Distinct Retrotransposon Evolution Profile in the Genome of Rabbit (*Oryctolagus cuniculus*)

Naisu Yang[1], Bohao Zhao[1], Yang Chen[1], Enrico D'Alessandro[2], Cai Chen[1], Ting Ji[1], Xinsheng Wu[1,*], and Chengyi Song[1,*]

[1]College of Animal Science & Technology, Yangzhou University, Jiangsu, China

[2]Department of Veterinary Science, University of Messina, Italy

*Corresponding authors: E-mails: xswu@yzu.edu.cn; cysong@yzu.edu.cn.

Accepted: 12 July 2021

Abstract

Although the rabbit genome has already been annotated, it is mobilome remains largely unknown. Here, multiple pipelines were used to de novo mine and annotate the mobilome in rabbit. Four families and 19 subfamilies of LINE1s, two families and nine subfamilies of SINEs, and 12 ERV families were defined in rabbit based on sequence identity, structural organization, and phylogenetic tree. The analysis of insertion age and polymerase chain reaction suggests that a number of families are very young and may remain active, such as L1B, L1D, OcuSINEA, and OcuERV1. RepeatMasker annotation revealed a distinct transposable element landscape within the genome, with approximately two million copies of SINEs, representing the greatest proportion of the genome (19.61%), followed by LINEs (15.44%), and LTRs (4.11%), respectively, considerably different from most other mammal mobilomes except hedgehog and tree shrew, in which LINEs have the highest proportion. Furthermore, a very high rate of insertion polymorphisms (>85%) for the youngest subfamily (OcuSINEA1) was identified by polymerase chain reaction. The majority of retrotransposon insertions overlapped with protein-coding regions (>80%) and lncRNA (90%) genes. Genomic distribution bias was observed for retrotransposons, with those immediately upstream (−1 kb) and downstream (1 kb) of genes significantly depleted. Local GC content in 50-kb widows had significantly negative correlations with LINE (r_s=−0.996) and LTR (r_s=−0.829) insertions. The current study revealed a distinct mobilome landscape in rabbit, which will assist in the elucidation of the evolution of the genome of lagomorphs, and even other mammals.

Key words: rabbit, lagomorph, genome, mobilome, retrotransposon, evolution.

Significance

Transposable elements (constituting the major part of the mobilome) account for approximately half of the mammalian genome and are believed to play pivotal roles in genome evolution, although their distribution within the rabbit genome remains vague. Here, we systematically characterized the evolutionary profile of mobilome (mainly represented by retrotransposons), including long-interspersed nuclear elements, short-interspersed nuclear elements, and endogenous retroviruses in the rabbit using multiple de novo mining pipelines, revealing a mobilome landscape in the rabbit distinct from the most surveyed mammalian genomes. The findings of this study suggest possible roles for transposable elements in the evolution of the rabbit genome and provide a better understanding of genomic evolution in lagomorphs.

Introduction

Mobilome, which has been defined previously (Siefert 2009), includes transposable elements (TEs) or transposons, plasmids, bacteriophage, and self-splicing molecular parasites. TEs, can be classified as either retrotransposons or DNA transposons, and are extensively distributed in nature, playing important

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

roles in the evolution of the genomes of different organisms (Böhne et al. 2008; Oliver and Greene 2009; Chalopin et al. 2015; Chuong et al. 2017; Bourque et al. 2018). Retrotransposons are the major parasitic elements in mammalian genomes. They are subdivided into two large categories distinguished by the presence or absence of long-terminal repeats (LTRs): LTR retrotransposons and non-LTR retrotransposons. LTR retrotransposons have LTRs at both ends and comprise five superfamilies: Copia, Gypsy, BEL, DIRS, and endogenous retroviruses (ERVs). By contrast, non-LTR retrotransposons lack LTRs at each terminal end and include long-interspersed nuclear elements (LINEs) and short-interspersed nuclear elements (SINEs) (Kazazian 2004; Wicker et al. 2007). Retrotransposons account for more than one-third to nearly half of the surveyed mammalian genome, LINEs being the most abundant type (genomic coverage), followed by SINEs and LTR retrotransposons (Mandal and Kazazian 2008). LINEs account for 20.42% of the human genome, 19.21% of the mice, and 29.17% of the opossum genome. LINEs and SINEs continue to be active elements in the majority of mammalian genomes, playing a role in shaping their evolution (Cordaux and Batzer 2009; Shpyleva et al. 2018).

Rabbits (*Oryctolagus cuniculus*) and pikas (*Ochotona princeps*), both classified in the taxonomic order of lagomorphs (Chapman and Flux 2008), with rodents belong to the Glires clade (Douzery and Huchon 2004), after primates the closest phylogenetic relative to humans (Dutta and Sengupta 2018). Rabbits have remained relatively unchanged during the past 40 Myr, the majority of changes occurring after colonization of North America during that period (Smith 2021). Additionally, most of the evolutionary events revolved around domestication (~1,400 years ago) and resistance to disease, changes that were influenced by human and environmental factors (Clutton-Brock 1999; Carneiro et al. 2014). Rabbits are monogastric herbivores, widely distributed around the globe and used in the production of meat, fur, and wool, particularly in China. In addition, rabbits are among the most common experimental animals in biomedical studies because of their moderate size, gentle disposition, superior reproductive performance, short life span, and their generational gap (Wang et al. 2017; Fan et al. 2018; Ma et al. 2018; Wu et al. 2018). The genome of the rabbit was firstly sequenced and assembled (Orycun 2.0, 2.66 Gb) in 2014. Genes coding for proteins (20,318) and lncRNAs (14,165) have been well-defined and annotated, whereas information about the repeating content in the genome of the species is very limited (Carneiro et al. 2014). Recently, ERVs were annotated and compared in the wild and domestic rabbit populations (Rivas-Carrillo et al. 2018). However, the entire TE landscape in this species remains largely unknown. In the current study, the major retrotransposon groups (including LINEs, SINEs, and ERVs) in the rabbit reference genome (Orycun 2.0) were de novo mined and annotated using multiple pipelines. Their intradiversity, activity, density, and evolutionary dynamics in the genome were systematically characterized. We also analyzed the distribution bias of retrotransposons in the genome and their intersections with genes. The study presents a genome-wide landscape of retrotransposons in the rabbit which may assist in elucidating the coevolution of the mobilome and genome in lagomorphs in addition to helping comprehend the evolution of the genome.

Results

Evolution of L1s in the Rabbit Genome

To systematically investigate the evolutionary profile of L1s in the rabbit genome, the RepBase (Bao et al. 2015) and L1Base databases were used to identify L1s which were merged with those mined using MGEScan-nonLTR (Rho and Tang 2009). Firstly, 2,974 L1 elements identified in the rabbit genome by MGEScan-non-LTR, which may have been truncated, were aligned against the rabbit genome using the BLAST-like alignment tool (BLAT) (Kent 2002) to obtain their genomic positions. The bedtools toolset was subsequently used to extend the sequence an additional 2,500 bp along the 5′-untranslated region (5′-UTR) and 500 bp along the 3′-UTR to obtain the full length of these elements. Thus, 4,296 L1 elements downloaded from the L1Base database (Penzkofer et al. 2016) were merged with these L1s and any redundancy removed. Finally, 5,076 L1 elements with unique positions in the rabbit genome were obtained and classified into 19 distinct subfamilies (17 new and two known subfamilies) depending on their consensus sequences. Old subfamilies with less than ten copies and high levels of divergence were discarded. Two subfamilies (L1A_OC and L1A2_OC) that had been deposited in RepBase and also identified by the MGEScan-non-LTR pipeline, were included in the 19 distinct subfamilies, whereas two subfamilies (L1B_OC and L1C_OC) identified in RepBase, were fragmented and shorter than normal L1s, were not detected by the protocol described here and excluded from further analysis. The 19 subfamilies were further classified into four distinct families (termed L1A, L1B, L1C, and L1D), based on the polygenic consensus tree and their structural organization (fig. 1*A*, table 1, and supplementary table S1, Supplementary Material online).

The details of the L1 families, including their names, classification, characteristics, GC content, and copy number, are summarized in table 1 and supplementary table S1, Supplementary Material online. The consensus sequences of each subfamily are supplied in an additional file (supplementary data set S1, Supplementary Material online). The total length of the consensus sequences varied between 6,805 and 7,817 bp, whereas the length of the 5′-UTR varied from 473 to 1,248 bp, and the 3′-UTR (excluding poly-A sequences) varied from 1,201 to 1,581 bp. The intergenic region (IGR) of the four families (L1A, L1B, L1C, and L1D) was 50–51, 6–27, 68–69, and 81 bp, respectively. The lengths of IGRs in four

