Transposable element competition in shaping the deer mouse genome

Landen Gozashti

Department of Organismic & Evolutionary Biology, Department of Molecular & Cellular Biology, Museum of Comparative Zoology and Howard Hughes Medical Institute, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA

Abstract

The genomic landscape of transposable elements (TEs) varies dramatically across species, with some TEs demonstrating greater success in colonizing particular lineages than others. In mammals, LINE retrotransposons typically occupy more of the genome than any other TE. Here, we report an unusual genomic landscape of TEs in the deer mouse (*Peromyscus maniculatus*), a model for studying the genomic basis of adaptation. In contrast to nearly all previously examined mammalian species, LTR elements occupy more of the deer mouse genome than LINEs (11% and 10% respectively). This pattern reflects a combination of reduced lineage-specific LINE gain in addition to a massive invasion of lineage-specific endogenous retroviruses (ERVs). Deer mouse ERVs exhibit diverse origins spanning the retroviral phylogeny and one likely arose *de novo* from an exogenous Feline Leukemia Virus within the last ~18 million years. Nonetheless, a few lineage-specific subfamilies account for a considerable proportion of ERV content, with the top five most abundant subfamilies occupying a remarkable ~2% of the genome. We also observe a massive expansion of lineage-specific Kruppel-associated box domain-containing zinc finger genes (KZNFs) which likely occurred to suppress ERV activity, and whose expansion may have been partially facilitated by ectopic recombination between ERVs. Significantly, we find evidence that ERVs directly impacted the evolutionary trajectory of LINEs by outcompeting them for genomic space and interrupting formerly intact LINE machinery. This model likely explains the unique genomic landscape of TEs observed in the deer mouse, and may also explain patterns of LINE inactivity in other mammals. Together, our results illuminate the genomic ecology that shaped the deer mouse genome's unique TE landscape, opening up a range of opportunities to investigate the evolutionary processes that give rise to variation in mammalian genome structure.

Introduction

Transposable elements (TEs) are parasitic genetic elements capable of mobilizing in genomes and function as important drivers of genome evolution [1–3]. In mammals, for example, TEs account for at least 20% of the genome and have been exapted for significant functional innovations [4–6]. Nonetheless, TEs generate mutations when they insert into new positions in the genome and thus represent a significant burden to host fitness. This cost is compounded by the fact that TEs can contain gene regulatory sequences and cause structural rearrangements even after they have lost the ability to move [1,7]. The evolutionary success of a given TE lineage is dictated by its ability to replicate faster than the host genome but limited by its cost to host fitness [3,8]. Thus, TE lineages are in a constant coevolutionary conflict with each other as well as their host [9,10]. Furthermore, hosts have evolved various ways to suppress TE activity [11]. These genetic conflicts embody the "ecology of the genome", and play an important role in shaping the genomic landscape of TEs in a given species, as well as its genome structure more broadly [9,10].

TEs are remarkably diverse and TE landscapes vary dramatically across species [12]. TEs are classified into two broad categories based on their transposition mechanism: class I elements (retrotransposons) which mobilize through an RNA intermediate and class II elements (DNA transposons) which do not. Most eukaryotic lineages harbor a diversity of TEs from multiple taxonomic subgroups within each of these broad classes [1,12]. However, some phylogenetic groups display contrasting patterns in which TE landscapes are relatively similar across species [13,14]. One of these clades is mammals [5]. In most mammalian species, DNA transposons cannot actively mobilize and only exist as relics of anciently active elements [5]. Actively

mobilizing retrotransposons include long terminal repeat retrotransposons (LTR) which are mostly endogenous retroviruses (ERVs), as well as non-LTR retrotransposons represented by long interspersed nuclear elements (LINEs) and their nonautonomous counterparts, short interspersed nuclear elements (SINEs) [5,15]. Furthermore, in nearly all examined mammalian genomes, LINEs are the most abundant TEs, and most of this observed abundance is represented by a single family, L1, which typically occupies hundreds of megabases of the genome [5]. This observed dominance of a single LINE is a trademark of mammalian genome structure [5]. However, the lack of genomes displaying alternative TE landscapes has limited our ability to investigate the evolutionary conflicts underlying mammalian genome structure evolution, as well as the evolutionary reasons for the maintenance of LINE dominance [5,16].

The North American deer mouse, *Peromyscus maniculatus,* is the most abundant mammal in North America and an important model for studying the genetic basis of adaptation [17]. Previous studies in the deer mouse and closely related species used polymerase chain reaction (PCR) methods to explore TE abundance and reported evidence for an unprecedented expansion of mysTR endogenous retroviruses (ERVs) [18,19]. However, the landscape of TEs in the deer mouse remains unexplored on a genomic scale.

Here, we report a highly distinct genomic landscape of TEs in the deer mouse genome. We find that, in contrast to nearly all examined mammalian genomes, LTR retrotransposons are more abundant in the deer mouse genome than LINEs. This landscape is driven by a combination of reduced lineage-specific LINE activity as well as a massive invasion of ERVs. Deer mouse ERVs display recent origins from diverse retroviruses and one ERV lineage likely arose from an

exogenous Feline Leukemia Virus within the past 16 million years. However, mysTR related ERV subfamilies account for a major proportion of lineage-specific ERV content with the top 5 most abundant subfamilies occupying a remarkable >2% of the genome. Several of the most abundant ERV subfamilies are nonautonomous and likely parasitize off the machinery of full-length ERVs for mobilization. While negative selection plays an important role in shaping the distribution of ERVs in the genome, we also find ERVs which display patterns consistent with regulatory function. Furthermore, we report a massive lineage-specific expansion of Kruppel-associated box domain-containing zinc finger genes (KZNFs) which reflects the relative time of ERV invasions. We find that young lineage-specific KZNFs colocalize with lineage-specific ERVs and hypothesize ERVs may play a role in facilitating the expansion of the KZNFs which suppress them. Moreover, we observed direct evidence for ERV-mediated LINE interruption. Together, these results illuminate the genomic ecology that shaped the deer mouse genome, opening up a range of opportunities to investigate the evolutionary processes that give rise to variation in mammalian genome structure as well as those that preserve its conservation across diverse species.

Results and Discussion

Deer mice exhibit a unique landscape of transposable elements

To evaluate the genomic landscape of TEs in the deer mouse genome, we first generated a lineage-specific TE library *de novo* from the *P. maniculatus* genome using a combination of systematic and manual methods (see Methods). We identified 48 LINE, 28 SINE, and 118 LTR deer-mouse specific subfamilies (Figure 1A, Supplementary Table 1). We then merged this lineage-specific TE library with all curated mammalian TEs from the Dfam database [20] and annotated the genome using the combined library. Our annotation revealed a distinct genomic

landscape of TEs in the deer mouse, relative to other mammals, in which LTR elements occupy more of the genome than LINEs (Figure 1A). Specifically, LTR elements occupy ~11% of the genome, followed by LINEs (~10%), SINEs (7%), and other TEs (<2%) (Figure 1A; Supplementary Table 2). In total, TEs occupy ~30% of the deer mouse genome, reflecting an increase in TE content relative to other species in the rodent Family *Cricetidae*, such as the grasshopper mouse (*Onychomys torridus*, 24%) and prairie vole (*Microtus ochrogaster*, 17%) (Figure 1A), but a considerable reduction relative to house mouse (*Mus musculus,* >40%), although these differences also likely capture differences in genome assembly and TE annotation quality [21,22]. Most of the difference in TE content between the deer mouse and house mouse can be attributed to decreased LINE content in the deer mouse, whereas most of the difference in TE content among cricetid species can be attributed to LTR elements.

