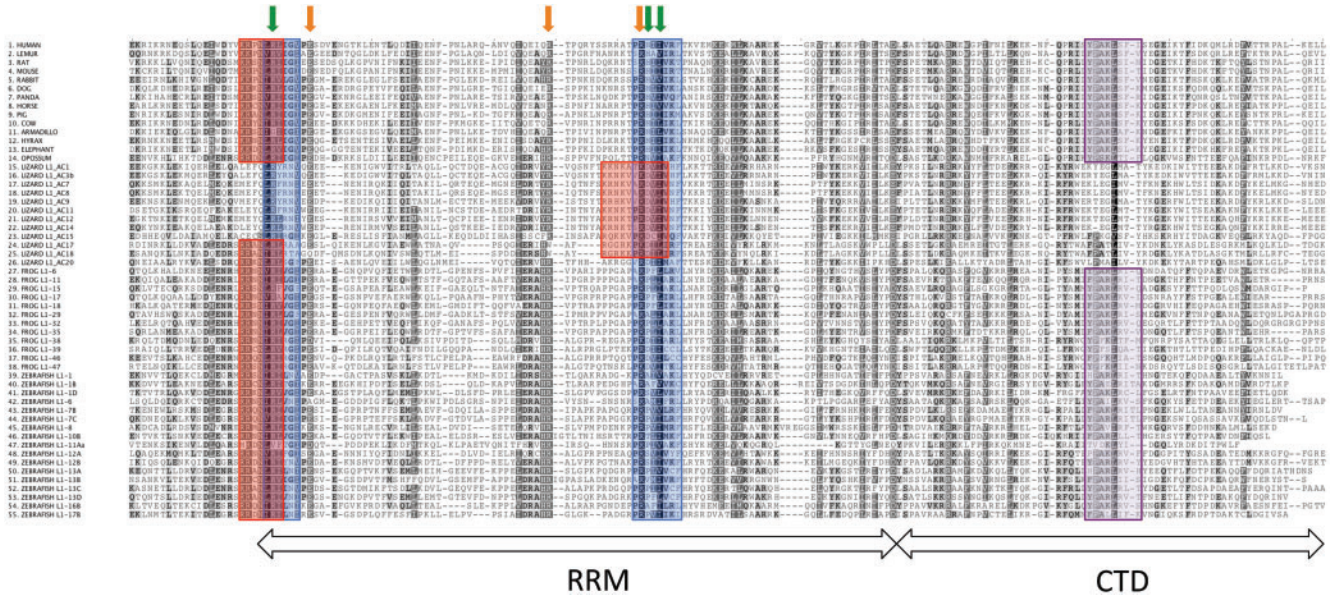

Fig. 9.—Amino acid alignment of the RRM and CTD. The two RRMs are boxed in blue, the amino acids forming the stabilizing salt bridge are indicated with orange arrows, the residues providing RNA-binding side chains are indicated with green arrows, and the PDPK docking sites are boxed in red and the PP1 docking site in purple.

Afrotheria and all other mammals. The IGR varies considerably in base composition and in sequence among organisms. In mammals, long IGRs tend to be AT-rich. The pig IGR contains several poly-A stretches and the opossum several imperfect short tandem repeats, whereas the hyrax and elephant IGR do not contain any repeats (supplementary material S4, Supplementary Material online). The long IGRs of lizard clade 2 do not contain any repeats but differ from the rest of the element by a high GC content (~46%) due to the presence of several G-rich stretches. Most frog IGRs do not contain repeats and their base composition (~44.2%) is similar to the base composition of ORF2 (~43.7%). Fish IGRs have a very low GC content (~34% on average) and all but two families contain several T-rich short repeats (supplementary material S4, Supplementary Material online).

We examined if repeats, as well as non-repeated sequences in the IGR, could be involved in the formation of secondary structure at the RNA level. The analysis of RNA secondary structure did not reveal any obvious shared patterns among IGRs, even within species. In some species, RNAfold identified stem-loop structures of various lengths but other IGRs did not show such structures (data not shown).