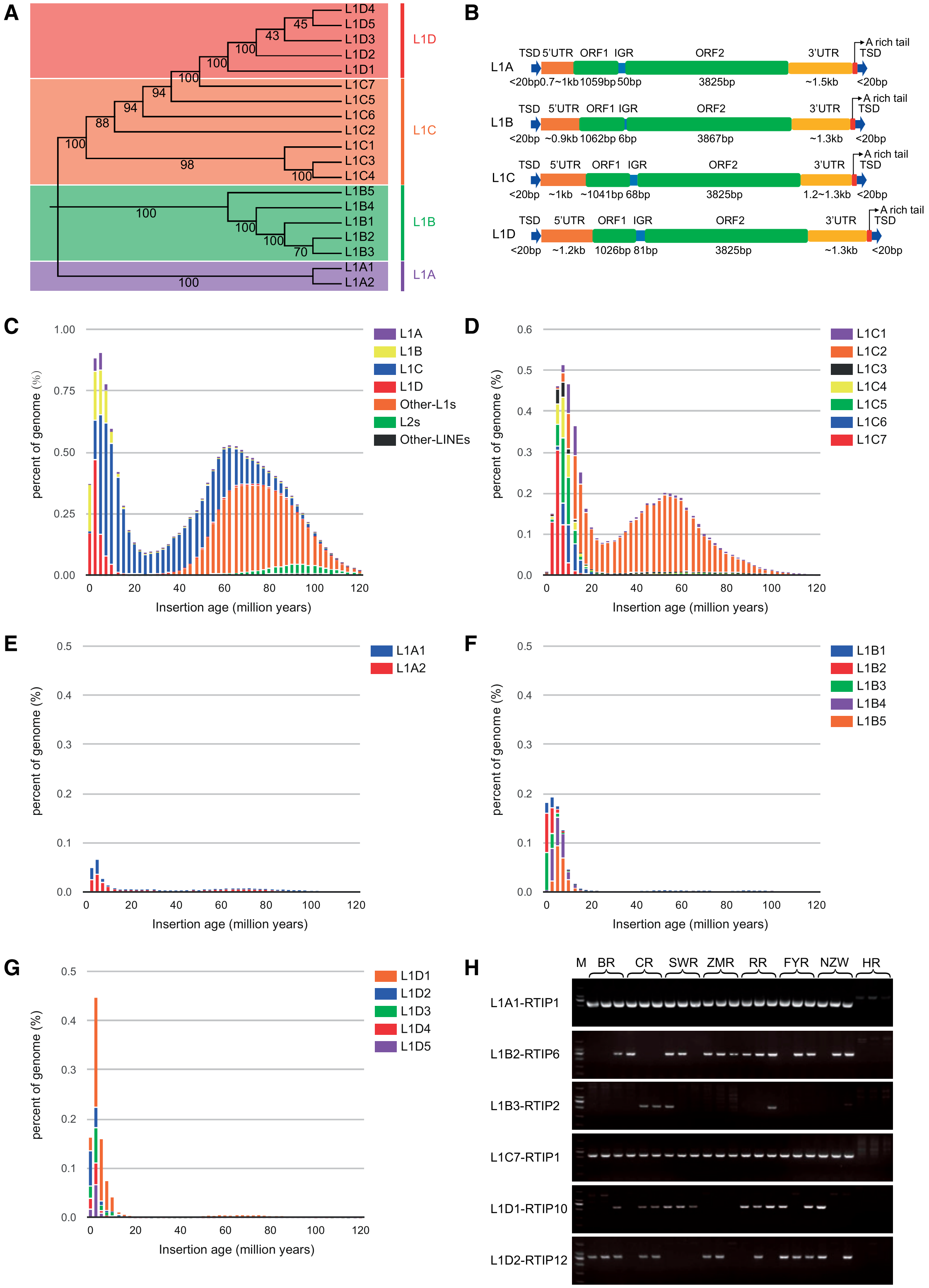

Fig. 1.—Evolution and activity analysis of L1s in the rabbit. (*A*) Maximum likelihood phylogenetic tree of the L1s revealed four families of L1s, namely L1A, L1B, L1C, and L1D that had 2, 5, 7, and 5 subfamilies respectively and (*B*) structural schematics of L1A, L1B, L1C, and L1D families. (*C*–*G*) Age distribution of L1s and its four families (L1A–L1D); (*H*) representative results of insertion polymorphism detection of putatively active L1s by PCR. Lanes are M (2,000 bp marker) and PCR measurements for eight breeds of rabbit, each breed comprising three individuals. BR, Belgian rabbit; CR, California rabbit; SWR, Sichuan white rabbit; ZMR, ZIKA meat-type rabbit; RR, Rex rabbit; FYR, Fujian yellow rabbit; NZW, New Zealand white rabbits; HR, hare.

Table 1

Classification of L1 Families in the Rabbit Genome

L1 Family	Subfamily Number	Length (bp)						Putatively Active L1 Number
		Consensus	5′-UTR	ORF1	IGR	ORF2	3′-UTR (No PolyA)	
L1A	2	7,235–7,610	753–1,094	1,005–1,059	50–51	3,825–3,828	1,508–1,581	2
L1B	5	6,805–7,296	473–938	1,041–1,062	6–27	3,867	1,201–1,375	28
L1C	7	7,183–7,514	841–1,071	1,038–1,065	68–69	3,825	1,201–1,508	3
L1D	5	7,401–7,617	1,009–1,248	1,026	81	3,825	1,368–1,377	38

L1B subfamilies were very short, with 6 bp for L1B1, L1B2, L1B3, and L1B4, and 27 bp for L1B5. The lengths of the two open reading frames (ORF1 and ORF2) were relatively conservative across these subfamilies. The copy numbers of L1 elements, the numbers of subfamilies, evolutionary time, and the number of putatively active L1 elements varied significantly between the families. The L1A family represented the lowest diversity, with only two subfamilies detected, whereas this family also displayed low activity with only two copies being putatively active in the genome. Five, seven, and five subfamilies were identified in the L1B, L1C, and L1D families respectively, but putatively active L1s were mainly detected in the L1B and L1D families, suggesting that these two families represent recent activity and some copies may still active, whereas L1C tended to be dead with very few putatively active copies detected (table 1 and supplementary table S1, Supplementary Material online). In total, 71 putatively active L1 elements with a typical structure of mammalian L1 were identified, the majority belonging to L1B (28 copies) or L1D (38 copies). The structural organization and lengths of ORF1, IGR, and ORF2 were highly conserved within the LIB and L1D families. Putatively active L1s in the LIB family possessed a 1,062 bp/354 aa ORF1, 6 bp IGR, and 3,867 bp/1,289 aa ORF2, whereas putatively active L1s of the L1D family had a 1,026 bp/342 aa ORF1, 81 bp IGR, and 3,825 bp/1,275 aa ORF2. In addition, LIB family numbers were approximately 7.2 kb in length in total, including a ∼900 bp 5′-UTR and ∼1.3 kb 3′-UTR. However, compared with L1B elements, putatively active L1D elements exhibited a longer 5′-UTR (from 1,009 to 1,248 bp) and a 3′-UTR of a similar length (∼1.3 kb), with a total length of ∼7.5 kb (fig. 1*B*, table 1, and supplementary table S1, Supplementary Material online).

Insertion age analysis revealed differential evolutionary profiles of LINEs in the rabbit. L1C family and other L1s (mammalian common L1s) dominated the ancient evolution of LINEs in the rabbit genome, whereas L1A, L1B, and L1D dominated their recent evolution, very low ancient copies of L2s and other LINEs were identified in genome, and they were extinct in recent evolution (fig. 1*C*). In detail, only the L1C2 subfamily of the L1C family experienced a relatively old and long expansion in the rabbit genome, with peak activity 52.5 Ma, followed by a substantial decrease in activity over the last 7.5 Myr, whereas all other subfamilies of L1C (L1C1 and L1C3–C7) exhibited recent expansion over the previous 20 Myr (fig. 1*D*). Additionally, L1A, L1B, and L1D are very recently evolved families and coevolved within 12.5 Myr (fig. 1*E–G*). Furthermore, both L1B and L1D contain many intact copies, which harbor ORF1 and ORF2 with coding capability, indicating that a number of L1 elements of L1B and L1D may still be active. Based on insertion age analysis, L1B2 and L1B3 may represent the youngest subfamilies across the L1B family (fig. 1*F*), whereas L1D1 and L1D2 may be the youngest across the L1D family (fig. 1*G*), suppositions well-supported by higher numbers of intact copies of L1 identified in these subfamilies and the high rates of insertion polymorphisms of L1B (39.29%, 11/28) and L1D (26.32%, 10/38) across eight breeds of rabbits (fig. 1*H* and supplementary tables S1 and S2 and fig. S3, Supplementary Material online).

Evolution of SINEs in the Rabbit Genome

RepeatModeler and SINE_Scan software were used to detect SINEs in the rabbit genome, identifying eight and three SINE consensuses, respectively. They were merged with four SINE consensus sequences from Repbase, and finally, nine SINE consensus sequences (five new and four known) were obtained after removal of the redundant data. These SINE sequences were classified into two families (OcuSINEA and OcuSINEB), depending on sequence alignment (supplementary fig. S1, Supplementary Material online) and the phylogenetic tree (fig. 2*A*). The OcuSINEA family contained four subfamilies (OcuSINEA1, OcuSINEA2, OcuSINEA3, and OcuSINEA4), whereas the OcuSINEB family contained five (OcuSINEB1, OcuSINEB2, OcuSINEB3, OcuSINEB4, and OcuSINEB5) (fig. 2*A* and supplementary data set S2, Supplementary Material online). Elements from the OcuSINEA and OcuSINEB families were approximately 330 bp in length (excluding the poly-A tail) (fig. 2*B* and supplementary fig. S1, table S3, and data set S2, Supplementary Material online), flanked by 8–30 bp of target site duplication (TSD) generated during retrotransposition. Alignments of OcuSINEA revealed that the four subfamilies were highly conserved with few base pair differences, whereas the

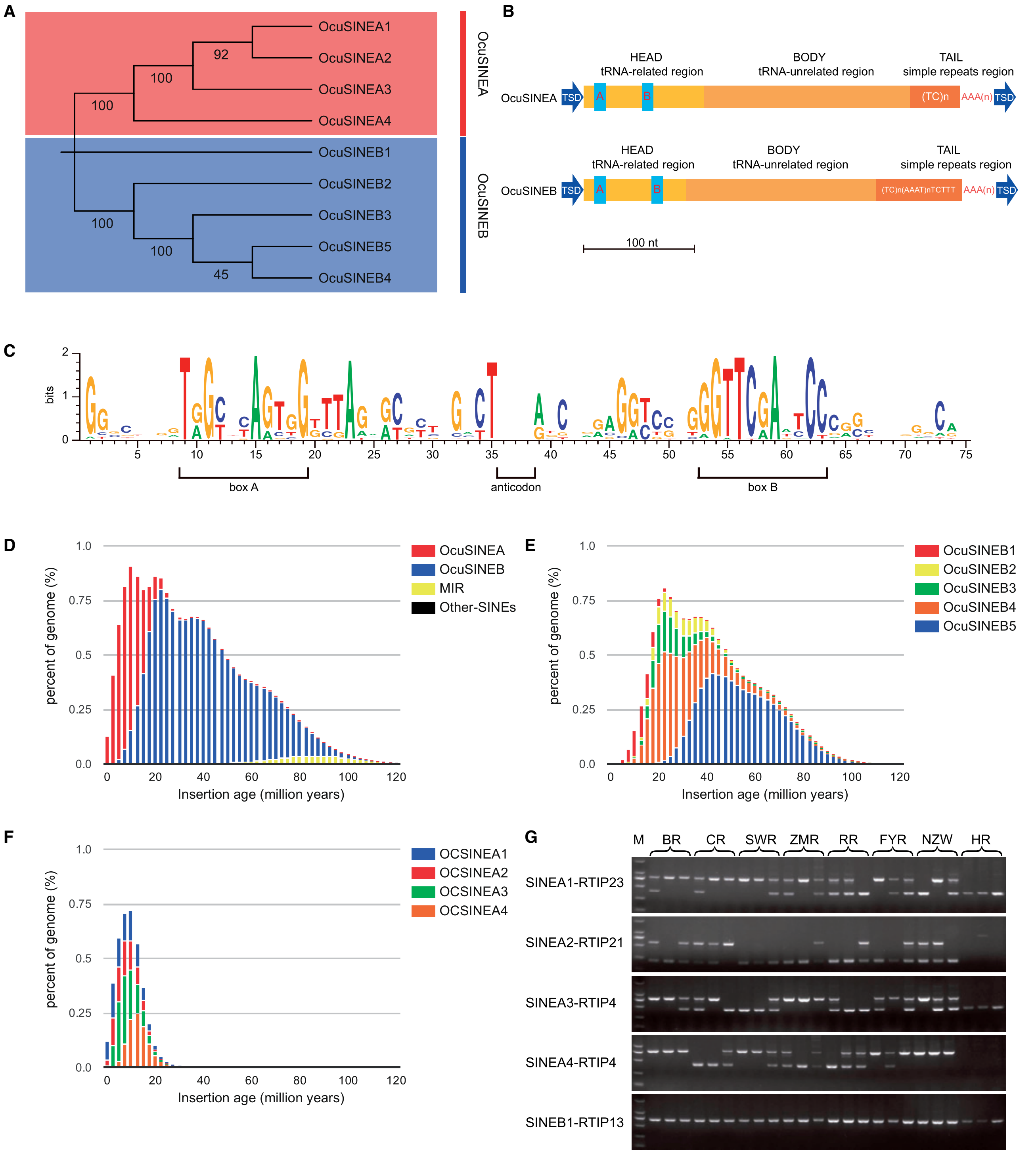

Fig. 2.—Evolution and activity analysis of SINEs in the rabbit. (*A*) Maximum likelihood phylogenetic tree of SINEs indicated the presence of two families (OcuSINEA and OcuSINEB), and four (OcuSINEA1–OcuSINEA4) and five (OcuSINEB1–OcuSINEB5) subfamilies, respectively. (*B*) Structural schematics of OcuSINEA and OcuSINEB. (*C*) Sequence graphics of 483 rabbit tRNA genes were generated by the WebLogo 2.8.2 tool. (*D–F*) Age distribution of SINEs, OCSINEA, and OCSINEB, and their subfamilies. (*G*) Representative SINE insertion polymorphism detection results by PCR for each subfamily. Eight rabbit breeds were used, including BR, Belgian rabbit; CR, California rabbit; RR, Rex rabbit; FYR, Fujian yellow rabbit; NZW, New Zealand white rabbit; HR, hare. M, 2,000 bp marker.

subfamilies of OcuSINEB were more divergent (supplementary fig. S1, Supplementary Material online).