Based on these observations, we hypothesized that the unique TE landscape of deer mice is the result of a combination of reduced lineage-specific LINE gain and a large proliferation of lineage-specific LTR elements. To investigate this possibility, we first compared genomic representation as a function of within-subfamily divergence across LINEs, SINEs, and LTR elements (Figure 1B-D). Consistent with our hypothesis, we observe reduced representation of LINEs with lower divergence from the consensus, suggesting decreased LINE accumulation in the deer mouse lineage on more recent timescales (Figure 1B). However, despite this decline in the accumulation of LINEs, we still find multiple candidate LINEs with intact protein machinery suggesting that LINEs are still active, consistent with previous reports of LINE activity in *Peromyscus* species (Supplementary Table 3) [23]. We also observe evidence for lineage-specific SINE activity (Figure 1C). Since SINEs parasitize LINE machinery for mobilization, the

observation of recently active SINEs suggests that potentially mobile LINEs still exist in the genome. In addition, we find a recent lineage-specific proliferation of LTR elements (Figure 1D): LTR elements are significantly overrepresented among the youngest TEs in the genome (<1% divergence from the consensus, two-sided Fisher's exact test P<0.00001). Furthermore, the observed decline of LINE gains in the genome coincides with the peak of LTR gains in the genome (Figure 1B-C). Together, these results suggest that both reduced LINE gain and lineage-specific LTR proliferation have contributed to the deer mouse's unique TE landscape, and that the two might be associated.

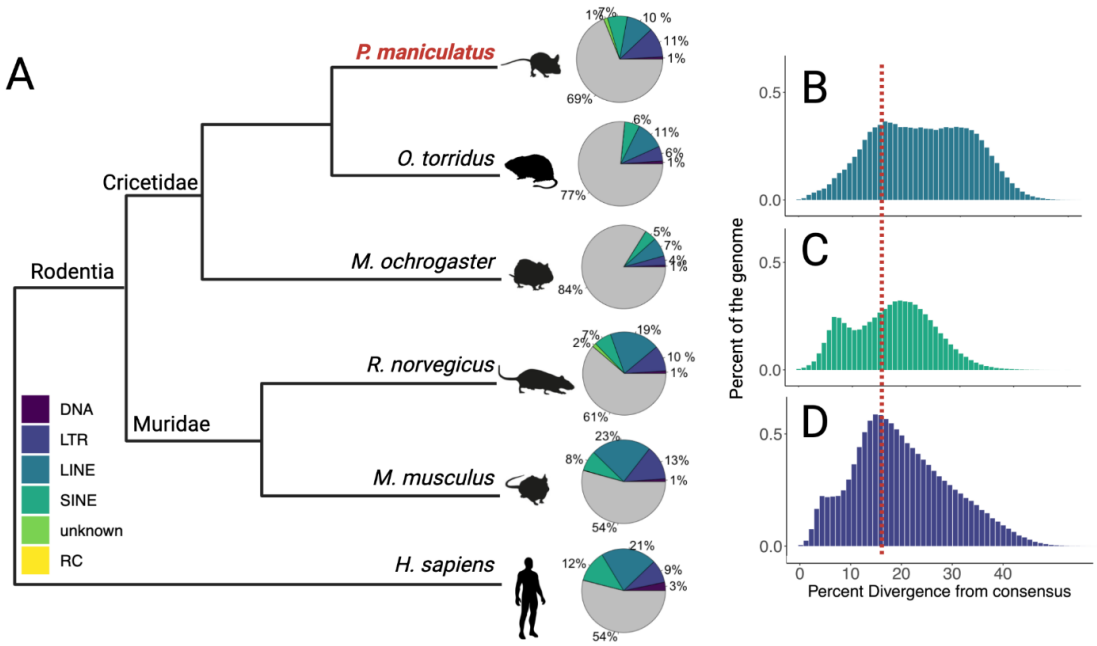

Figure 1: Transposable element landscape: (**A**) Cladogram highlighting the relationship of *P. maniculatus* (red) to other mammalian species,. Pie charts show the relative percentage of TE subfamilies for each species. Color corresponds to the percent of the genome attributed to each type of TE (see legend); Gray represents the percent of the genome that is not occupied by TEs. Note, in *P. maniculatus* LTRs (dark blue) occupy more of the genome than LINEs (aqua blue). (**B-D**) Percent of the genome as a function of CpG corrected Kimura divergence from the consensus for each TE subfamily of (**B**) LINEs, (**C**) SINEs, and (**D**) LTR elements. The red dotted line represents the start of observed LINE decline in all plots.

DNA loss fails to explain reduced LINE content in the deer mouse genome

In addition to gain, TE loss is an important driver of genomic TE content. Although we observe evidence for a decline of both LINE activity and fixation, the observed low LINE content in the deer mouse genome, relative to *M. musculus,* could also have resulted from higher rates of loss in the deer mouse (Supplementary Figure 1A-B). To investigate this possibility, we calculated the DNA loss coefficient k (following [24]), using the formula $E = A\ e\text{-}kt$, where E is the amount of extant ancestral DNA in the species considered, A is the ancestral assembly size, and t is time. Larger values of k suggest higher rates of lineage-specific DNA loss. We calculated a k coefficient of ~0.0047 for deer mice, a value similar to, and in fact slightly lower than the k value estimated for *M. musculus* (~0.006), suggesting that the reduced LINE content observed in the deer mouse genome cannot be explained by significantly higher rates of loss in deer mice (Supplementary Figure 1C).

Nonetheless, it remains possible that the rate of LINE loss in the deer mouse does not reflect the genome-wide rate of loss; rather, LINEs are lost at a higher rate. To investigate this possibility, we compared the proportions of DNA attributed to ancient mammalian LINEs present in the common ancestor of the deer mouse and house mouse as well as lineage-specific elements. If the relative absence of LINEs in the deer mouse is due to high rates of loss, we expect to observe a decreased amount of DNA attributed to ancient LINEs in the deer mouse relative to house mouse. To the contrary, we find that although LINEs contribute to nearly twice as much of the house mouse genome as the deer mouse genome, ancient LINEs are significantly less represented in the house mouse genome (Supplementary Figure 1D, two-way Fisher's Exact Test, P=0.006). These results suggest that the low LINE content observed in the deer mouse

cannot be explained by high rates of LINE-specific loss and instead, is more likely the result of reduced rates of gain.

Nonautonomous ERVK-like elements are abundant in the deer mouse genome

Most mammalian LTR retrotransposons are endogenous retroviruses (ERVs). ERVs are divided into three broad classes depending on their retroviral origins: ERV1, ERVK and ERVL with the additional subgroup of nonautonomous ERVL-MaLRs [20,25]. In the deer mouse, we observe a pattern in which ERVK and ERVL-MaLR elements together account for over 80% of genomic ERV content, consistent with previous observations in other rodent species [5,20] (Figure 2A). Lineage-specific elements represent over half of genomic ERV content, further suggesting that the deer mouse has experienced a lineage-specific ERV expansion. However, when we compare the relative genomic proportion of ERVs across lineage-specific subfamilies, we find that ERVKs account for a disproportionately large proportion of ERV content relative to shared elements (Fisher's exact test, $P<0.00001$), representing over 75% of observed lineage-specific ERV sequence in the genome (Figure 2B, Supplementary Table 1).

ERVKs are usually autonomous elements that encode their own machinery for mobilization [26]. However, we carefully inspected deer mouse-specific ERV subfamilies, annotating *gag*, *pro*, *pol* and *env* genes as well as the protein domains required for autonomous transposition and found that several of the most abundant ERVK families did not possess any internal open reading frames (ORFs) or expected protein domains, suggesting that they might be nonautonomous. Many ERVs contained gaps which interrupted or truncated their internal sequences, making it challenging to reconstruct full length elements and assess the presence or absence of coding machinery. In light of this, we required that a putatively nonautonomous ERV subfamily display

at least 5 full length copies with no gaps for it to be classified as nonautonomous, regardless of its consensus sequence length or content. Even with this conservative filter, we found that the most abundant deer mouse-specific ERV subfamilies are nonautonomous ERVK-like elements (Figure 2C; Supplementary Table 4). Furthermore, for the subset of ERVKs in which we could confidently reconstruct full length sequences and assess autonomy, nonautonomous elements occupy significantly more of the genome than autonomous ones (Fisher's exact test $p<0.0001$; Supplementary Table 4). Overall, our results suggest that ERVKs and their nonautonomous counterparts have had a significant impact on the deer mouse's unique genome structure.