Because the presence of an internal ribosomal entry site (IRES) upstream of ORF2 has previously been proposed (Li et al. 2006), we examined the possibility that the IGRs contain IRES for translation of the downstream ORF2. The IRESPred webserver, which detects IRES in a sequence of interest by searching for sequence and structural features found in known nuclear and viral IRES sequence (Kolekar et al. 2016), predicts the presence of IRES in 32 out of 36 long IGR sequences. The program failed to predict the presence of IRES in only two frog and two fish elements. The nature and position of the putative IRES was further examined by the VIPS server (Hong et al. 2013), which search for structural similarity with known viral IRES. VIPS detected IRES in 31 long IGR including three of the four mammals (pig, elephant, and opossum). In all cases, similarity was detected with the IRES of cripavirus, a virus belonging to the dicistroviridae family, and the regions predicted to act as IRES were located <10 bp from ORF2 start codon in all but two elements (fig. 11 and supplementary material S4, Supplementary Material online). The dicistroviridae IRES consists of three stem-loop structures that interact directly with the 40S and 60S ribosomal subunits, without requiring protein factors (Pfingsten and Kieft 2008; Nakashima and Uchiumi 2009). Figure 11B shows an example of dicistroviridae IRES RNA structure together with the secondary structure of the predicted IRES of three L1 IGRs, which also exhibit the three stem-loop structures typical of cripavirus IRES. Despite the uncertainty associated with IRES prediction program, it is significant that two different algorithms predicted the presence of IRES, and that VIPS identified the same type of viral IRES in the same position, independently of the sequence, base composition and length of the IGR.

Evolution of ORF2

ORF2 encodes the reverse transcriptase domain necessary for retrotransposition and, not surprisingly, it is the most conserved region of L1. There is very little variation in the length