OcuSINE elements possess a head, a body, and a tail in structure (fig. 2*B*). The head is a tRNA-related region, which is an evolutionary element derived from a tRNA synthesized by RNA Polymerase III (Vassetzky and Kramerov 2013). Both OcuSINEA and OcuSINEB subfamilies begin with GC enriched regions (12–13 bp), followed by box A (10–11 bp), a spacer (42 bp in OcuSINEA and 33 bp in OcuSINEB), and box B (11 bp) (fig. 2*C* and supplementary fig. S1, Supplementary Material online). The spacer between A and B promoter boxes represents the RNA polymerase III internal control region (Geiduschek and Tocchini-Valentini 1988). The body of SINEs have a tRNA-unrelated origin, with a function and origin which remain unknown, but generally sharing considerable homology with corresponding LINEs which thus allow SINEs to parasitically co-opt endonucleases coded by LINEs (Mandal et al. 2004; Kumari et al. 2011). The 3′ tail of SINEs comprises a simple repeat of varying length, ending with an A-rich region. There is a long TC-motif immediately upstream of the A-rich tail in the OcuSINEA family. In addition, in the OcuSINEB family, the A-rich tail is followed by multiple polyadenylation signals, AAAT, and a TCTTT sequence, an efficient terminator for RNA pol III (Borodulina and Kramerov 2001, 2008).

Evolutionary dynamics analysis of rabbit-specific SINEs, Mammalian-wide interspersed repeats (MIR), and other SINE families (common rodent), indicate that OcuSINEA and OcuSINEB families account for the vast majority of SINEs in the rabbit genome. MIR represented a small fraction of the rabbit genome, whereas other SINE families were almost absent in the genome of the rabbit (fig. 2*D*). OcuSINEA and OcuSINEB displayed a differential proliferation history. The OcuSINEB family has experienced a long evolutionary history (from 7 to 110 Ma), but activity recently decreased significantly and it has become almost extinct in the last 7 Myr (fig. 2*D* and *E*). Compared with OcuSINEB, OcuSINEA has a very recent evolutionary history with activity mainly in the last 25 Myr, some elements were predicted to be active in recent 2 Myr (fig. 2*F*). In addition, the majority of subfamilies of OcuSINEA have fewer copy numbers than those of OcuSINEB (supplementary table S3, Supplementary Material online). All OcuSINEA subfamilies are polymorphic across rabbit breeds, as revealed by polymerase chain reaction (PCR), confirming that this family may remain active, with OcuSINEA1 having the highest activity with the most recent age of insertion (fig. 2*F*) and extremely high polymorphisms, with more than 85% of OcuSINEA1 insertions (88%, 44/50) being polymorphic, followed by OcuSINEA2 (84%, 42/50), OcuSINEA3 (58%, 29/50), and OcuSINEA4 (56%, 28/50) (supplementary table S4, Supplementary Material online). In contrast, all OcuSINEB subfamily elements were nonpolymorphic (supplementary table S4, Supplementary Material online) and appear to be fixed within the rabbit genome (fig. 2*G* and supplementary table S4, Supplementary Material online).

Table 2
Number of ERVs Detected by LTRHarvest and Retrotector in the Rabbit Genome

Structure	Number of Detected Elements	
	LTRHarvest	Retrotector
Total	26,751	945
ERV containing RT (~700 bp)	167	186
ERV containing *gag* (~1,600 bp)	57	41
ERV containing *pol* (~2,600 bp)	46	29
ERV containing *pro* (~1,000 bp)	60	50
ERV containing *gag, pol*, and *pro*	35	23
Putatively active nonredundant ERVs	36	

Evolution of ERVs in the Rabbit Genome

ERV-derived elements in the rabbit genome were mined using LTRharvest and RetroTector pipelines. In total, 26,751 and 945 ERVs were obtained using LTRharvest and RetroTector, respectively. ERV-derived elements with undetectable reverse transcriptase (RT) regions were then removed, resulting in 167 and 186 ERVs with intact RT regions remaining that had been identified using LTRharvest and RetroTector (table 2), which were submitted for phylogenetic analysis following removal of redundancies. Based on phylogenetic analysis, 12 families of rabbit ERVs (OcuERV1-12, six new and six known families deposited in Repbase) were identified (fig. 3*A*). Three were classified as Class 1 gamma retroviruses, whereas five were alpha retroviruses, three were beta retroviruses, and one was a Class 2 lenti retrovirus (fig. 3*A* and supplementary table S5, Supplementary Material online). The consensus or representative sequence of each family is detailed in supplementary data set S3, Supplementary Material online.

The OcuERV families varied from 5,923 to 9,816 bp in length, with an LTR varying in length from 292 to 671 bp (supplementary table S5, Supplementary Material online). A primer binding site (PBS) of 18 nt downstream and adjacent to the 5′-LTR was identified that was complementary to a specific zone in the 3′-end of a tRNA normally provided by the host cell to begin retrotranscription (Beerens et al. 2001; Damgaard et al. 2004). Alpha retroviruses in the rabbit genome prefer tRNA-His as a primer (supplementary table S5, Supplementary Material online). Additionally, a small region of polypurine tract (PPT) found upstream and adjacent to the 3′-LTR, was responsible for starting the synthesis of the proviral (+) DNA strand (Figiel et al. 2018). To identify the functional domains for all OcuERV families, we translated all six frames of all sequences in each family, and searched against the Pfam database with hmmsearch (Johnson et al. 2010), from which we successfully identified multiple retrovirally relevant domains, including group antigens (Gag), proteases

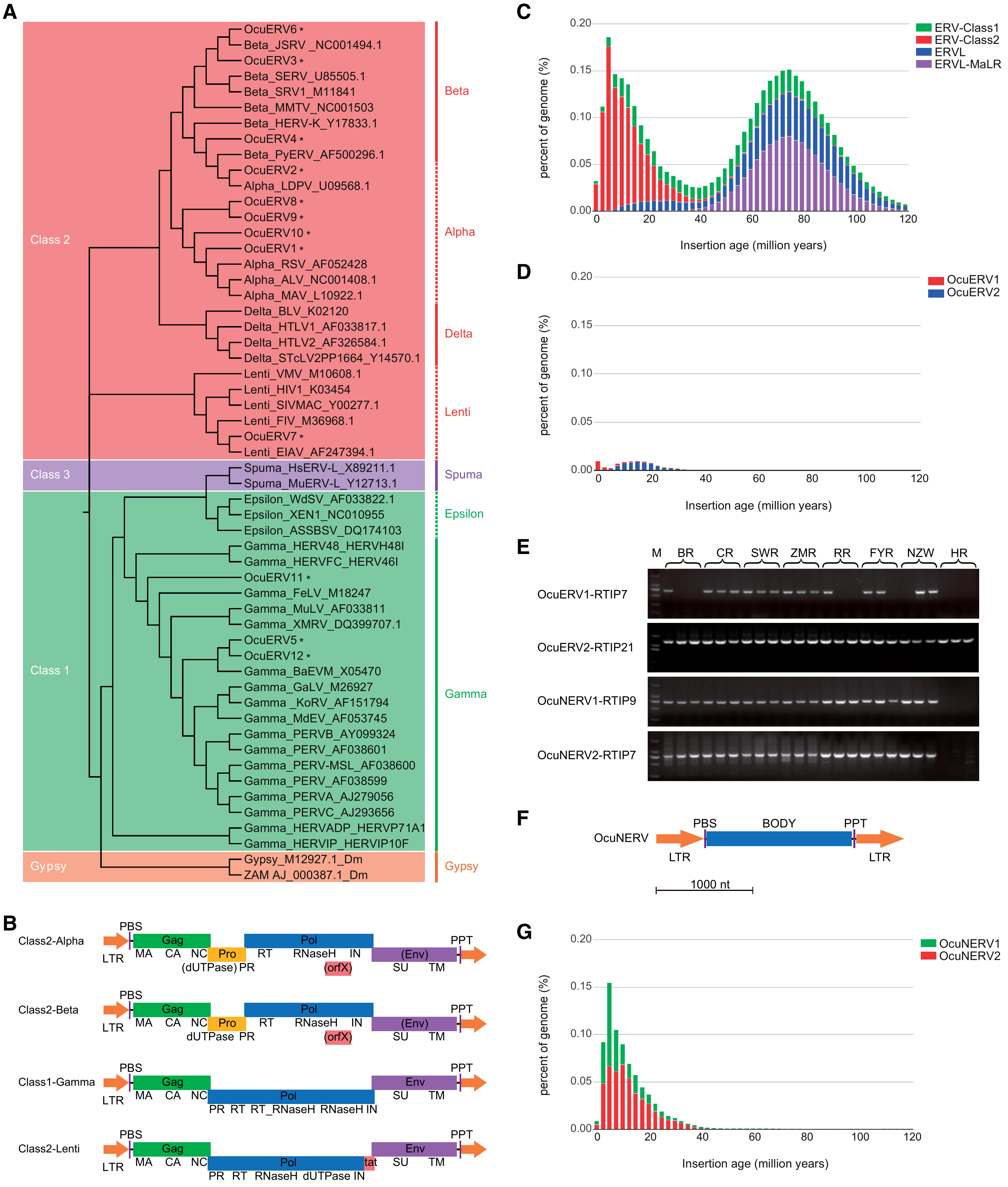

Fig. 3.—Evolution and activity analysis of OcuERVs in rabbits. (*A*) ERVs in the rabbit genome were clustered into four classes including 12 families (OcuERV1–OcuERV12) based on known ERV sequences. (*B*) Structural schematics of four ERV genera in the rabbit. (*C*) Age distribution of ERVs in the rabbit genome. (*D*) Age distribution of the OcuERV1 and OcuERV2 families. (*E*) Representative results of insertion polymorphism detection for 36 putatively active OcuERVs and 20 nonautonomous ERVs (OcuNERVs) by PCR. M, 2,000 bp marker; BR, Belgian Rabbit; CR, California Rabbit; RR, Rex rabbit; FYR, Fujian yellow rabbit; NZW, New Zealand white rabbit; HR, hare. (*F*) Structural schematic of OcuNERV (nonautonomous ERV). (*G*) Age distribution of OcuNERV1 and OcuNERV2.