Several autonomous and nonautonomous ERVKs show sequence homology to mysTR

Nonautonomous TEs parasitize autonomous elements for mobilization. Studies on the nonautonomous ERVK, ETn, in *Mus* showed that ETn exhibits regions of homology to fully coding MusD ERVKs, suggesting that ETn likely arose from the ancestors of Mus-D [27] and now hijacks MusD machinery for mobilization via complementation in trans [28]. Given the massive copy numbers of nonautonomous ERVKs in the deer mouse genome, we next searched for related autonomous elements which might facilitate nonautonomous element mobilization. In addition to several prolific nonautonomous subfamilies, we observe three autonomous ERVK subfamilies, Pman_ERVK_4.503, Pman_ERVK_6.7639 and Pman_ERVK_5.247, that together occupy a remarkable 1% of the genome (Figure 2C). These subfamilies show sequence homology to mysTR, a previously identified ERV [18].

Previous studies failed to identify full-length mysTR copies with intact *gag* and *pro-pol* genes required for mobilization, raising questions about mysTR's overall autonomy and current ability

to mobilize, although these studies lacked the genomic resources to analyze mysTR sequences on a genome-wide scale [18,29]. Our genome-wide analysis reveals multiple candidate intact copies of mysTR related ERVs which display ORFs with homology to *gag* and *pro-pol* genes and contain expected protein domains, suggesting that these ERVs are indeed autonomous and still active in the deer mouse (Supplementary Table 4). Furthermore, we hypothesized that observed nonautonomous ERVKs represent mys elements that are also related to mysTR and may now use mysTR protein machinery for mobilization [19,30]. Indeed, similarly to ETn and MusD, we find regions of homology between autonomous mysTR related ERVK subfamilies and nonautonomous ERVK subfamilies (Figure 2D; Supplementary Table 4). The most conserved region of nucleotide sequence homology between these subfamilies is just downstream of the *pro-pol* gene and upstream of the 3' LTR (Figure 2D). Interestingly, candidate nonautonomous and autonomous mysTR-related subfamilies do not display strong homology outside of this region, suggesting that ERVKLs may have evolved through an internal recombination event in an autonomous ERVK [27]. Maintenance of sequence similarity in this region is also consistent with strong selection due to a possible role in ERVK mobilization, although its function remains unknown.

Autonomous ERV subfamilies are younger than autonomous ones

Autonomous ERV subfamilies are generally younger than nonautonomous ones. Since ERV LTRs are identical upon insertion, LTR sequence identity can provide an estimate for how recently an ERV inserted. To investigate the evolutionary dynamics of deer mouse ERVs, we compared the distributions of LTR identity across lineage-specific ERV subfamilies. We find that nonautonomous ERVs are significantly older than autonomous ERVs (Mann Whitney U, P=0.03; Figure 2E). This pattern is likely explained by multiple non-mutually exclusive processes. First,

nonautonomous ERV insertions probably rise to fixation more often than fully autonomous insertions since they are less deleterious to their host. This is expected since nonautonomous ERVs are shorter than their autonomous counterparts and lack coding transcripts which can be toxic to the host. Second, since the coding capacity of autonomous ERVs dictates the ability of both autonomous and nonautonomous ERVs to mobilize, the activity of parasitic nonautonomous subfamilies is likely associated and limited by that of their autonomous targets [31]. Consistent with this, we also observe that the youngest ERV subfamilies in the deer mouse genome are autonomous (Figure 2F).

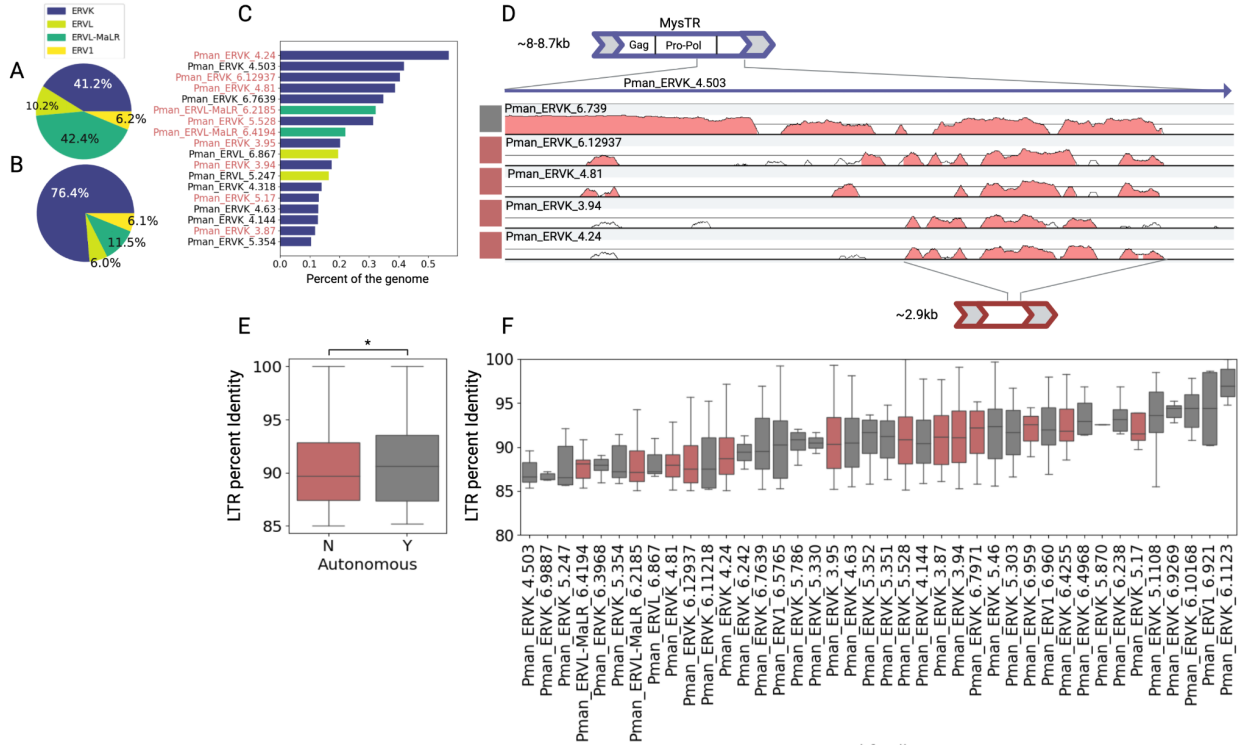

Figure 2: Relative contribution of different broad ERV classes to ERV content in the deer mouse across (**A**) all ERVs and (**B**) lineage-specific ERVs. (**C**) Respective genomic occupancy across the 18 most common lineage-specific ERV subfamilies. Red font denotes nonautonomous subfamilies. (**D**) VISTA plot showing regions of homology between autonomous mysTR subfamilies and nonautonomous subfamilies as well as their respective position. Gray boxes on the left denote autonomous subfamilies and red boxes denote nonautonomous ones. (**E**)

Comparison LTR percent identity aggregated across all lineage-specific autonomous and nonautonomous elements suggests that autonomous elements are significantly younger (with * denoting Man Whitney U $p<0.05$). (**F**) Comparison of LTR percent identity across all ERV subfamilies which display at least one candidate full-length copy suggests that the youngest subfamilies are autonomous.

Diverse origins of endogenous retroviruses in the deer mouse

ERVs in the deer mouse genome show diverse origins. ERVs arise in a species when an exogenous retrovirus infects the germline, and new families of ERVs evolve *de novo* more frequently than other autonomous mammalian TEs such as LINEs [26]. To investigate the retroviral origins of ERVs in the deer mouse, we focused on full-length ERVs across all identified subfamilies with flanking LTRs, *pol* genes and full-length reverse transcriptase (RT) domains (>450bp) which we used for classification and phylogenetic analysis. We initially identified 148 candidate full-length ERVs with *pol* genes and evidence of an RT domain. However, many ERVs contained ambiguous sites or gaps which interrupted or truncated the RT domain, leaving only 52 ERVs which met our conservative requirements (Supplementary Table 5-6). Thus, we note that our reported ERV diversity is likely an underestimate.