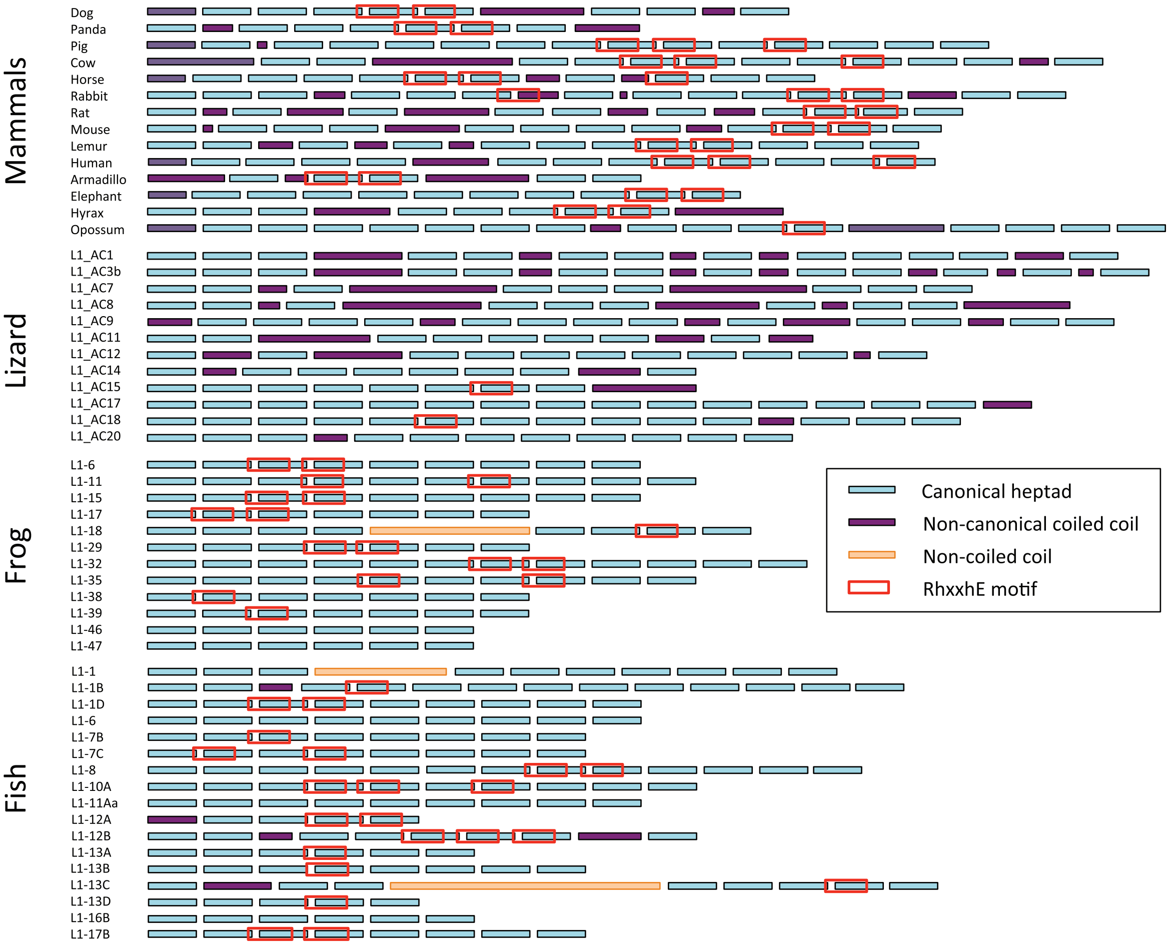

Fig. 10.—Schematic structure of the CCD of ORF1. The structure of the coiled coils is based on the analysis with a 28 residues window width.

of this ORF (3516–3882 a.a.) and all elements contain conserved endonuclease and reverse transcriptase domains. The cysteine-rich motif located at the C-terminus, which has been shown to be essential for retrotransposition (Moran et al. 1996; Doucet et al. 2010), is found in all elements, with a consensus $CX_3CX_7HX_4C$ in mammals and in a small number of non-mammalian families and $CX_2CX_8HX_4C$ in all other non-vertebrate L1s. The PCNA-interacting box located between the endonuclease and reverse transcriptase domain, which is also necessary for retrotransposition (Taylor et al. 2013), is also present in all species, with slight variation in consensus among species.

Evolution of the 3′UTR

Finally, we focused our attention on the 3′UTR, which has been shown to contain a conserved poly-purine tract of unknown function in mammalian L1 (Howell and Usdin 1997). There is considerable variation in the length of the 3′UTR but mammalian L1s tend to have longer 3′UTRs than other vertebrates (table 2). The main difference among vertebrates resides in the GC content, mammalian 3′UTRs being enriched in GC (46.3%) relative to other vertebrates (24.7% in zebrafish to 35.0% in frog). This difference is mainly due to the presence in mammals of a G-rich poly-purine tract. This poly-purine tract is surprisingly absent in rabbit, which also has the lowest GC content in mammals. With few exceptions (the anole L1AC_17 and 20 families), other vertebrate 3′UTRs lack a G-rich tract but they always contain repeated regions. In lizard and frog, these repeated regions can take the form of a C-rich poly-pyrimidine tract, of a T-rich repetitive region or of a combination of poly-C and poly-T tracts. All zebrafish 3′UTRs contain long poly-T tracts that occupy most of the length of the UTR and which can form T-rich microsatellites. In all species, L1 ends with a canonical poly-adenylation signal (AATAA A) followed by a poly-A tail.

Discussion

We identified a number of differences in the sequence of active L1 among vertebrates including (1) a stronger A bias on the positive strand in mammals and lizard than in frog and