(Pro), polymerases (Pol), and envelope proteins (Env). However, all had multiple defects, containing numerous in-frame stop codons and frameshift mutations (fig. 3*B* and supplementary fig. S2, Supplementary Material online), suggesting that the majority of ERV families had decayed within the rabbit genome and so were transfection incompetent. Conversely, OcuERV1 and OcuERV2 families, belonging to gamma ERVs, displayed high sequence identity with those that were full length and LTRs, had low divergence, and tended to be recent insertions into the genome. A total of 36 copies of nonredundant OcuERV1 candidates still maintained the coding capacity of retrovirus-relevant domains, including Gag, Pro, and Pol proteins, although the Env domain, which plays important role in ERV transfection (Malik et al. 2000), had decayed (table 2 and supplementary fig. S2 and table S5, Supplementary Material online). supporting the hypothesis that these families may still be retrotransposition competent and be capable of transposition in the genome.

Evolutionary dynamic analysis of rabbit-specific ERVs and mammalian common ERVs (ERVL and ERVL-MaLR) revealed that they had experienced dramatic differential expansion profiles in the rabbit genome. Overall, the evolution of ERVs in the rabbit genome could be categorized as having two stages (ancient and recent). In the ancient stage from about 120 to 40 Ma, Class 1 ERVs and Class III ERVs (ERVL and ERVL-MaLR) coevolved in the rabbit genome, but dominated with Class III ERVs having peak activity around 75 Ma, followed by a significant recent decrease, with the extinction of ERVL-MaLR about 40 Ma, and ERVL about 8 Ma (fig. 3*C*). Other LTR retrotransposons, including Gypsy and Copia, were almost absent in the genome of the rabbit and displayed extremely low copy number (supplementary table S6, Supplementary Material online). In recent times, Class I and Class II ERVs coevolved in the genome of the rabbit, but with significant accumulation of Class II ERVs and weak proliferation of Class I ERVs at this stage (fig. 3*C*).

Although insertion age analysis demonstrated that the ERV1 family of class II ERVs was inserted during the previous 5 Myr, they currently may display activity (fig. 3*D*). The polymorphisms of 36 young ERV1 and 24 ERV2 insertions were evaluated by PCR using eight rabbit breed genomic DNA samples, whereas two-thirds of ERV1 insertions (64%, 23/36) were polymorphic (fig. 3*E* and supplementary table S7 and fig. S3, Supplementary Material online), strongly supporting the hypothesis that the OcuERV1 family was transposition competent and able to transpose within the genome, although no insertion polymorphisms were found for the OcuERv2 family (fig. 3*E* and supplementary table S7, Supplementary Material online).

Further analysis revealed that Class II ERVs in the rabbit genome included not only long elements of the OcuERV family but also a type of short nonautonomous ERV element, termed OcuNERVs, which also provided a substantial contribution to the recent burst of class II ERVs in the rabbit genome, with approximately 34,000 occurrences (supplementary table S8, Supplementary Material online). OcuNERVs represent nonautonomous retrotransposons which rely on the RT and integrase (IN) of OcuERV autonomous transposons to become reverse transcribed and integrate into the genome (Havecker et al. 2004; McCarthy and McDonald 2004). Further analysis revealed that OcuNERVs had a total length of 2.5 kb and consisted of two LTRs (~500 bp), PBS (specific for tRNA-His), and PPT, also located downstream and upstream of 5′-LTR and 3′-LTR regions, respectively (fig. 3*F*). The body of the OcuNERVs was approximately 1.5 kb, whereas no protein motif was detected. OcuNERVs can be further subdivided into two subfamilies (OcuNERV1 and OcuNERV2) based on sequence similarity (supplementary data set S4, Supplementary Material online). Insertion age analysis indicated that OcuNERV1 and OcuNERV2 expanded over the last 40 Myr, peaking at approximately 5 Ma, but with activity declining dramatically over the last million years (fig. 3*G*). No polymorphism was detected for 20 OcuNERV insertions by PCR (fig. 3*E* and supplementary table S9, Supplementary Material online), indicating that these elements were fixed within the genome.

Overall, these data suggest that the OcuERV1 family may be transposition competent in genome of rabbit, and still play a role in shaping the evolution of the genome, whereas OcuNERVs have lost retrotransposition activity although they displayed quite high recent activity.

Distinct Retrotransposon Landscape in the Rabbit Compared with Other Mammals

The landscape of TEs in rabbit was annotated using the RepeatMasker program using a custom library, including the de novo identified elements and known repeats from the RepBase and Dfam databases, as summarized in figure 4*A* and supplementary table S5, Supplementary Material online. The data demonstrate that the genomic coverage of TEs in the rabbit is generally similar to that of the majority of mammals. Four types of TE (LINE, LTR, SINE, and DNA) accounted for 41.60% (~1,083 Mb) of the rabbit genome, with retrotransposons representing the vast majority, at 39.16% of the genome (~1,020 Mb), whereas DNA TEs, mainly representing by hAT and Tc1/mariner superfamilies (supplementary table S6, Supplementary Material online) with extinct activities, accounted for the minority, at 2.44% (~63.6 Mb). However, a distinct TE landscape in the rabbit was obtained for three types of retrotransposons compared with the most surveyed mammal genomes, in which LINEs have the highest genomic coverage, followed by SINEs and LTRs in primates, carnivores, rodents, odd and even-toed ungulates, and bats (Smit et al. 2013–2015). Although SINEs are the most abundant retrotransposon in the rabbit, representing approximately two million copies and accounting for 19.61% (~510.5 Mb) of the rabbit genome, followed by LINEs,

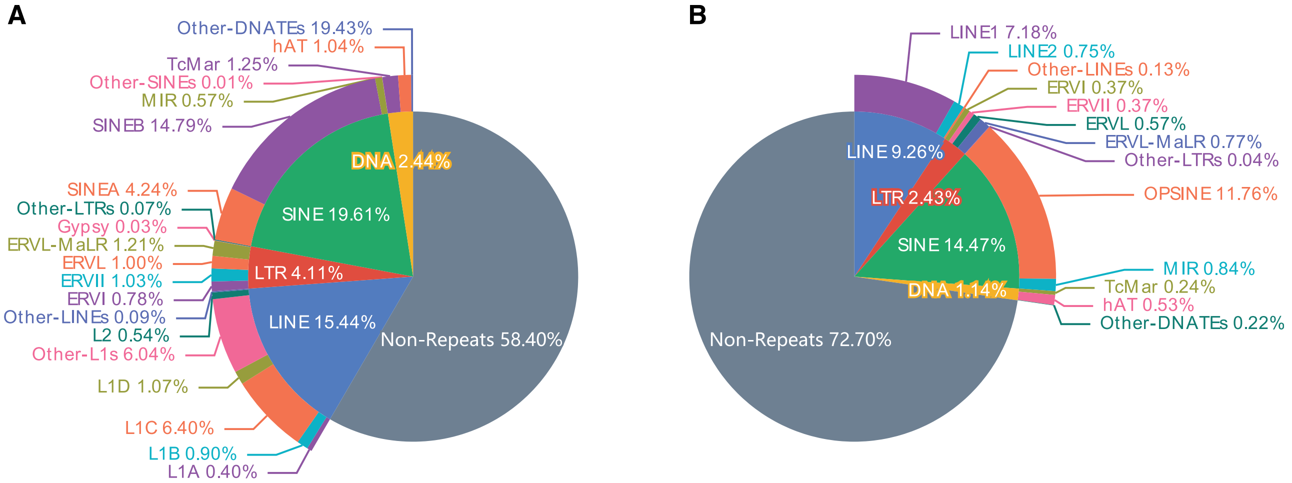

Fig. 4.—Mobilome annotation in the rabbit (*A*) and pika (*B*). Inner circle represents the percentage of nonrepeats and four classes of TEs annotated in the genome, including SINE (short-interspersed nuclear elements), LINE (long-interspersed nuclear elements), LTR (long-terminal repeats), and DNA (DNA transposons). The outer circle indicates the families and the proportion of each in each class of TE.

accounting for 15.44% of the genomic sequences (402 Mb) with more than 84,000 occurrences. Conversely, LTRs displayed substantially lower abundance, representing 4.11% of genomic coverage (~107 Mb) with 335,570 occurrences compared with SINEs and LINEs (fig. 4*A* and supplementary table S5, Supplementary Material online). A very similar trend was observed in the close species, pika, where SINEs also represented the highest occurrence and genomic coverage (14.47%), followed by LINEs (9.26%), and LTRs (2.43%) although these types of retrotransposons occupied a substantially lower proportion of genomic sequence compared with that in rabbit and other mammals (fig. 4*B* and supplementary table S10, Supplementary Material online).

Extensive Impact of Retrotransposons on lncRNAs and Protein-Coding Genes in Rabbits

To investigate the impact of retrotransposons on lncRNAs and protein-coding genes in rabbits, intersection analysis was conducted between genic regions and TE insertions. Overall, 36.76% (1,293,053) and 7.77% (273,436) of TE insertions (3,517,829) overlapped with protein-coding and lncRNA genes, respectively, with 2.26% (79,333) of the TE insertions overlapping the protein-coding and lncRNA genes (fig. 5*A* and supplementary table S11, Supplementary Material online). Retrotransposons displayed extensive impact on the lncRNA and protein-coding genes. In general, approximately 80% (16,296) of protein-coding genes and 90% (12,926) of lncRNA genes contained retrotransposon (RTn) insertions (fig. 5*B* and supplementary table S12, Supplementary Material online). SINEs displayed the greatest impact on lncRNAs and protein-coding genes, followed by LINEs and LTRs. Specifically, 77.09% (15,664) of protein-coding genes and 84.24% (11,932) of lncRNA genes contained SINE insertions, occupying approximately 46% (890,155) of the total SINE insertions (1,953,795) in the genome, respectively. Of protein-coding genes, 59.11% (12,009) harbored LINE insertions, and 58.79% (8,328) of lncRNA genes, representing approximately 37% (311,221) of all LINE insertions (841,004). A total of 47.58% (9,668) of protein-coding genes and 52.71% (7,466) of lncRNA genes overlapped with LTR insertions, occupying approximately 32% (106,124) of all LTR insertions (335,570) (fig. 5*A* and *B*; supplementary tables S12 and S13, Supplementary Material online).

TE coverages for different genic features revealed many biases in their TE composition (fig. 5*C* and supplementary table S13, Supplementary Material online). Significant depletion of TEs in transcribed regions (including the 5′-UTR, exons, Coding sequence CDS, and 3′-UTR) of protein-coding genes was observed, with less than 5% of sequence coverage in these regions, although protein-coding gene introns (37.72%), lncRNA introns (41.73%), and exons (37.72%) did not display significant bias toward the distribution of TEs, with a similar level to the mean TE coverage (39.57%) in the genome (fig. 5C and supplementary table S13, Supplementary Material online). In addition, we also analyzed TE coverage for each kb region within the sequences 10 kb upstream and 10 kb downstream of these genes, which usually contain regulatory sequences (Sharan et al. 2007). Significant depletion of TEs was observed in the majority of regions 1 kb upstream and 1 kb downstream of protein-coding, lncRNA, and other genes, except for the 1 kb upstream of lncRNA genes (fig. 5*D–F* and supplementary table S14, Supplementary Material online).