We initially used a hidden Markov model (HMM) [32] approach to classify ERVs based on their RT domains (Supplementary Table 6). This approach suggests that of the 52 deer mouse ERVs with full-length RT domains, 11 are derived from gammaretroviruses (ERV1), 39 from betaretroviruses (ERVK) and 2 from spumaretroviruses (ERVL) (Supplementary Table 6). Phylogenetic analysis of RT domains from these ERVs and other known retroviruses supports these initial classifications and shows that deer mouse ERVs form 14 distinct clusters representing at least 14 independent origins spanning retroviral diversity (Figure 3A). Most of these are derived from diverse betaretroviruses (9 of the 14), consistent with previous

observations in other rodents [33,34]. Additionally, four ERV clusters show evidence of gammaretroviral origin and one ERV cluster shows evidence of spumaretroviral origin (Figure 3A).

Searches for deer mouse ERVs in grasshopper mouse, prairie vole, and house mouse suggest that most (9 of 14) deer mouse ERVs arose before the divergence between the deer mouse and grasshopper mouse (~5-11 MYA) [35,36], but after the divergence between their ancestral lineage and the prairie vole (~18 MYA) [37,38], and are thus lineage-specific relative to *M. musculus* (Figure 3B). Additionally, one ERV (Beta_Pman-ERV_cluster-5) evolved even more recently, after the divergence between the deer mouse and grasshopper mouse. Moreover, LTR identity for ERVs in each respective cluster generally concur with these results (Supplementary Table 6). Given the recent origins of deer mouse ERVs and since ERVs frequently arise in new species through the invasion and subsequent endogenization of exogenous retroviruses, we searched for lineage-specific ERVs that display strong homology to a known exogenous retrovirus. We identified one potential case of a recent endogenization of an exogenous Feline Leukemia Virus (FeLV) or a closely related virus in the ancestor of the deer mouse and grasshopper mouse within the last 18 million years (Figure 3C) [37,38].

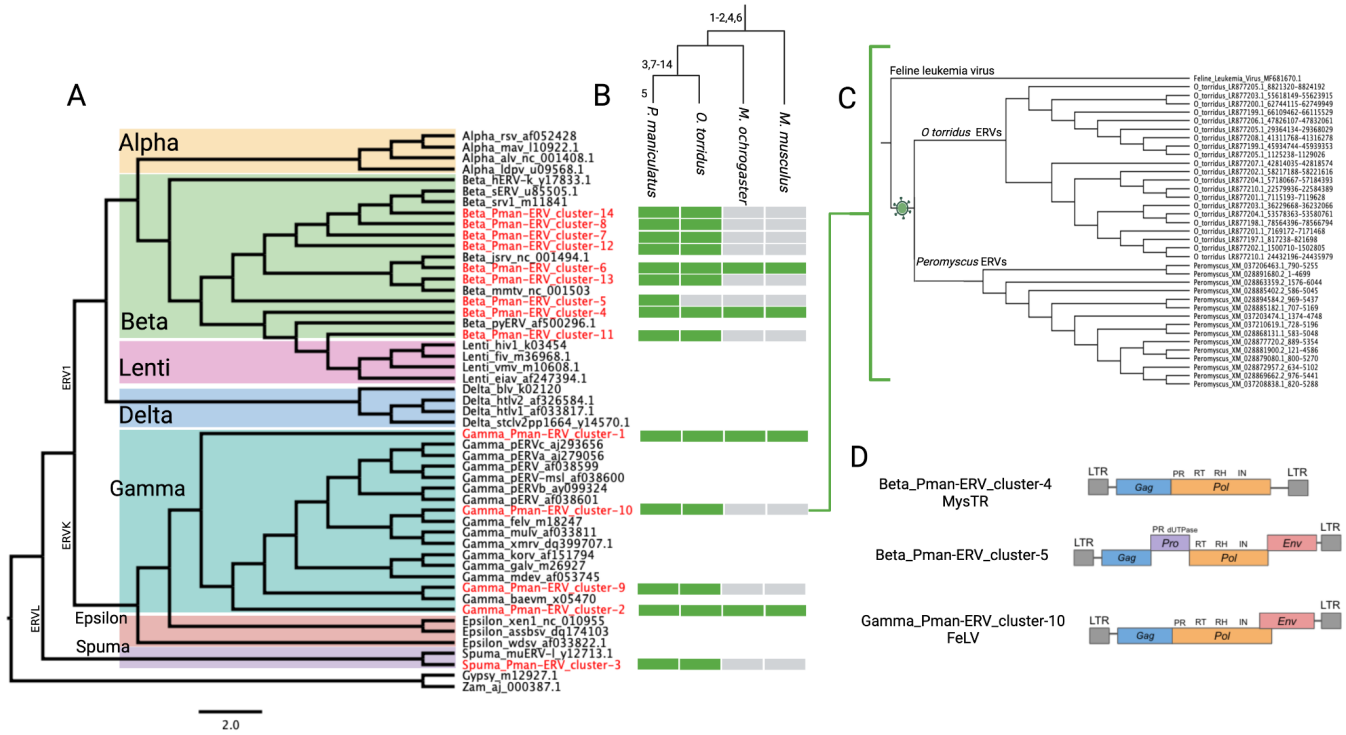

Figure 3: Origins of ERVs in the deer mouse genome: (**A**) Phylogeny constructed with *RT* domains of deer mouse ERVs and other publicly available endogenous and exogenous retroviruses for context. Deer mouse ERVs form 14 distinct clusters spanning broad retroviral diversity. (**B**) Cladogram showing the approximate time of origin of *P. maniculatus* ERVs based on presence or absence in species at different phylogenetic distances. ERV cluster numbers are annotated on the branch corresponding to their approximate origin. For each species, green boxes represent presence and gray boxes represent absence for each respective ERV cluster found in *P. maniculatus*. (**C**) Neighbor-joining tree showing the phylogenetic relationship between Pman-ERV_cluster-10 copies in *P. maniculatus,* related ERVs in *O. torridus*, and feline leukemia virus, suggesting that Pman-ERV_cluster-10 originated via cross-species infection of an exogenous feline leukemia virus before the divergence between *P. maniculatus* and *O. torridus* but after the divergence between their ancestral lineage and *M. ochrogaster*. (**D**) Structure of ERVs found in *P. maniculatus* with ORFs denoted by colored boxes and important protein domains annotated. Overlapping boxes denote overlapping ORFs.

Some deer mouse ERVs may still be infectious

Although ERVs only require *gag* and *pol* genes to mobilize in the germline, ERVs with intact *env* genes can also infect other cells. Given the recent evolution of ERVs in the deer mouse, we inspected all intact ERVs as well as ERV subfamily consensus sequences for intact *env* genes.

We find no evidence for *env* genes in mysTR related subfamilies, consistent with previous studies on mysTR [18] (Figure 3D; Supplementary Table 4). However, we find putatively intact *env* genes in multiple other ERV clusters suggesting that some deer mouse ERVs are still capable of horizontal transmission (Supplementary Table 4, 6). One of these is Gamma_Pman-ERV_cluster-10, suggesting that Gamma_Pman-ERV_cluster-10 ERVs are still infections, as has been previously observed for other Leukemia Viruses [39,40] (Figure 3D). We also observe evidence of an intact *env* gene for ERVs within the Beta_Pman-ERV_cluster-5. Beta_Pman-ERV_cluster-5 ERVs are absent in *O. torridus* and thus represent some of the most recent ERVs to arise in the deer mouse (Figure 3B, D). Beta_Pman-ERV_cluster-5 *env* genes show sequence homology to the *env* IAP elements in *M. musculus*, which are also capable of intracellular transmission [41], suggesting possible origin from a similar retrovirus (Supplementary Table 4, 6).