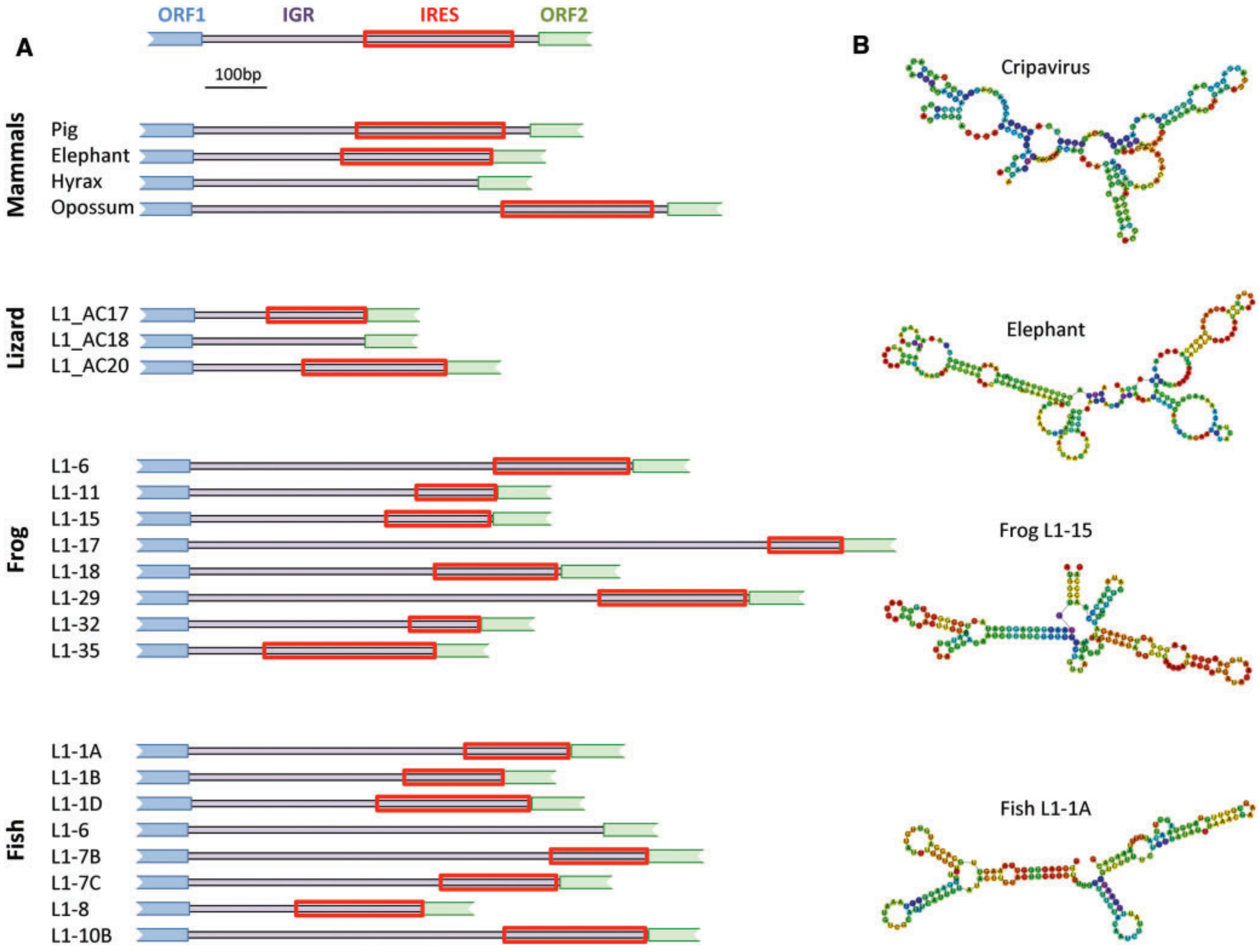

Fig. 11.—(*A*) Schematic structure of the IGR showing the position of the predicted IRES. (*B*) RNA structure of the predicted IRES of the elephant, frog L1-15 and zebrafish L1-1A compared with the IRES of a dicistroviridae, the cripavirus-1 infecting the insect *Homalodisca coagulata* (GenBank accession number KT207917).

fish, (2) the independent evolution of long GC-rich 5′UTRs in amniotes, (3) the loss of the IGR in amniotes, (4) species-specific repeated motifs in the 3′UTR, (5) a higher level of evolutionary conservation of the 5′UTR and ORF1 in non-mammalian vertebrates than in mammals. Although our sampling of mammalian genomes is probably representative of L1 diversity in this vertebrate class, the same is not true for other vertebrate lineages (reptiles, amphibians, fish), which are represented by a single model species. Comparative studies have shown that the profile of diversity and abundance of transposable elements differs greatly among fish, amphibians and reptiles species (Castoe et al. 2011; Chalopin et al. 2015; Sun et al. 2015) and additional studies on other representatives of these groups will be necessary to determine if the evolutionary trends we describe here apply widely across non-mammalian vertebrates. Preliminary analyses of the few teleostean consensus deposited in Repbase (salmon, medaka, fugu) suggest that L1 in these species share the structure and base composition as the zebrafish L1 (unpublished observations).

Functional Implications

In all species examined here, we found that the ORFs are enriched in adenine (and thymine in zebrafish ORF2) at all three positions of codons, resulting in the use of sub-optimal codons for translation and a biased amino acid composition of ORF1p and ORF2p. The compositional bias of L1 is similar to the bias reported in lentiviral retroviruses, which have adenine-rich genomes (van Hemert and Berkhout 1995), use sub-optimal codons (Jenkins and Holmes 2003) and encode lysine-rich proteins (Berkhout and van Hemert 1994). It is believed that the cause of this bias in lentiviridae is G-to-A hypermutation during reverse transcription (Vartanian et al. 1994; Deforche et al. 2007) but sequence editing by restriction factors of the

APOBEC3 family could also contribute to the bias (Lecossier et al. 2003). Our data do not allow us to determine if the same mechanisms are at play in L1. It is however well documented that APOBEC3 proteins play a role in inhibiting L1 retrotransposition (Schumann 2007). A search of the lizard genome (at genome.ucsc.edu) revealed the presence of several homologues of mammalian *APOBEC3* genes but these genes are absent from the genome of the frog and fish (Conticello et al. 2005). Since organisms that lack *APOBEC3* genes have a less biased base composition, it is tempting to speculate a role of APOBEC3 sequence editing in the adenine enrichment of L1 in amniotes.