Impact of Genomic GC Content on the Retrotransposon Insertion Profile

To investigate the impact of GC content on retrotransposon insertion configurations, the GC content of different genomic

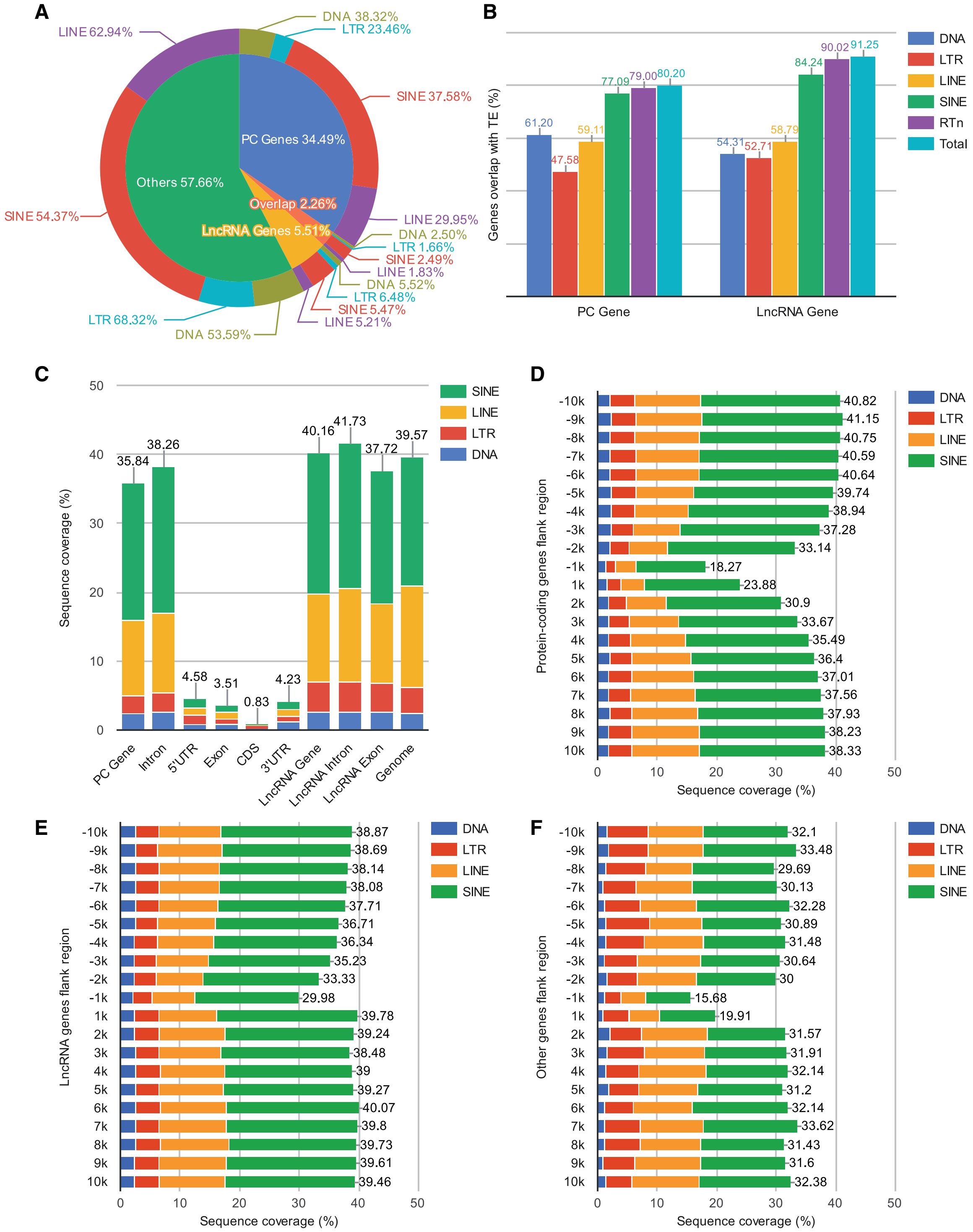

Fig. 5.—Distribution of TEs relative to genic regions. (*A*) The proportion of TEs overlapping with protein-coding genes (PC genes), and long noncoding genes (lncRNA genes). Inside the pie chart represents the proportion of TEs distributed within these regions, whereas outside represents the proportion of different types of TE in each region. (*B*) Proportion of genes overlapped with different types of TE in each PC and lncRNA gene. (*C*) Proportion of genic features overlapped with TEs. Detailed information is available in supplementary table S13, Supplementary Material online. (*D–F*) Density of TEs in the flank regions of protein-coding genes and lncRNA genes, respectively. The ordinate axis represents the position relative to the gene, negative and positive numbers representing upstream and downstream of the gene, respectively. DNA, DNA transposons, LTR, long-terminal repeats; LINE, long-interspersed nuclear elements; SINE, short-interspersed nuclear elements; RTn, all retrotransposons. Detailed information is available in supplementary table S14, Supplementary Material online.

features was calculated, which were then associated with the retrotransposon insertion profiles in genomic regions. Local GC content in 50-kb windows had significantly negative correlations with LINE (r_s=−0.996) and LTR (r_s=−0.829) insertions, but very weakly negative correlations with SINE (r_s=−0.240) and DNA repeat (r_s=−0.134) insertions (fig. 6*A*). This suggested that the local GC content may play roles in shaping LINE and LTR insertions in genome, but has limited impact on the genomic distributions of SINE and DNA TEs. In particular, significant enrichment of LINEs in chromosome X was observed, with 491 TE insertions per Mb, whereas other chromosomes only contained TE insertions ranging from 212 to 361 insertions (fig. 6*B* and supplementary table S15, Supplementary Material online). The impact of GC content on TE insertion patterns in the flanking regions (10 kb upstream and 10 kb downstream) of protein-coding and lncRNA genes was investigated. We found that the closest upstream region (−1 kb) of protein-coding and lncRNA genes, close to the core promoter which tended to be enriched in gene regulatory elements (Sharan et al. 2007), displayed a higher GC content than the other flanking regions. Accordingly, all TE insertions appeared to be repressed in this region, particularly for SINE insertions, where it decreased from more than one insertion per kb to approximately 0.6 insertions per kb in protein-coding genes and approximately 0.8 insertions per kb in lncRNA genes (fig. 6*C* and supplementary table S16, Supplementary Material online). Enrichment of SINEs was observed in the 1 kb downstream of lncRNA genes, and depletion of SINEs and LINEs in the 1 kb downstream of protein-coding genes, apparently unassociated with the GC content of this region (fig. 6*C* and supplementary table S16, Supplementary Material online). In addition, compared with other genic regions (protein-coding genes, and protein-coding gene introns, lncRNA genes, and lncRNA introns and exons) and chromosomes, significantly higher GC content was observed in the 5′-UTRs, CDS, and exons (fig. 6*D* and supplementary table S17, Supplementary Material online), where retrotransposons in these regions were significantly depleted (fig. 6*C* and supplementary table S16, Supplementary Material online), indicating again that GC content may play roles in the accumulation of retrotransposons in these regions. However, selection could also play a role in the depletion of TE insertions in these genic regions (Brunet and Doolittle 2015), since it is commonly accepted that TEs contribute to the genetic adaptation (Chénais et al. 2012; Cosby et al. 2019).

Discussion

Contribution of the Mobilome to Genome Size Variations of Lagomorphs

It is commonly accepted that the TE landscape plays an important role in determining the size of the genome. Usually, larger genomes contain a higher TE content and vice versa (Hawkins et al. 2006; Chalopin et al. 2015; Zhou et al. 2020), and in both plant and animal kingdoms, genome size is positively correlated with the size of the mobilome in many lineages (Hawkins et al. 2006; Gao et al. 2016). Repetitive sequences are the most prevalent feature of mammalian genomes, occupying about half of the genomic sequence. However, differential TE profiles have also been observed in different lineages of mammals (Mandal and Kazazian 2008; Platt et al. 2018). It has been suggested that the smaller genomes of both dogs (2.40 Gb) and cats (2.52 Gb) are due to less lineage-specific repeat sequences compared with that of the human (3.04 Gb) and mouse (2.73 Gb) (Lindblad-Toh et al. 2005; Pontius et al. 2007). In the present study, mobilome annotation in the rabbit revealed that repeats accounted for approximately 42% of the rabbit genome (2.74 Gb), lower than that of the human (45%) (International Human Genome Sequencing Consortium 2001), but higher than mice (38%) (Waterston et al. 2002), generally agreeing with the trend in genome size difference. In the pika, the closest sequenced species of lagomorph to the rabbit (Chapman and Flux 2008), TEs only account for approximately 24% of the genomic sequence (fig. 4*B*), possibly representing the lowest genomic coverage in mammals, and also largely explains it having the smallest size of genome (2.23 Gb). Low TE coverage in the genome may be explained by a lower level of TE amplification and/or fast TE loss, whereas the rate of TE loss is relative to the mutation rate of the genome (Waterston et al. 2002). In unicellular organisms, mutation rates of RNA and DNA viruses vary inversely with genome size (Drake 1991; Drake et al. 1998; Sung et al. 2012; Bradwell et al. 2013), thus the mutation rate may play a role in shaping the evolution both of the mobilome and genome in animals. It has been suggested that the very large genomes of salamanders is due to slow DNA loss of TEs (Sun et al. 2012). So far, there are three reports of mutation rate estimations (1.62×10^{-9}, 1.74×10^{-9}, and 2.35×10^{-9}) for rabbits (Carneiro et al. 2009, 2011, 2012), in which the mean mutation rate (1.99×10^{-9}) is approximately four times faster than that in humans (0.5×10^{-9}), but three times slower than in mice (5.4×10^{-9}) (Scally and Durbin 2012; Fu et al. 2014; Uchimura et al. 2015; Scally 2016), which appears consistent with the differences in sizes of the mobilomes and genomes, supporting again the possibility that mutation rate may shape the evolution of the mobilome and genome. Furthermore, this hypothesis is also well-supported by the distribution of MIRs in these species, which are relatively ancient, and these MIRs may amplify prior to mammalian radiation, then transfer to the genomes of rabbits, pikas, mice, and humans. We found that the quantity of MIRs in the pika (18.8 Mb), rabbit (14.8 Mb) (supplementary tables S10 and S11, Supplementary Material online), and mouse (14.1 Mb) genomes (Waterston et al. 2002) was considerably lower than in humans (69.4 Mb) (International Human Genome Sequencing Consortium 2001), supporting