Negative selection plays an important role in shaping TE distributions in the deer mouse genome

Most TE insertions are deleterious or neutral and the distribution of TEs in a genome is primarily shaped by negative selection. In the deer mouse genome, TEs are generally well represented in coding genes, accounting for nearly 25% of genic nucleotides, but are relatively absent from coding exons, suggesting strong purifying selection on new insertions in coding exons (Figure 4A). We also observe considerable representation of TEs in long noncoding RNAs (lncRNAs), consistent with observations in other species [42]. Comparison of TE occupancies across chromosomes reveals that ERVs and LINEs exhibit higher occupancies on the X chromosome (Figure 4B). This pattern is not observed for SINEs and likely reflects the more frequent removal

of longer TEs such as LINEs and ERVs on autosomes by recombination [43,44]. ERV insertions around protein coding genes are also usually deleterious since ERVs contain complex internal regulatory elements which can disrupt gene expression. Consistent with this, ERVs are generally distant from genes, and even more distant from genes in the same orientation (Mann Whitney U Test, $P<0.0001$; Figure 4C). It is worth noting that this bias is most pronounced for mysTR related ERV subfamilies, suggesting that these ERVs are highly deleterious.

Some ERV subfamilies display patterns consistent with possible regulatory function

While most ERV subfamilies show patterns suggesting deleterious effects on neighboring genes, others display patterns consistent with possible regulatory function. ERV LTRs are often co-opted for important regulatory functions over evolutionary time [45]. We find some ERV subfamilies that are enriched within the 5-kb region upstream of genes, suggesting selection on these ERVs to minimally affect neighboring gene expression or the possibility of host co-option (Figure 4D). These ERVs also display significantly higher within-subfamily divergence relative to other lineage-specific deer-mouse ERVs (Mann Whitney U Test, $P=0.0048$), suggesting that they primarily represent inactive subfamilies (Figure 4E). Additionally, some subfamilies, including MT2B1 and ORR1B1, represent lineages of ancestrally shared elements which have been co-opted for regulatory functions in other mammalian species [46]. Functional enrichment tests on neighboring genes for each of these subfamilies yield few enriched categories. Regardless, given the frequent and recurrent lineage-specific ERV co-option events observed across mammals [47–49], these ERVs represent promising candidates for independent co-option events in the deer mouse.

ERVs accumulate in "hotspots" which are enriched in Kruppel-associated box domain-containing zinc finger genes

The distribution of ERVs in the genome is largely biased towards specific regions, or "hotspots", which are enriched in Kruppel-associated box domain-containing zinc finger genes (KZNFs). We define "hotspots" as regions of the genome in the top 95th percentile of ERV density, where ERV density is the proportion of nucleotides attributed to ERVs in a given 100-kb genomic window (Figure 4F). Lineage-specific ERVKs constitute over 70% of ERVs in hotspots, suggesting that these genomic structures are likely also lineage specific. ERV hotspots are largely devoid of genes, and we observe a strong negative correlation between gene density and ERV density overall (GLM, $P<0.0001$). However, we do observe some genes in ERV hotspots. We performed a gene ontology (GO) enrichment analysis for genes in ERV hotspots and found significant enrichment for one biological process term: "regulation of transcription, DNA-templated" (Fisher's Exact Test, $Q<0.00001$). Scrutiny of genes overlapping ERV hotspots which match this GO term reveals that ~85% (100/118) are deer mouse-specific Kruppel-associated box domain containing zinc finger genes (KZNFs) (Figure 4C). We find that deer mouse specific KZNFs specifically are enriched in ERV hotspots, with ~32% (100/312) of KZNFs occurring in ERV hotspots, despite the fact that ERV hotspots only represent <5% of the genome (Fisher's Exact Test, $P<0.00001$).

Coevolution of ERVs and KZNFs in the deer mouse genome

The primary function of KZNFs is to suppress ERV activity [11,50,51]. KZNF gene clusters evolve rapidly through a birth-death model under positive selection and often expand in response to lineage-specific ERV activity [52–54]. The colocalization of KZNF genes and ERVs in

genomic space is curious and has been observed previously in the house mouse [55]. Although this observation could simply be explained by relaxed selection on nonessential KZNF genes, an alternative hypothesis is that ERVs contribute to KZNF gene family evolution by facilitating rapid gene turnover in these regions. Indeed, ERVs facilitate structural rearrangements via ectopic recombination and ERV-rich regions of the genome are often highly plastic. Interestingly, lineage-specific KZNF duplicates in ERV hotspots exhibit significantly lower divergence compared to other KZNFs, suggesting that genes in ERV hotspots duplicated more recently (Figure 4G, Mann Whitney U, P=0.0043). This observation supports a hypothesis in which KZNFs in ERV hotspots duplicate more often, although the evolutionary processes driving this pattern remain unclear.

The relative timing and magnitude of lineage-specific KZNF gene family expansion in ERV hotspots reflects that of lineage-specific ERVK activity. We compared the distribution of candidate gene duplicate divergence for KZNFs in ERV hotspots and KZNFs outside ERV hotspots with the distribution of within-subfamily ERVK divergence (Figure 4H). The distribution of duplicate divergence for KZNFs in ERV hotspots suggests that the largest KZNF expansion occurred just before or around the same time of peak ERVK activity (Figure 4H). Indeed, the median percent divergence for lineage-specific KZNF gene duplicates in ERV hotspots is ~17.2% while the within-subfamily divergence for the top 3 most abundant ERVs in the *P. maniculatus* genome is ~17.4%. This pattern is consistent with a KZNF expansion in ERV hotspots in response to highly active lineage-specific ERVKs. In contrast, the distribution of duplicate divergence for KZNFs not overlapping ERV hotspots shows little evidence for a relationship to lineage-specific ERVK activity (Figure 4H), further suggesting that the observed

colocalization between ERVKs and KZNFs may have some evolutionary significance. Furthermore, some KZNF gene clusters display much larger expansions than others. A cluster on chromosome one contains >90 genes, representing ~1/3 of annotated lineage-specific KZNFs in the genome (Figure 4F). This observation suggests that KZNFs in this chromosome 1 cluster in particular might play an important role in suppressing ERVs in the deer mouse. We observe another case on chromosome 22, which displays a custer of 48 lineage-specific genes. Since specific KZNF clusters often bind to specific ERV families, the massive invasion of closely related ERVs in the deer mouse predicts expansions of closely related KZNFs [52]. Together, these results suggest that KZNFs in the deer mouse underwent a massive expansion in response to lineage-specific ERV activity.