The most striking difference among vertebrates L1 resides in the length, structure and level of conservation of the 5′UTR. Vertebrates 5′UTRs fall into two types: the long GC-rich 5′UTR of mammals and lizard clade 1 and the much shorter 5′UTR of lizard clade 2, frog and fish. Although similar in length and base composition, the long 5′UTR of mammals and lizard differ drastically in their mode of evolution. The mammalian 5′UTR shows very little homology among species past the YY1 transcription initiation site (Athanikar et al. 2004). This is due to the frequent acquisition of novel, non-homologous 5′UTR during the evolution of mammals (Adey, Schichman, et al. 1994; Khan et al. 2006; Sookdeo et al. 2013). Presumably, the acquisition of a novel 5′UTR by an L1 family allows this family to avoid sequence-specific repression of transcription, resulting in an arms race between L1, which is escaping repression by acquiring new promoters, and the host which must evolve repressors of the novel 5′UTR. This scenario is consistent with the coevolution between the KZNF transcriptional silencer and L1 in primates (Jacobs et al. 2014). In contrast, the lizard clade 1 5′UTRs can be aligned over most of their length and do not show sign of replacement. The only exception is family L1_AC9, which carries a non-homologous 5′UTR. Similarly we failed to find evidence of replacement of the short 5′UTRs of fish, frog and lizard clade 2, which are highly conserved in length and exhibit strong conservation of several motifs across highly divergent families. This suggests that the transcription of L1 elements within a species, as well as the regulation of transcription by the host, relies on the same biochemical machinery and that the arms race between the promoter sequence and host repressors is an evolutionary feature specific of mammals.

Although the general structure of ORF1 is conserved among organisms, we found substantial differences in the rate of evolution of the CCD. In mammals, the CCD is evolving very rapidly in sequence and structurally, which is consistent with adaptive evolution in response to a host repressor (Boissinot and Furano 2001). This results in very diverse and non-alignable CCDs, composed of an alternation of canonical and disrupted heptads (for an example in mouse, see Sookdeo et al. 2013). In contrast, the CCD in lizard is relatively conserved among families and the structure of the CCD in frog and fish is composed of a perfect succession of canonical heptads. This suggests that the CCD might not be evolving adaptively in non-mammalian vertebrates, implying that the arms race hypothesized in mammals does not exist in non-mammals or does not involve an interaction between a host factor and the CCD.

Another major difference between vertebrate L1 is the ubiquitous presence of an IGR in frog, fish, lizard clade 2 and some basal mammals (opossum and afrotheria). An obvious implication of the presence or absence of an IGR is the effect this region will have on the translation of the ORFs. The IGRs differ considerably in length and base composition, yet the two IRES detection programs we used (IRESPred and VIPS) suggest the presence of IRES in the vast majority of elements with long IGR. Considering the uncertainty of *in silico* IRES predictions, it will be necessary to validate experimentally the presence of functional IRES in the IGR. It is interesting to note that the presence of a functional IRES upstream of ORF2 has been postulated in mouse L1 (Li et al. 2006). Studies in human however demonstrated that the region upstream of ORF2 was not necessary for efficient retrotransposition (Alisch et al. 2006), leading to the suggestion that ORF2 was translated by an unusual termination/re-initiation mechanism. In the context of a long IGR, the possibility for spurious re-initiation seems significant given the length of the IGR and would constitute a very inefficient and risky mechanism to translate ORF2. It should be noted that these experimental studies were conducted on mouse and human L1, which have no or very small IGRs. The only functional study performed in a species with an IGR was done on an ancestral megabat L1 (Yang et al. 2014). It was shown that the IGR was dispensable and in fact inhibits retrotransposition. However since the mobility of the megabat L1 was tested in a human cell line, it is plausible that these experiments do not recapitulate L1 retrotransposition in its native environment. Furthermore, the wide distribution of an IGR across vertebrates and the persistence of an L1 element with a long IGR in megabat (Yang et al. 2014) contradict a strong negative impact of this region on retrotransposition in natural conditions.