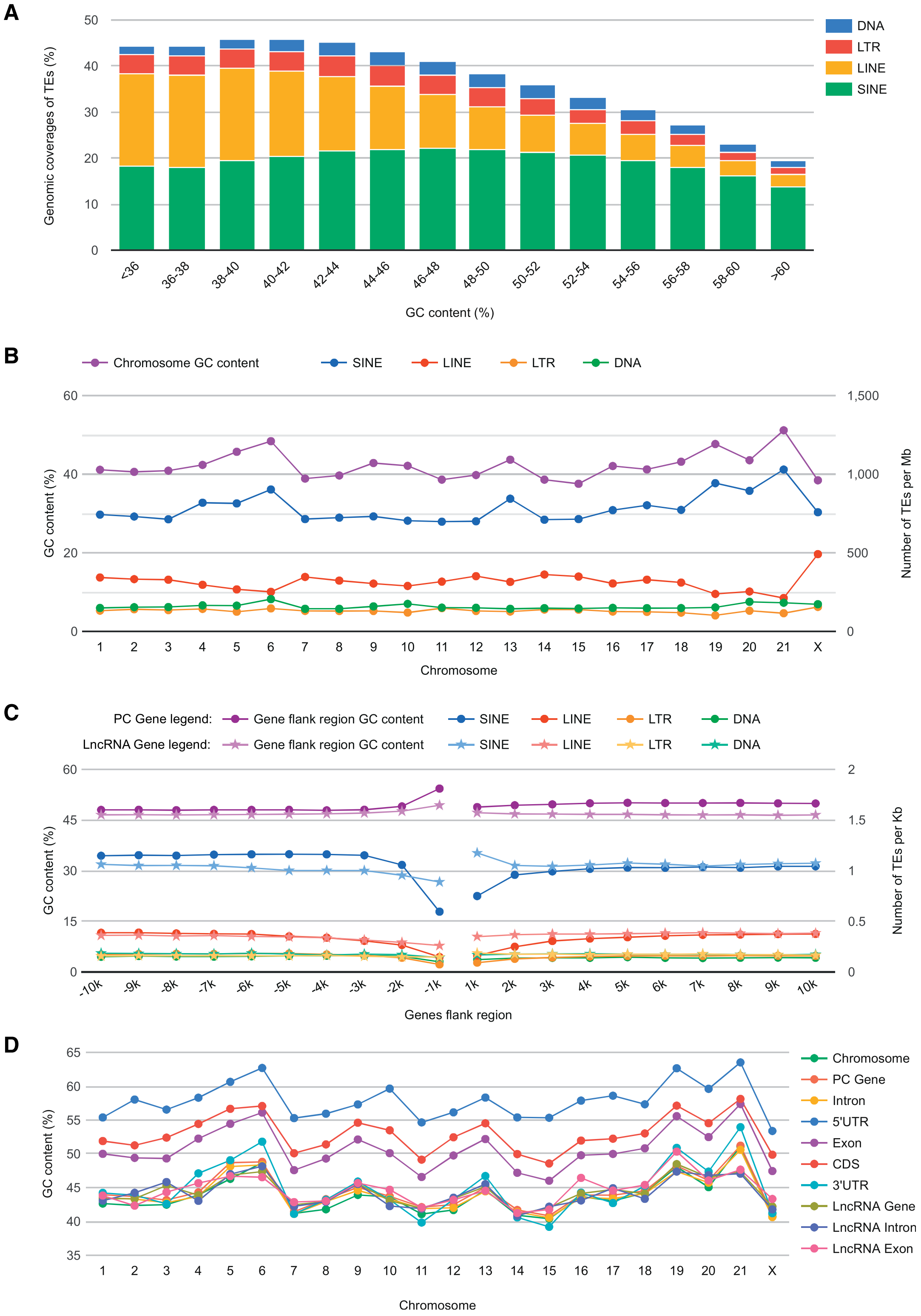

Fig. 6.—Impact of genomic GC content on retrotransposon distribution in the genome. (*A*) Density of the major repeat classes as a function of local GC content, in windows of 50 kb. (*B*) GC content (excluding TEs) and TE density of each chromosome. (*C*) GC contents (excluding TEs) and TE density in the flank regions of protein-coding (PC) genes and long noncoding (lncRNA) genes. (*D*) GC content of different genomic features in each chromosome. DNA, DNA transposons, LTR, long-terminal repeats; LINE, long-interspersed nuclear elements; SINE, short-interspersed nuclear elements; RTn, retrotransposons; CDS, coding sequence; 5′-UTR, 5′-untranslated region; 3′-UTR, 3′-untranslated region.

the supposition that the rate of loss of MIR sequences in the pika, rabbit, and mouse was faster than in the human genome after mammalian radiation, indicating that fast DNA loss of TEs may also contribute to a relatively small genome size.

Distinct Retrotransposon Landscape in Lagomorphs

Three major classes of retrotransposons, including LINEs, SINEs, and LTRs/ERV, dominate mammalian genomes. LINEs represent the most abundant type of repeats by greatest genomic coverage in almost all annotated mobilomes of mammals, including the human, gorilla, chimpanzee, orangutan, gibbon, representing primates, and cat, dog, panda, and seal for carnivores, mouse, rat, squirrel and vole for rodents, odd-toed ungulates (horse), even-toed ungulates (alpaca, cow, and pig), and bats (Smit et al. 2013–2015), followed by SINEs and LTRs. However, in the current study, annotation revealed a distinct retrotransposon landscape in the genomes of lagomorphs (rabbit and pika), together with hedgehog and tree shrew (http://www.repeatmasker.org/, last accessed July 25, 2021), compared with the most surveyed mammalian genomes, where SINEs represented the highest genomic element in these lineages, followed by LINEs, and LTRs, respectively. SINEs displayed the fastest and most successful amplification in the genomes of both the rabbit and pika. Currently, rabbit-specific SINEs (approximately two million occurrences) accounted for up to one-fifth (495.5 Mb) of genomic sequences, representing not only the highest genomic coverage but also the largest genome size masked compared with the distribution of SINEs in all mobilome-annotated mammals (Platt et al. 2018). Furthermore, the data also demonstrated that more than 85% of OcuSINEA insertions, the youngest family of SINEs, were polymorphic, substantially higher than that of the pig (25%) (Chen et al. 2021), indicating transposition activity of recent SINEs is significantly higher than in the pig, which may also contribute to their large amplification in the rabbit. On the other hand, only 0.57% (14.8 Mb) and 0.57% (18.8 Mb) of nonlineage specific SINEs (mainly MIRs) were identified in the rabbit and pika genomes, respectively, indicating that the majority of MIRs have been lost in this lineage, playing limited roles in shaping the recent evolution of lagomorph genomes.

The rabbit-specific SINEs identified here were classified into two families (OcuSINEA, OcuSINEB) and nine subfamilies (OcuSINEA1-4 and OcuSINEB1-5) (fig. 2*A* and supplementary fig. S1, Supplementary Material online) based on sequence similarity and length. All rabbit-specific SINEs were derived from tRNA sequences and harbored a conservative internal polymerase III promoter (box A and box B) (fig. 2*C* and supplementary fig. S1, Supplementary Material online). Conversely, the only active Alu SINE in the human genome was derived from 7SL RNA (International Human Genome Sequencing Consortium 2001), whereas the mouse has both tRNA-derived and 7SL-derived SINEs (Waterston et al. 2002). The length of putatively active SINEs in the rabbit was approximately 330 bp (excluded the poly-A tail), which is longer than the active SINEs in mouse (B1: 135 bp and B2: 175 bp) and human (Alu: 280 bp), respectively. In general, SINEs, ranging from 100 to 700 base pairs in length (Vassetzky and Kramerov 2013), were too short and do not have the protein-coding capacity. although a number of SINEs were active and able to reverse-transcribe and insert into new positions in the mammalian genome through active partnership with LINEs (Hancks and Kazazian 2012). Active LINEs can encode proteins and assist SINEs in being reverse-transcribed and integrated back into the genome, thus a number of SINEs and LINEs have coevolved, such as MIR and LINE2, which have coevolved and dominated the ancient genome in mammals (Smit et al. 1995; Smit 1996). Here, the coevolution of MIRs and LINEs in rabbits was observed between 120 and 60 Ma (figs. 1*C* and 2*D*). The activity of LINE2 began from approximately 120 Ma, peaking at 95 Ma, finally becoming extinct around 60 Ma, followed by that of MIR activity. With the extinction of LINE2s and MIRs, rabbit-specific LINE1s and SINEs emerged and coevolved over the last 100 Myr (figs. 1*D* and 2*E*). Furthermore, rabbit-specific L1s were classified into four distinct families with differential evolution profiles, which have also been observed in pigs (Chen et al. 2019), mice (Waterston et al. 2002), and humans (International Human Genome Sequencing Consortium 2001). In addition, only one family of pigs and human L1s is active, with approximately 100 intact copies in their genomes (Chen et al. 2019), whereas two families (L1B and L1D) of rabbit L1s are putatively active, with approximately 30 and 40 intact copies identified in the families of L1B and L1D, respectively, possibly also contributing to the high activity and large accumulation of SINEs in the rabbit. The number of intact copies of L1s in these species (rabbit, pig, and human) is significantly lower than in rodents, with approximately 3,000 intact L1 copies identified in mice (Waterston et al. 2002).

ERVs are also the major retrotransposons within mammalian genomes, constituting 5–10% of genomic sequences (International Human Genome Sequencing Consortium 2001; Belshaw et al. 2004; McCarthy and McDonald 2004; Nelson et al. 2004). Here, we found ERVs displayed low genomic coverages in both the rabbit (~4%) and pika (~2%), substantially lower than in the majority of other mammalian genomes (International Human Genome Sequencing Consortium 2001; Waterston et al. 2002; Lindblad-Toh et al. 2005; Li et al. 2010; Chen et al. 2019). High ERV activity has been observed in rodents, in which all three classes (even MaLR) have active members in the mouse (Waterston et al. 2002), whereas no full length of ERV was identified in the rabbit, suggesting that the activity of ERVs in this species is limited, very similar to ERVs in humans (International Human

Genome Sequencing Consortium 2001) and pigs (Chen et al. 2019).

Distribution Bias of Retrotransposons in the Rabbit Genome

The extensive intersection of retrotransposons with protein-coding and lncRNA genes, and significant distribution bias of retrotransposons was observed in the genome of the rabbit. Similar to observations in the human (Burns and Boeke 2012) and pig (Chen et al. 2019), approximately 80% of protein-coding genes and more than 90% of lncRNA genes contained retrotransposon-derived sequences, most retrotransposons tending to insert into introns, whereas exons of protein-coding regions were almost devoid of retrotransposons, but significant deletions in lncRNA exons were not observed (figs. 4*D–F* and 5*C*). Furthermore, regions immediately upstream and downstream (1 kb) of both protein-coding and lncRNA genes tended to not have retrotransposon insertions. These data indicate that retrotransposon insertions in genomes experienced strong selection, whereas a number of genic regions were not tolerable for retrotransposon insertions. This is particularly the case for a number of key regulator genes, such as *hox* genes, which have been extensively studied and it has been suggested that they play a crucial role in the regulation of cell differentiation and embryonic development (Krumlauf 1994; Pearson et al. 2005; Sheth et al. 2012), and are the poorest regions for repeats in the human and murine genomes (International Human Genome Sequencing Consortium 2001; Waterston et al. 2002). A similar trend was observed for the four *hox* gene clusters in the rabbit genome, where repeats were almost totally depleted, with less than 1% of sequences represented by repeats (data not shown). These data suggest again that retrotransposon insertions were dependent on strong purification selection with retrotransposon distribution bias existing in the genome. Additionally, we found that local GC content had negative correlations with LINE and LTR insertions in genome. On the other hand, we also found that all retrotransposons, particularly SINEs, were highly depleted in the immediate upstream regions (−1 kb) of protein-coding and lncRNA genes, which tended to enrich gene regulatory elements and putatively core promoter regions (Sharan et al. 2007), with higher GC content than other flanking regions (fig. 5*C* and supplementary table S14, Supplementary Material online). These data suggest that GC content may also play a role in shaping the retrotransposon distribution in the genome.