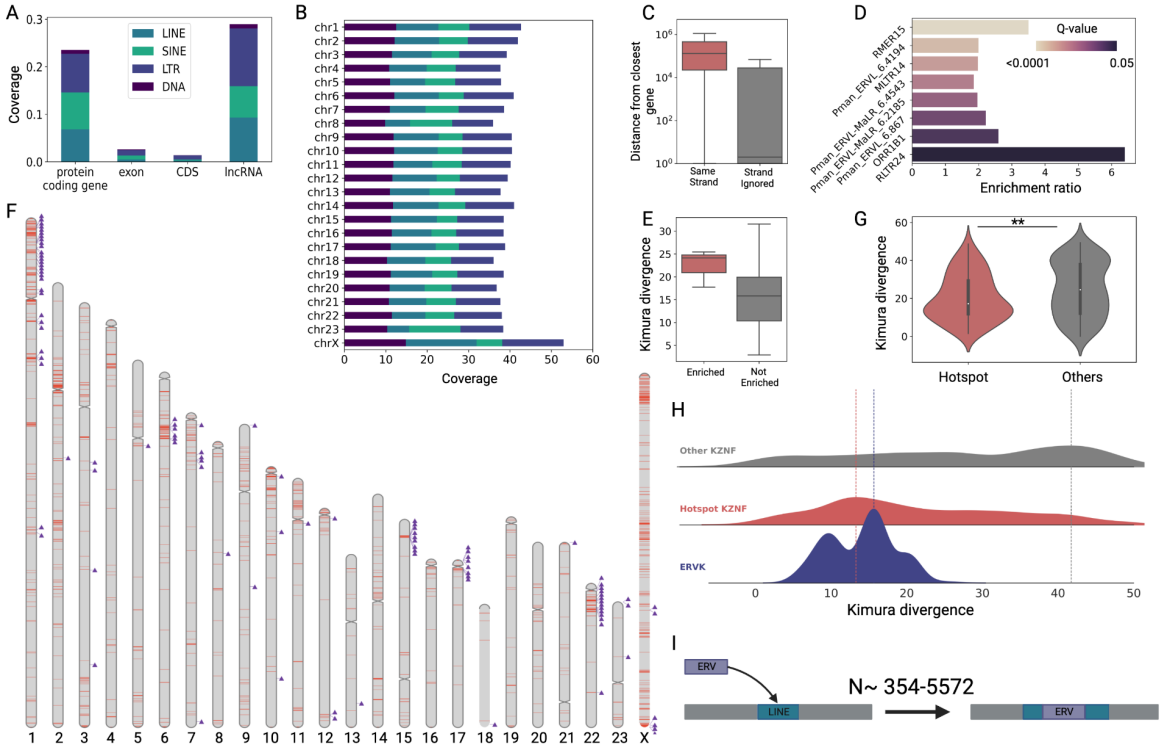

Figure 4: Genomic distribution of TEs in the deer mouse: (**A**) Respective coverage for different TE subclasses across genomic features. Coverage is defined as the proportion of nucleotides attributed to TEs for a given feature. (**B**) Respective coverage for TEs across chromosomes. (**C**) Box plots showing the distribution of ERV distances from the closest gene on the same strand versus when strand is ignored. (**D**) Enrichment ratio (number observed/expected) and bonferroni corrected Fisher's Exact Test P-values (Q-values) for ERV subfamilies enriched within the 5 kb upstream of genes in the same orientation. (**E**) Within-subfamily CpG corrected Kimura divergence for ERV subfamilies enriched within the 5 kb upstream of genes in the same orientation compared to all other ERV subfamilies. (**F**) Genomic distribution of ERV hotspots (red) across chromosomes. Lineage-specific KZNF genes are annotated with purple triangles and are enriched in ERV hotspots. **(G)** KZNFs in ERV hotspots show lower Kimura divergence than other KZNFs suggesting that they are younger. **(H)** Kernel density estimates for the distribution of Kimura divergences for KZNFs outside ERV hotspots (gray), KZNFs in ERV hotspots (red), and ERVKs (purple). Color coded dotted lines show the peak for each distribution. **(I)** Cartoon displaying an ERV insertion interrupting a formerly intact LINE. *N* represents the range of observed candidate instances of ERV-mediated LINE interruption.

Lineage-specific ERV insertions interrupt LINE sequences

In addition to evaluating ERV distributions with respect to genes, we also assessed ERV distributions with respect to other TEs. We were specifically interested in how the observed ERV invasion in the deer mouse might directly impact preexisting LINEs. ERVs could have directly impacted LINE activity in the deer mouse lineage by inserting into and interrupting transposition-competent LINEs. LINE families typically only have from a hundred to a few thousand "master genes" which are transposition-competent in mammalian genomes [5,56–58]. Furthermore, the LINE retrotransposition mechanism is fairly inefficient, and the vast majority of new LINE insertions are defective and incapable of mobilizing thereafter [59,60]. Thus, disruption of many master genes could have a considerable impact on the evolutionary trajectory of LINEs in a species.

To explore the direct impacts of ERV insertions on LINEs, we searched for ERV insertions directly flanked by LINE sequences from the same LINE subfamily. We then filtered for cases in which flanking LINE sequences conjoined at the correct coordinates with respect to the subfamily consensus, forming a full-length LINE. We also initially filtered for LINEs which did not contain any additional TE insertions. These results revealed 354 prospective lineage-specific ERV insertions which interrupt full-length LINEs. However, this number is likely a considerable underestimate, since it does not include fragmented LINEs that have accumulated multiple indels. If we include those as well, we find 5572 prospective ERV insertions interrupting LINEs, 900 of which are attributed to the two most abundant ERVK related subfamilies (Pman_ERV2_4.503 and Pman_ERV2_4.24). Interestingly, within-subfamily percent divergences for these subfamilies (16.05% and 15.86%) suggests that they invaded just before the decline of LINE gain (Figure 1B, ~15%). We postulate that this association is no

coincidence. Punctuated LINE interruptions on these scales (354 - >5000 LINE interruptions) would eliminate most functional LINEs in many mammalian species, and even on much smaller scales, could have a catastrophic effect on the evolutionary trajectory of LINEs in a genome, especially given the poor success rates of the LINE retrotransposition mechanism in producing new fully functional LINE copies.

An "ecology of the genome" model for the evolution of the deer mouse genome

In the same way that species compete for space and resources, TEs compete with each other for sites in the genome as well as metabolic resources [9]. TEs can occupy specific niches which can allow them to coexist with limited competition, but TEs which occupy similar niches are more likely to drive one or another to extinction [9,10]. Furthermore, the relative success of a given TE also depends on host suppression mechanisms and their targets. For example, differential host targeting between two TE families in direct competition could limit the success of one family that would, in the absence of host defense mechanisms, be more fit than the other [10]. Also, because TEs could threaten to kill their host in the absence of host-mediated suppression, it is advantageous in some cases (for both the host and TEs) that host defenses evolve to suppress TE activity [10].

Our model for the evolution of the unique TE landscape observed in the deer mouse genome lends itself to the ideas explained above: the "ecology of the genome" [9]. We postulate that the introduction of mysTR related ERVs sparked a shift in the deer mouse TE landscape through the following processes (Figure 5): First, mysTR ERVs evaded host defenses upon immigration, which allowed them to expand to large copy numbers. This hypothesis is supported by the

observation that mysTR ERVs are highly divergent from other known retroviruses as well as the remarkable expansion of deer mouse-specific KZNFs following peak ERV activity [18]. In mammals, ERVs are the primary targets of KZNF suppression, whereas LINEs and SINEs are less frequently targeted, probably because ERV insertions are more deleterious [61,62]. These host defenses keep ERVs in check, despite evidence that LINEs and ERVs occupy overlapping niches. In fact, several lines of evidence suggest that LINEs and ERVs are in direct competition. First, many ERVs and LINEs both preferentially integrate into AT rich regions [63–66]. Thus, ERVs and LINEs often inhabit similar regions of the genome and insert within each other frequently [63]. Second, ERV insertions in LINEs (or vice versa) are likely invisible to selection and exhibit a higher rate of fixation relative to deleterious insertions [63]. Under these circumstances, in the absence of host defense mechanisms, we expect the primary driver of ERV or LINE success in the genome to be relative rates of gain (of transposition-competent copies). Thus, we postulate that the massive expansion of mysTR ERVs nearly drove LINEs to extinction in the deer mouse genome.

More generally, we propose that this model may explain the loss of LINE activity in other mammals. A subclade of sigmodontine rodents for example (~13-18 MYA diverged from the deer mouse [36–38,67]), represents one of the few mammalian lineages to have experienced LINE extinction [68]. Consistent with our model, previous studies suggest that LINE extinction in this lineage followed an invasion of mysTR related ERVs on a similar or possibly larger scale to that observed for the deer mouse [18]. The lack of genome assemblies for sigmodontine rodents makes it challenging to study TEs in these species for the time being. Future studies in these and other species that show contrasting patterns to the norm for mammalian genome

composition will shed further light on evolutionary conflicts that drive mammalian genome evolution.