Interestingly, the VIPS program found that the region of the IGR adjacent to ORF2 has some structural similarities with the IRES of dicistroviridae, a family of positive-stranded RNA viruses (Pfingsten and Kieft 2008; Nakashima and Uchiumi 2009). These viruses related to picornaviridaes infect invertebrates and have a linear genome consisting of two open-reading frames separated by an IGR, hence the name of the family. This dicistronic structure is very similar to the one of L1, although L1 is not related to this family of viruses. This raises the intriguing possibility that dicistroviridaes and L1 have independently evolved similar mechanisms for the translation of their second ORF.

The 3′UTR differ considerably in composition among organisms and the presence of a highly conserved poly-G tract in mammals is, in fact, a mammalian-specific feature. It was shown that the mammalian poly-G tract has the ability to form intra-strand tetraplexes but also stable RNA secondary structures (Howell and Usdin 1997). The exact role of the poly-G tract remains unclear and it has been shown that retrotransposition using an L1 vector with a disrupted 3′UTR can occur (Moran et al. 1996). Experiments will be required to determine if the 3′UTR of other vertebrates also form non-standard structures, at the DNA or RNA level. It is however puzzling that the highly divergent L1s of fish, frog and lizard, each have species-specific repeated motifs (a mixture of poly-C and poly-T in lizard and frog and long poly-T tracts in fish) suggesting that, not only the potential ability to generate unusual structure, but also the base composition of the repeats could be functionally important.

L1 Evolution and Host-L1 Interactions

One of the most striking observations is the overall conservation of L1 in sequence and structure within each vertebrate lineage. This is particularly obvious in frog and fish, which contain multiple deeply divergent L1 families that are similar in base composition, structure (presence of IGR) and sequence (high conservation of the 5′ termini and presence of similar repeats in the 3′UTR) within each species but different among species. Considering that L1 families have coexisted in these genomes since the origin of vertebrates, they had ample time to diversify functionally (by acquiring different promoters or evolving different base composition) in order to colonize distinct genomic niches and/or recruit different hosts factors. Yet, they did not. This is suggestive of a high level of adaptation of L1 to the host's genome wherein coexisting elements are subjected to the same functional constraints imposed by the host. For instance, we can speculate that all L1 families in reptiles, amphibians and fish rely on a highly conserved host factor for their transcription, and that any change in the promoter would be deleterious to L1 replication.

These differences also suggest that the mechanism of control of L1 in non-mammalian hosts is radically different than it is in mammals. In mammals, a number of processes have evolved to regulate L1 transposition and this regulation yielded the arms race exemplified by the frequent replacement of 5′UTR and adaptive evolution in ORF1 (Khan et al. 2006; Sookdeo et al. 2013). In non-mammalian vertebrates, we do not see any evidence for such an arms race. This is not surprising considering the diversity of transposable elements these genomes harbor. In addition to L1, the lizard genome hosts an even larger number of L2 families and other LINEs (CR1, RTE), numerous DNA transposons and LTR-retrotransposons (Novick et al. 2009; Alfoldi et al. 2011; Novick et al. 2011). In the frog and zebrafish, this diversity could be even higher (Hellsten et al. 2010; Howe et al. 2013; Chalopin et al. 2015). In these species, it is very unlikely that the host has evolved mechanisms of repressions that are specific to each type of transposable elements. A more efficient strategy would be to repress transposition non-specifically, the way DNA methylation is acting. Thus, from the point of view of the host, the different L1 lineages are functionally equivalent and are not repressed in a specific manner, thus removing the need for change. The absence of an arms race between L1 and its non-mammalian host is highly consistent with the phylogeny of L1, which does not show the cascade structure typical of mammalian L1, but fits the expectation of a stochastic birth and death model of evolution.