Conclusions

In the present study, de novo mining of retrotransposons in the rabbit genome was performed, allowing us to define their diversity at a family and subfamily level, and the structural organization, evolutionary dynamics, and distribution within the genome. Four families (L1A, L1B, L1C, and L1D) and 19 subfamilies of L1, two SINE families (OcuSINEA, OcuSINEB), nine subfamilies of SINEs, and 12 ERV families (OcuERV1–OcuERV12) were characterized in terms of sequence identity, structural organization, and the creation of a phylogenetic tree. Differential evolutionary dynamics across these families and subfamilies were recorded. Insertion age analysis and insertion polymorphism identification demonstrated that SINEA, L1B, L1D, and ERV1 are relatively young families, with some copies from these families may still active and able to jump within the genome. Our analysis revealed a distinct retrotransposon landscape in lagomorphs, represented by rabbits and pika, in which SINEs displayed the highest genome coverage, very different from most mobilome-annotated mammals. We also found that the majority of retrotransposons overlap with lncRNA (>90%) and protein-coding genes (>80%), with significant retrotransposon insertion bias observed for relative genic regions. The data are also consistent with GC content possibly shaping the distribution of retrotransposons in the genome. These findings describe retrotransposon evolution in the rabbit and provide a better understanding of genomic evolution in lagomorphs, and possibly mammals. Furthermore, putatively active retrotransposons, especially the most recent SINEs, can generate high polymorphic insertions, as identified in this study, which may allow future genetic marker development, which could have considerable potential for applications in quantitative trait locus (QTL) mapping and molecular breeding in rabbits, and so are worthy of additional evaluation.

Materials and Methods

Sampling

Ear samples from wild and domestic rabbit breeds were collected so that retrotransposon insertion polymorphisms could be identified by PCR. For wild rabbits, cape hares (*Lepus capensis*) were obtained from local hunters in Jiangsu, Anhui, Shandong, Sichuan, Henan, and Hebei Provinces, three individuals from each province used for DNA extraction. Three individuals of each domesticated rabbit breed, including the Californian rabbit, Sichuan white rabbit, ZIKA meat rabbit, Fujian yellow rabbit, and New Zealand white rabbit were obtained from the Sichuan Animal Sciences Academy. Three Belgian rabbits were obtained from Jiangsu Academy of Agricultural Sciences and the three Rex rabbits were provided by a rabbit breeding farm, Xinnong Rabbit Co., Ltd (Yuyao, Zhejiang Province). The total genomic DNA was then extracted from the ear using a TIANamp genomic DNA kit (Tiangen, Beijing, China) in accordance with the manufacturer's instructions. All experiments with rabbits were performed in accordance with the ethical regulations of the Animal Care and Use Committee of Yangzhou University, Yangzhou, China.

Retrotransposon Mining in the Rabbit Genome

A rabbit reference genome (OryCun2.0) was downloaded from Ensembl database release 94 (see URLs). The MGEScan-non-LTR utility (Rho and Tang 2009) was used for de novo identification of L1s in the rabbit genome, which identifies non-LTR retrotransposons based on the probabilistic models (Hidden Markov Model). However, most L1 elements identified by the MGEScan-non-LTR tool were incomplete, and thus they were mapped onto the rabbit genome again using the BLAT tool (Kent 2002) (−minIdentity=100, −minScore=200), after which the sequences were extended 2,500 bp for the 5′ flanks and 500 bp for the 3′ flanks to obtain full-length copies of in the genome using the slop command (−s, −l 2,500, −r 500) in bedtools (version 2.27.1) (Quinlan and Hall 2010). Additionally, available L1 genomic sequences of the rabbit were downloaded from the L1Base database (Penzkofer et al. 2016) and then merged with the L1 sequences identified above. Redundancy was removed (locus distance within 3,000 nt of the same strand). The getfasta command in bedtools was used to extract the L1 sequences using their genomic coordinates. These elements were clustered based on sequence similarity using usearch (see URLs), with only clusters having more than ten elements remaining for further analysis. The boundaries of these L1 elements were defined manually after alignment with Clustal W (version 2.1) (Larkin et al. 2007) software. Consensus sequences were derived for each cluster using Geneious Prime 2019.2.1 (see URLs).

SINEs were de novo identified using SINE_Scan (Mao and Wang 2017) and RepeatModeler (version 2.0) tools. RepeatModeler uses two core programs, RECON (version 1.08) and RepeatScout (version 1.06) to de novo recognize the repetitive sequences (see URLs). SINEs were mined using RepeatModeler three times, whereas SINE_Scan was used to identify SINE elements using default parameters. These SINEs (<500 bp in length), were combined with known SINEs downloaded from RepBase (Bao et al. 2015) (see URLs), merged then redundancy removed.

Full-length ERV retrotransposons were firstly identified in the rabbit genome using the LTRharvest (Ellinghaus et al. 2008) tool embedded in GenomeTools (Gremme et al. 2013) using the parameters: -motif tgca -minlenltr 100 -maxlenltr 5,000 -mindistltr 1,000 -maxdistltr 20,000 -similar 80 -motifmis 1 -mintsd 4 -maxtsd 20 -overlaps best. LTRharvest is a tool for the de novo detection of full length LTR retrotransposons (including ERV) (Ellinghaus et al. 2008). In addition, ERVs were also mined in the rabbit genome using RetroTector (Sperber et al. 2007), whereas all predicted ERVs were further annotated for protein domains using the HMMER tool (version 3.3, see URLs) and hidden Markov model profiles downloaded from the Pfam (El-Gebali et al. 2019) and GyDB databases (see URLs). tRNAscan-SE (Chan and Lowe 2019) (version 1.3.1) was used to search tRNA genes and construct a database of tRNAs to predict the location of each PBS.

Phylogenetic Analysis

In consideration of the presence of numerous frame-shift mutations and stop codons in ancient retrotransposon elements, DNA sequences of L1, SINE, and RT regions of ERV retrotransposons were used to construct multiple alignments and build a phylogenetic tree. Multiple alignments of the L1, OcuSINE, and RT consensus sequences were conducted using Clustal W. A maximum likelihood tree was generated from the alignments using MEGA X (Kumar et al. 2018) with a Kimura 2-parameter model and bootstrap values selected for 1,000 replications. Phylogenetic trees were further customized using the Evolview online tool (He et al. 2016) (see URLs). Reference RT sequences of ERVs, used for phylogenetic analysis and definitions for classification of rabbit ERVs, were downloaded from the NCBI nucleotide database.

TE Annotation of Rabbit Genomes

RepeatMasker (version 4.0.9) software with RMblast was used as a sequence search engine to annotate TEs in the rabbit genome (OryCun2.0) using a custom library, which combined the known repeats in the rabbit genome from the Repbase (version 20181026) (Bao et al. 2015) and Dfam databases (3.0) (Hubley et al. 2016), in which the cutoff value was set to 250.

The coordinates of protein-coding genes (20,318) and internal exons (236,742), mRNAs (31,990), 5′-UTRs (34,311), 3′-UTRs (22,665), and CDS (196,710), in addition to tRNA genes and other genes were obtained using NCBI annotation release 102 (see URLs). lncRNA transcripts (25,082) and the coordinates of lncRNA genes (11,135) were obtained by lncRNA-seq and data were deposited in the short read archive (SRA) of NCBI under bioproject number PRJNA479733. About 10-kb sequences both upstream and downstream of protein-coding and lncRNA genes were extended based on gene coordinates. The sequences of the genes and flanks were extracted from the genomes using the bedtools getfasta tool in accordance with these coordinates. Overlaps (>10 bp) of TEs with protein-coding and lncRNA genes and their internal features, including 10 kb flanking regions, were calculated using bedtools. Recording of the data of the sequences, modifications and extraction, data format conversions, result file parsing, removal of duplicate data, and TSD sequence recognition were executed using a custom Python (version 3.7.3) script. Data were visualized using bar and line charts generated using Google chart tools (see URLs) and pie charts generated using Echarts (Li et al. 2018) (see URLs).

Retrotransposon Insertion Polymorphism Detection Using PCR

A total of 71 primer pairs were designed for intact L1 copies (two for L1A, 28 for L1B, three for L1C, and 38 for L1D), 325 primer pairs for SINE families (50 primer pairs for each OcuSINEA subfamily and 25 primer pairs for each OcuSINEB subfamily), 36 primer pairs for full ERVs, and 20 primer pairs for OcuNERV family members (10 primer pairs for each subfamily) for insertion polymorphism detection using PCR with Primer3 software (Untergasser et al. 2012) (version 2.3.7). Flanking regions of SINEs, 5′-UTRs of LTRs and L1s, and their flanking region sequences were used as inputs for primer design. Product sizes varied from 400 to 800 bp. PCR reactions were performed using a T100 Thermal Cycler (Bio-Rad) using standard cycling conditions and PCR products were subjected to electrophoresis to check for presence or absence of respective inserts in wild and domesticated rabbits. All primer sequences, expected PCR product sizes, and the positions of corresponding retrotransposons are detailed in supplementary tables S2, S5, S7, and S9, Supplementary Material online.

Statistical Analyses and Insertion Time Estimation

Spearman rank correlation coefficients were used to determine the correlation between the distribution of TEs and local GC content using SPSS software (version 25.0; Chicago, IL).

The Perl script calcDivergencefromAlign.pl (obtained from the RepeatMasker package) was used to calculate Kimura divergence values (*K*), with a mean substitution rate (*r*) of 1.99×10^{-9} mutations per site per year used for insertion age analysis, in accordance with estimates by Carneiro et al. (2009, 2011, 2012). The formula $T = K/2r$ (Kimura 1980) was then employed to measure insertion time (*T*).