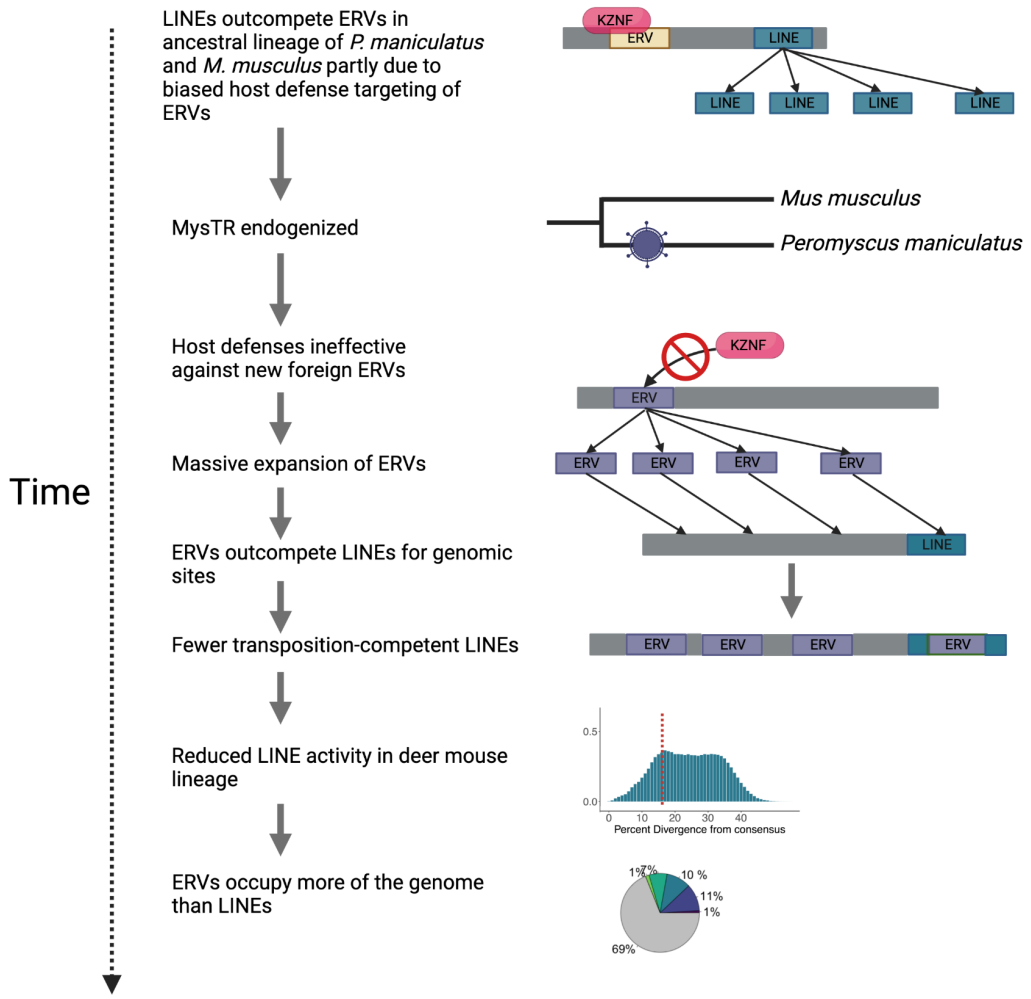

Figure 5: Model for deer mouse genome evolution.

Concluding remarks

Although TE landscapes differ drastically across species, most mammalian genomes are dominated by LINE retrotransposons. The deer mouse, *Peromyscus maniculatus*, represents one of the only exceptions to this pattern in which LTR elements occupy more of the genome than LINEs. We find that the unique genomic landscape of TEs in the deer mouse reflects a massive

expansion of ERVs as well as a dearth of LINE activity on recent timescales, and that the two are likely associated. Our results demonstrate a broad diversity of ERVs in the deer mouse genome and that the immigration of one particular divergent ERV, mysTR, played an integral role in driving the deer mouse unique TE landscape. Based on these findings as well as expectations from genomic ecological models, we postulate that the propensity for a mammalian genome to undergo a shift in TE content and/or experience LINE extinction is directly related to its susceptibility to invasion by divergent TEs (in this case ERVs). Furthermore, the accumulation of ERVs in specific hotspots raises additional questions about how TE-dense regions can affect mammalian genome evolution. Indeed, we would expect such regions to experience structural rearrangements more often than other regions of the genome. Previous studies in the deer mouse have identified massive inversions (>1Mb in length) which are polymorphic within populations [69,70]. Could these ERV hotspots have played a role in facilitating deer mouse inversions? We also observe enrichment of KZNF gene families which evolve rapidly via duplication under positive selection in ERV hotspots. Is the colocalization of KZNFs and ERVs advantageous for the host due to the increased propensity for KZNF gene family expansion? Or is this observation a mere coincidence? We show that KZNFs in ERV hotspots are indeed younger than other KZNFs, providing some support for the prior. However, within-population studies are essential to further elucidate this coevolutionary relationship. Together, our results have broad implications and open up a range of opportunities to investigate the evolutionary processes that give rise to variation in mammalian genome structure.

Materials and Methods

Obtaining relevant genomic data

We downloaded publically available TE annotations for *Homo sapiens* (GCF_000001405.40; genome contig N50 = 57,879,411; contig L50 = 18), *Mus musculus* (GCF_000001635.27; genome contig N50 = 59,462,871; contig L50 = 15), *Rattus norvegicus* (GCF_015227675.2; genome contig N50 = 29,198,295; contig L50 = 27), and *Microtus Ochrogaster* (GCF_000317375.1; genome contig N50 = 21,250; contig L50 = 29,205) from RepeatMaster (http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html) and for *Onychomys torridus* from NCBI (GCF_903995425.1; genome contig N50 = 2,276,141; contig L50 = 308) respectively. We used the *Peromyscus maniculatus* genome assembly available through NCBI (refseq GCF_003704035.1; contig N50 = 30,111; contig L50 = 23,323) for all genomic analyses. Retroviral sequences for ERV phylogenetic analysis were downloaded from NCBI. Genbank accession numbers and for these sequences are shown in Figure 1A.

TE discovery and annotation

We used a combination of systematic and manual techniques to identify and annotate TEs in the *P. maniculatus* genome. We started by using an approach similar to the EarlGrey pipeline (github.com/TobyBaril/EarlGrey/) [71]. We first identified known rodent TEs in the *P. maniculatus* genome using RepeatMasker (version 4.1.2) (https://www.repeatmasker.org/) with a curated set of rodent TEs from the DFAM database [20] and the flags *−nolow*, *-norna* and *−s*. Next, we constructed a *de novo* repeat library using RepeatModeler2 (version 2.0.1), with RECON (version 1.08) and RepeatScout (version 1.0.5) [72–74]. Maximum-length consensus sequences were generated for putative *de novo* TEs identified by RepeatModeler using an automated version of the "BLAST, Extract, Extend" process through EarlGray [21]. Briefly, EarlGray first performs a BLASTn search to obtain up to the top hits for each TE family [75]. Then, it aligns the 1000 base pairs of flanking retrieved sequences using MAFFT (version 7.453) [76]. Following this, alignments are trimmed using trimAl (version 1.4) with the options (*−gt* 0.6 *-cons* 60) [77]. Finally, consensus sequences are updated using EMBOSS cons (*−plurality* 3) [78]. This process is then repeated 5 times. Following this, we performed blastx [75] searches against all known *P. maniculatus* proteins with parameters (*-max_target_seqs* 25 *-culling_limit* 2 *-evalue* 10e-10) and filtered all TEs with unknown classifications that shared homology with proteins.

Following the automated processes described above, alignments for TE families were individually inspected using AliView [79] and poorly represented positions were manually trimmed as recommended by [80]. Families were also manually realigned using extract_align.py [21] and MAFFT (version 7.453) [76] and then reexamined. Manually curated TE families were then re-clustered using cd-hit-est *[81]* and families were merged based on the 80-80-80 rule criterion [82]. We also used TE-Aid (https://github.com/clemgoub/TE-Aid) to identify TE-associated ORFs and sequence features such as long terminal repeats when classifying TEs. We combined our final *de novo* TE library with the Rodent DFAM TE library [20] and annotated

TEs in the *P. maniculatus* genome using RepeatMasker (version 4.1.2) (https://www.repeatmasker.org/). To identify full-length LTR elements, we used LTR_FINDER [83] and LTRharvest [84] through EDTA_raw with the flag *-type ltr* (version 2.0.0) [85], which also report LTR divergence for each element. TE annotations were defragmented and refined using RepeatCraft with the flag *−loose* [86], and overlapping annotations were resolved in favor of the longer element using MGKit (version 0.4.1) filter-gff [87].