The apparent lack of an arms race between L1 and its host in non-mammalian vertebrates can find its origin in the population dynamics of L1 insertions in those genomes. In fish and reptiles, very young insertions are over-represented suggesting a low rate of fixation of L1, possibly because novel insertions are under stronger purifying selections in fish and reptiles than they are in mammals (Furano et al. 2004; Tollis and Boissinot 2012). This is particularly true for long elements, including full-length ones, which are found at extremely low frequency in natural populations of stickleback (Blass et al. 2012) and anole (Tollis and Boissinot 2013), and almost never reach fixation. Consequently, the number of full-length progenitors in a given genome is very small in these species. In contrast, full-length elements in mammals attain high frequency and can eventually reach fixation, although not to the same extent as short truncated elements do (Boissinot et al. 2001). This accumulation gives rise to genomes with hundred or thousand potentially active copies (DeBerardinis et al. 1998; Goodier et al. 2001; Beck et al. 2011; Streva et al. 2015). This difference between mammals and non-mammals is most likely due to a lower rate of ectopic recombination in mammals and thus a lower negative effect of long insertions, which are more likely to mediate deleterious chromosomal rearrangements (Furano et al. 2004; Myers et al. 2005; Song and Boissinot 2007). We can speculate that the accumulation of L1 progenitors in mammalian genomes could yield a higher rate of transposition in mammals, thus setting the stage for an arms race between L1 and its host. In contrast, the very small number of progenitors in a given fish or reptile genome could result in a low transposition rate that would be insufficiently deleterious to trigger the evolution of a specific response by the host. The hypothesis of a differential rate of transposition among vertebrates will require experimental evidence. The model species analyzed here constitute excellent systems to address this question.

Supplementary Material

Supplementary materials S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/)

Acknowledgments

We thank Robert Ruggiero and three anonymous reviewers for their helpful comments on the manuscript.

Literature Cited

Adey NB, Schichman SA, et al. 1994. Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol Biol Evol. 11:778–789.

Adey NB, Tollefsbol TO, Sparks AB, Edgell MH, Hutchison CA III. 1994. Molecular resurrection of an extinct ancestral promoter for mouse L1. Proc Natl Acad Sci U S A. 91:1569–1573.

Alfoldi J, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477:587–591.

Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV. 2006. Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev. 20:210–224.

Athanikar JN, Badge RM, Moran JV. 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res. 32:3846–3855.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet. 12:187–215.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Berkhout B, van Hemert FJ. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res. 22:1705–1711.

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, Davis J, Entezam A, Petrov D, Furano AV. 2006. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A. 103:9590–9594.

Boissinot S, Entezam A, Furano AV. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol. 18:926–935.

Boissinot S, Furano AV. 2001. Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol. 18:2186–2194.

Callahan KE, Hickman AB, Jones CE, Ghirlando R, Furano AV. 2012. Polymerization and nucleic acid-binding properties of human L1 ORF1 protein. Nucleic Acids Res. 40:813–827.

Castoe TA, et al. 2011. Discovery of highly divergent repeat landscapes in snake genomes using high throughput sequencing. Genome Biol Evol. 3:641–653.

Chalopin D, Naville M, Plard F, Galiana D, Volff JN. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 7:567–580.

Conticello SG, Thomas CJ, Petersen-Mahrt SK, Neuberger MS. 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol. 22:367–377.

Cook PR, Jones CE, Furano AV. 2015. Phosphorylation of ORF1p is required for L1 retrotransposition. Proc Natl Acad Sci U S A. 112:4298–4303.

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

DeBerardinis RJ, Goodier JL, Ostertag EM, Kazazian HH. Jr 1998. Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. Nat Genet. 20:288–290.

DeBerardinis RJ, Kazazian HH. Jr 1999. Analysis of the promoter from an expanding mouse retrotransposon subfamily. Genomics 56:317–323.

Deforche K, et al. 2007. Estimating the relative contribution of dNTP pool imbalance and APOBEC3G/3F editing to HIV evolution in vivo. J Comput Biol. 14:1105–1114.

Denli AM, et al. 2015. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. Cell 163:583–593.

Doucet AJ, et al. 2010. Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet. 6:e1001150.

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.

Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905–916.

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 D, Boissinot S. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet. 20:9–14.

Goodier JL, Ostertag EM, Du K, Kazazian HH. Jr. 2001. A novel active L1 retrotransposon subfamily in the mouse. Genome Res. 11:1677–1685.

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite. Nucleic Acids Res. 36:W70–W74.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429:268–274.

Hellsten U, et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science 328:633–636.

Hong JJ, Wu TY, Chang TY, Chen CY. 2013. Viral IRES prediction system—a web server for prediction of the IRES secondary structure in silico. PLoS One 8:e79288.

Howe K, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.

Howell R, Usdin K. 1997. The ability to form intrastrand tetraplexes is an evolutionarily conserved feature of the 3′ end of L1 retrotransposons. Mol Biol Evol. 14:144–155.

Jacobs FM, et al. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516:242–245.

Januszyk K, et al. 2007. Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. J Biol Chem. 282:24893–24904.

Jenkins GM, Holmes EC. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92:1–7.