URLs

Ensembl database: https://www.ensembl.org (last accessed July 25, 2021); MGEScan-non-LTR: https://github.com/MGEScan/mgescan (last accessed July 25, 2021); BLAT: http://genome.ucsc.edu/ (last accessed July 25, 2021); bedtools: https://bedtools.readthedocs.io/en/latest/ (last accessed July 25, 2021); L1Base database: http://l1base.charite.de/l1base.php (last accessed July 25, 2021); usearch: http://www.drive5.com/usearch/ (last accessed July 25, 2021); Clustal W: http://www.clustal.org/clustal2/ (last accessed July 25, 2021); Geneious Prime: https://www.geneious.com (last accessed July 25, 2021); RepeatModeler: http://www.repeatmasker.org/RepeatModeler/ (last accessed July 25, 2021); SINE_Scan: https://github.com/maohlzj/SINE_Scan (last accessed July 25, 2021); RECON: http://eddylab.org/software/recon/ (last accessed July 25, 2021); RepeatScout: https://bix.ucsd.edu/repeatscout/ (last accessed July 25, 2021); RepBase: https://www.girinst.org/repbase/ (last accessed July 25, 2021); GenomeTools and LTRharvest: http://genometools.org/ (last accessed July 25, 2021); RetroTector: https://github.com/PatricJernLab/RetroTector (last accessed July 25, 2021); HMMER: http://www.hmmer.org/ (last accessed July 25, 2021); Pfam: http://www.pfam.org/ (last accessed July 25, 2021); GyDB: http://gydb.org/ (last accessed July 25, 2021); tRNAscan-SE: http://lowelab.ucsc.edu/tRNAscan-SE/ (last accessed July 25, 2021); MEGA X: https://www.megasoftware.net/ (last accessed July 25, 2021); Evolview: http://www.evolgenius.info/evolview/ (last accessed July 25, 2021); NCBI: https://www.ncbi.nlm.nih.gov/ (last accessed July 25, 2021); RepeatMasker: http://www.repeatmasker.org/ (last accessed July 25, 2021); Dfam: https://dfam.org/ (last accessed July 25, 2021); NCBI rabbit genome annotation release 102: https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation_releases/9986/102/GCF_000003625.3_OryCun2.0/ (last accessed July 25, 2021); Python: https://www.python.org/ (last accessed July 25, 2021); Google chart tools: https://developers.google.com/chart/ (last accessed July 25, 2021); Echarts: https://echarts.apache.org/ (last accessed July 25, 2021).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by Modern Agricultural Industrial System Special Funding (CARS-43-A-1), National Natural Science Foundation of China (Grant No. 32072724), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Data Availability

All data required for evaluation of the conclusions in this article are present either in the main text or Supplementary Material online.

Literature Cited

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

Beerens N, Groot F, Berkhout B. 2001. Initiation of HIV-1 reverse transcription is regulated by a primer activation signal. J Biol Chem. 276(33):31247–31256.

Belshaw R, et al. 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A. 101(14):4894–4899.

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff J-N. 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res. 16(1):203–215.

Borodulina OR, Kramerov DA. 2001. Short interspersed elements (SINEs) from insectivores. Two classes of mammalian SINEs distinguished by A-rich tail structure. Mamm Genome. 12(10):779–786.

Borodulina OR, Kramerov DA. 2008. Transcripts synthesized by RNA polymerase III can be polyadenylated in an AAUAAA-dependent manner. RNA 14(9):1865–1873.

Bourque G, et al. 2018. Ten things you should know about transposable elements. Genome Biol. 19(1):1–12.

Bradwell K, Combe M, Domingo-Calap P, Sanjuán R. 2013. Correlation between mutation rate and genome size in riboviruses: mutation rate of bacteriophage Qβ. Genetics 195(1):243–251.

Brunet TD, Doolittle WF. 2015. Multilevel selection theory and the evolutionary functions of transposable elements. Genome Biol Evol. 7(8):2445–2457.

Burns KH, Boeke JD. 2012. Human transposon tectonics. Cell 149(4):740–752.

Carneiro M, et al. 2011. The genetic structure of domestic rabbits. Mol Biol Evol. 28(6):1801–1816.

Carneiro M, et al. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Mol Biol Evol. 29(7):1837–1849.

Carneiro M, et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345(6200):1074–1079.

Carneiro M, Ferrand N, Nachman MW. 2009. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). Genetics 181(2):593–606.

Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 7(2):567–580.

Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol. 1962:1–14.

Chapman JA, Flux JE. 2008. Introduction to the lagomorpha. Lagomorph Biology. New York: Springer. p. 1–9.

Chen C, et al. 2019. Retrotransposons evolution and impact on lncRNA and protein coding genes in pigs. Mob DNA. 10:19.

Chen C, et al. 2021. SINE jumping contributes to large-scale polymorphisms in the pig genomes. Mob DNA. 12(1):17.

Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene 509(1):7–15.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18(2):71–86.

Clutton-Brock J. 1999. A natural history of domesticated mammals. Cambridge: Cambridge University Press.

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

Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict, cooperation, and cooption. Genes Dev. 33(17–18):1098–1116.

Damgaard CK, Andersen ES, Knudsen B, Gorodkin J, Kjems J. 2004. RNA interactions in the 5′ region of the HIV-1 genome. J Mol Biol. 336(2):369–379.

Douzery EJ, Huchon D. 2004. Rabbits, if anything, are likely Glires. Mol Phylogenet Evol. 33(3):922–935.

Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A. 88(16):7160–7164.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148(4):1667–1686.

Dutta S, Sengupta P. 2018. Rabbits and men: relating their ages. J Basic Clin Physiol Pharmacol. 29(5):427–435.

El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9:18.

Fan J, et al. 2018. Principles and applications of rabbit models for atherosclerosis research. J Atheroscler Thromb. 25(3):213–220.

Figiel M, et al. 2018. Mechanism of polypurine tract primer generation by HIV-1 reverse transcriptase. J Biol Chem. 293(1):191–202.

Fu Q, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514(7523):445–449.

Gao B, et al. 2016. The contribution of transposable elements to size variations between four teleost genomes. Mob DNA. 7(1):1–16.

Geiduschek EP, Tocchini-Valentini GP. 1988. Transcription by RNA polymerase III. Annu Rev Biochem. 57:873–914.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 10(3):645–656.

Hancks DC, Kazazian HH Jr. 2012. Active human retrotransposons: variation and disease. Curr Opin Genet Dev. 22(3):191–203.

Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. Genome Biol. 5(6):225–226.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. 16(10):1252–1261.

He Z, et al. 2016. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. 44(W1):W236–W241.

Hubley R, et al. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44(D1):D81–D89.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431–438.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. Science 303(5664):1626–1632.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12(4):656–664.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16(2):111–120.

Krumlauf R. 1994. *Hox* genes in vertebrate development. Cell 78(2):191–201.

Kumar S, et al. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35(6):1547–1549.

Kumari V, et al. 2011. Differential distribution of a SINE element in the *Entamoeba histolytica* and *Entamoeba dispar* genomes: role of the LINE-encoded endonuclease. BMC Genomics 12:267–212.

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948.

Li D, et al. 2018. ECharts: a declarative framework for rapid construction of web-based visualization. Vis Inform. 2(2):136–146.

Li R, et al. 2010. The sequence and *de novo* assembly of the giant panda genome. Nature 463(7279):311–317.

Lindblad-Toh K, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069):803–819.

Ma M, et al. 2018. Research on rapid gelatinization of rabbit skin collagen as effect of acid treatment. Food Hydrocolloids. 77:945–951.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 10(9):1307–1318.

Mandal PK, Bagchi A, Bhattacharya A, Bhattacharya S. 2004. An Entamoeba histolytica LINE/SINE pair inserts at common target sites cleaved by the restriction enzyme-like LINE-encoded endonuclease. Eukaryot Cell. 3(1):170–179.

Mandal PK, Kazazian HH. 2008. SnapShot: vertebrate transposons. Cell 135(1):192–192.

Mao H, Wang H. 2017. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics 33(5):743–745.

McCarthy EM, McDonald JF. 2004. Long terminal repeat retrotransposons of *Mus musculus*. Genome Biol. 5(3):R14–R18.

Nelson PN, et al. 2004. Human endogenous retroviruses: transposable elements with potential? Clin Exp Immunol. 138(1):1–9.

Oliver KR, Greene WK. 2009. Transposable elements: powerful facilitators of evolution. Bioessays 31(7):703–714.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating *Hox* gene functions during animal body patterning. Nat Rev Genet. 6(12):893–904.

Penzkofer T, et al. 2016. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acid Res. 45(D1):D68–D73.

Platt RN, Vandewege MW, Ray DA. 2018. Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. 26(1–2):25–43.

Pontius JU, et al. 2007. Initial sequence and comparative analysis of the cat genome. Genome Res. 17(11):1675–1689.

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

Rho M, Tang H. 2009. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Res. 37(21):e143.

Rivas-Carrillo SD, Pettersson ME, Rubin C-J, Jern P. 2018. Whole-genome comparison of endogenous retrovirus segregation across wild and domestic host species populations. Proc Natl Acad Sci U S A. 115(43):11012–11017.

Scally A. 2016. The mutation rate in human evolution and demographic inference. Curr Opin Genet Dev. 41:36–43.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 13(10):745–753.

Sharan R, Karni S, Felder Y. 2007. Analysis of biological networks: transcriptional networks–promoter sequence analysis. Available from: http://www.cs.tau.ac.il/~roded/courses/bnet-a06/lec11.pdf. Accessed July 25, 2021.

Sheth R, et al. 2012. *Hox* genes regulate digit patterning by controlling the wavelength of a Turing-type mechanism. Science 338(6113):1476–1480.

Shpyleva S, Melnyk S, Pavliv O, Pogribny I, James SJ. 2018. Overexpression of LINE-1 retrotransposons in autism brain. Mol Neurobiol. 55(2):1740–1749.

Siefert JL. 2009. Defining the mobilome. Horizontal Gene Transfer. 13–27.

Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. Available from: http://www.repeatmasker.org. Accessed July 25, 2021.

Smit AF. 1996. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev. 6(6):743–748.

Smit AF, Tóth G, Riggs AD, Jurka J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol. 246(3):401–417.

Smith AT. 2021. Rabbit. Encyclopedia Britannica. Available from: https://www.britannica.com/animal/rabbit. Accessed July 25, 2021.

Sperber GO, Airola T, Jern P, Blomberg J. 2007. Automated recognition of retroviral sequences in genomic data—RetroTector. Nucleic Acids Res. 35(15):4964–4976.

Sun C, López Arriaza JR, Mueller RL. 2012. Slow DNA loss in the gigantic genomes of salamanders. Genome Biol Evol. 4(12):1340–1348.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. 109(45):18488–18492.

Uchimura A, et al. 2015. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. 25(8):1125–1134.

Untergasser A, et al. 2012. Primer3—new capabilities and interfaces. Nucleic Acids Res. 40(15):e115.

Vassetzky NS, Kramerov DA. 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. 41(Database issue):D83–D89.

Wang L, et al. 2017. Regression of atherosclerosis with apple procyanidins by activating the ATP-binding cassette subfamily A member 1 in a rabbit model. Atherosclerosis 258:56–64.

Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520–562.

Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973–982.

Wu L, Yao Q, Lin P, Li Y, Li H. 2018. Comparative transcriptomics reveals specific responding genes associated with atherosclerosis in rabbit and mouse models. PLoS One 13(8):e0201618.

Zhou W, Liang G, Molloy PL, Jones PA. 2020. DNA methylation enables transposable element-driven genome expansion. Proc Natl Acad Sci U S A. 117(32):19359–19366.

Associate editor: Ellen Pritham