Kammerer RA, et al. 2005. A conserved trimerization motif controls the topology of short coiled coils. Proc Natl Acad Sci U S A. 102:13891–13896.

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.

Kolekar P, Pataskar A, Kulkarni-Kale U, Pal J, Kulkarni A. 2016. IRESPred: web server for prediction of cellular and viral internal ribosome entry site (IRES). Sci Rep. 6:27436.

Kolosha VO, Martin SL. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci U S A. 94:10155–10160.

Kulpa DA, Moran JV. 2005. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Hum Mol Genet. 14:3237–3248.

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

Lecossier D, Bouchonnet F, Clavel F, Hance AJ. 2003. Hypermutation of HIV-1 DNA in the absence of the Vif protein. Science 300:1112.

Li PW, Li J, Timmerman SL, Krushel LA, Martin SL. 2006. The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic Acids Res. 34:853–864.

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

Martin SL, Branciforte D, Keller D, Bain DL. 2003. Trimeric structure for an essential protein in L1 retrotransposition. Proc Natl Acad Sci U S A. 100:13815–13820.

Martin SL, Bushman FD. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol. 21:467–475.

Mathias SL, Scott AF, Kazazian HH, Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. Science 254:1808–1810.

Meredith RW, et al. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.

Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. Genome Biol. 3:REVIEWS0004.

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.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.

Nakashima N, Uchiumi T. 2009. Functional analysis of structural motifs in dicistroviruses. Virus Res. 139:137–147.

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.

Novick PA, Smith JD, Floumanhaft M, Ray DA, Boissinot S. 2011. The evolution and diversity of DNA transposons in the genome of the Lizard *Anolis carolinensis*. Genome Biol Evol. 3:1–14.

Perepelitsa-Belancio V, Deininger PL. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet. 35:363–366.

Pfingsten JS, Kieft JS. 2008. RNA structure-based ribosome recruitment: lessons from the Dicistroviridae intergenic region IRESes. RNA 14:1255–1263.

Pickeral OK, Makalowski W, Boguski MS, Boeke JD. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 10:411–415.

Puigbo P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct. 3:38.

Richardson SR, et al. 2015. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol Spectr. 3:MDNA3–M0061.

Rishishwar L, Tellez Villa CE, Jordan IK. 2015. Transposable element polymorphisms recapitulate human evolution. Mob DNA 6:21.

Schumann GG. 2007. APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition. Biochem Soc Trans. 35:637–642.

Severynse DM, Hutchison CA, 3rd, Edgell MH. 1992. Identification of transcriptional regulatory activity within the 5′ A-type monomer sequence of the mouse LINE-1 retroposon. Mamm Genome 2:41–50.

Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.

Smit AF, Toth G, Riggs AD, Jurka J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol. 246:401–417.

Song M, Boissinot S. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. Gene 390:206–213.

Sookdeo A, Hepp CM, McClure MA, Boissinot S. 2013. Revisiting the evolution of mouse LINE-1 in the genomic era. Mob DNA 4:3.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol. 21:1973–1985.

Streva VA, et al. 2015. Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. BMC Genomics 16:220.

Sun YB, et al. 2015. Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes. Proc Natl Acad Sci U S A. 112:E1257–E1262.

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

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

Taylor MS, et al. 2013. Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell 155:1034–1048.

Tollis M, Boissinot S. 2012. The evolutionary dynamics of transposable elements in eukaryote genomes. Genome Dyn. 7:68–91.

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.

van Hemert FJ, Berkhout B. 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol. 41:132–140.

Vartanian JP, Meyerhans A, Sala M, Wain-Hobson S. 1994. G–>A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc Natl Acad Sci U S A. 91:3092–3096.

Warren IA, et al. 2015. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosome Res. 23:505–531.

Witherspoon DJ, et al. 2006. Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions. Hum Hered. 62:30–46.

Wright F. 1990. The ‘effective number of codons’ used in a gene. Gene 87:23–29.

Xu Q, Minor DL. Jr 2009. Crystal structure of a trimeric form of the K(V)7.1 (KCNQ1) A-domain tail coiled-coil reveals structural plasticity and context dependent changes in a putative coiled-coil trimerization motif. Protein Sci. 18:2100–2114.

Yang L, Brunsfeld J, Scott L, Wichman H. 2014. Reviving the dead: history and reactivation of an extinct l1. PLoS Genet. 10:e1004395.

Associate editor: Mar Alba