Identifying functional machinery for putatively autonomous TEs

To identify the protein machinery of potentially autonomous LINE and LTR elements, we extracted all LINE elements longer than 2700bp and LTR elements longer than 5000bp from the *P. maniculatus* genome. Then, we also used TE-Aid (https://github.com/clemgoub/TE-Aid) to identify ORFs in each retrieved LTR and LINE element with homology to known TE genes. We used hmmer [32] and relevant hmms available from GyDB [88] and PFAM [89] to identify retroviral protein domains as well as NCBI's conserved domain search tool [90,91].

Calculating *k* coefficients

We calculated the DNA loss coefficient k [24], using the formula $E = A\ e\text{-}kt$, where E is the amount of extant ancestral DNA in the species considered, A is the ancestral assembly size, and t is time. We calculated E for each species by subtracting the amount of genomic DNA attributed lineage-specific TEs from the amount of DNA attributed to ancient shared mammalian TEs (retrieved from [92]). We used 2.8Gb for A and 100 million years for t as in [92].

Identifying nonautonomous ERV-like elements

Since the *P. maniculatus* genome was produced primarily with short reads, most ERVs have internal gaps or strings of low quality or ambiguous nucleotides. Thus, in order to decipher nonautonomous ERV-like elements from autonomous ERVs, we used a strict criterion. For a given ERV subfamily to be considered nonautonomous, we required at least 5 full-length copies which lack identifiable ORFs as well as ambiguous nucleotides.

Comparing nonautonomous and autonomous ERVK sequences

We performed global pairwise alignments between nonautonomous and autonomous ERVK consensus sequences using AVID with default parameters [93]. We visualized alignments using VISTA [94].

ERV classification and phylogenetic analysis

We used two complementary approaches to classify *P. maniculatus* ERVs. First, we examined e-value statistics in the output from our GyDB hmm scans to discern which viral reverse transcriptase domain hmm best fit each ERV. In addition, we also used a phylogenetic approach. We annotated *P. maniculatus* ERVs with their viral origin as predicted by our hmm scans. Next, we downloaded several endogenous and exogenous retroviruses from NCBI (accessions shown

in Figure 3A), extracted their RT domains and annotated them with their respective viral clade. Then, we filtered sequences with large strings of ambiguous characters, performed a multiple sequence alignment of RT genes using MAFFT [76], and generated a maximum likelihood based phylogeny using IQ-TREE *[95]* with a GTR+G model (general time reversible model with unequal rates and unequal base frequencies and discrete gamma rate heterogeneity) [96,97]. We analyzed and edited the resulting phylogeny using ete3 (version 3.1.2) [98], collapsing clusters of *P. maniculatus* ERVs into representative nodes. We visualized the phylogenetic tree using the Interactive Tree Of Life (ITOL) [99] and FigTree (version 1.4.4) (https://github.com/rambaut/figtree).

Searching for deer mouse ERVs in other species

To search for deer mouse ERVs in the house mouse, prairie vole and grasshopper mouse genomes, we performed local BLASTn [75] queries for each full-length deer mouse ERV to each respective genome . We ran BLASTn [75] with the flag *-outfmt 6* and required a minimum alignment length of 400bp and minimum percent identity of 75 to limit possible erroneous hits. As a proof of concept, we also made sure our results were consistent with expectations based on LTR divergences. For example, we would not expect an ERV with highly divergent LTRs (a signature of a more ancient insertion) to be specific to *P. maniculatus*. We also performed broader BLASTn queries against NCBI's nucleotide database. Queries for Gamma_Pman-ERV_cluster-10 sequences only yielded high-confidence hits in deer mouse species, the grasshopper mouse, and the Feline Leukemia Virus (FeLV) reference genome. A neighbor-joining phylogeny constructed from deer mouse Gamma_Pman-ERV_cluster-10 sequences, homologous ERVs in the grasshopper mouse genome, and the FeLV reference suggests a scenario in which Gamma_Pman-ERV_cluster-10 originated in the common ancestor the deer mouse and grasshopper mouse from FeLV or another closely related exogenous virus between 11 and 16 million years ago [36].

TE distribution analysis

We used bedtools intersect [100] to find overlaps between TE annotations and gene feature annotations. We used bedtools closest [100] with the parameter *-s* to identify TE distances from the nearest gene on the same strand and again with default parameters to ignore strand. All functional enrichment tests were performed using goatools [101]. We also tested for jenrichment or depletion of TEs 5kb upstream of genes in the same orientation. Specifically, for each TE subfamily, we randomized all TE locations on each chromosome and compared the number of TEs within 5kb of genes upstream in the same orientation with the observed value. We repeated this 10,000 times to obtain a p-value. Functional enrichment Fisher's Exact p-values and permutation test p-values were adjusted using the bonferroni method to obtain q-values. We used bedtools coverage [100] to calculate ERV and gene density along 100kb windows in the genome. ERV hotspots were defined as windows which exhibit ERV densities within the top 95th percentile. Figure 4F was produced using RIdeogram [102].

KZNF gene family analysis

We defined deer mouse-specific zinc finger (ZF) genes as genes which do not have recognizable orthologs as annotated by NCBI. We employed hmmscan [32] using KRAB hmms downloaded from PFAM [89] to identify KRAB domain-containing ZFs (KZNF). Then, we performed a multiple sequence alignment of all KZNFs using Clustal omega [103] with the parameters *--use-kimura* and *--full* in order to simultaneously produce a pairwise Kimura divergence matrix across all genes. We constructed a subsequent phylogeny using IQ-Tree *[95]* with a general time reversible model. To test for phylogenetic clustering of KZNF that overlapped ERV hotspots, we used phyloclust through RRphylo R package [104] with 100 simulations. Since KZNF genes evolve via a birth-death process, we defined duplicate genes as genes which exhibit the lowest divergence among all pairwise comparisons.

References

1. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19: 199.

2. Kazazian HH Jr. Mobile elements: drivers of genome evolution. Science. 2004;303: 1626–1632.

3. Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature. 1980;284: 604–607.

4. van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 2003;19: 530–536.

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

6. Senft AD, Macfarlan TS. Transposable elements shape the evolution of mammalian development. Nat Rev Genet. 2021;22: 691–711.

7. Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res. 2018;26: 5–23.

8. Ford Doolittle W, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284

9. Brookfield JFY. The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet. 2005;6: 128–136.

10. Venner S, Feschotte C, Biémont C. Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet. 2009;25: 317–323.

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

12. Wells JN, Feschotte C. A field guide to eukaryotic transposable elements. Annu Rev Genet. 2020;54: 539–561.

13. Sotero-Caio CG, Platt RN 2nd, Suh A, Ray DA. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol. 2017;9: 161–177.

14. Abrusán G, Krambeck H-J. Competition may determine the diversity of transposable elements. Theor Popul Biol. 2006;70: 364–375.

15. Eickbush TH. Transposing without ends: the non-LTR retrotransposable elements. New Biol. 1992;4: 430–440.

16. Furano AV, Duvernell DD, Boissinot S. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet. 2004;20: 9–14.

17. Bedford NL, Hoekstra HE. *Peromyscus* mice as a model for studying natural variation. Elife. 2015;4. doi:10.7554/eLife.06813

18. Cantrell MA, Ederer MM, Erickson IK, Swier VJ, Baker RJ, Wichman HA. MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. J Virol. 2005;79: 14698–14707.

19. Wichman HA, Potter SS, Pine DS. Mys, a family of mammalian transposable elements isolated by phylogenetic screening. Nature. 1985;317: 77–81.

20. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44: D81–9.

21. Platt RN 2nd, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016;8: 403–410.

22. Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, et al. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. Mol Ecol Resour. 2021;21: 263–286.

23. Casavant NC, Sherman AN, Wichman HA. Two persistent LINE-1 lineages in Peromyscus have unequal rates of evolution. Genetics. 1996;142: 1289–1298.

24. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438: 803–819.

25. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, et al. Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology. 2018;15: 59.

26. Mager DL, Stoye JP. Mammalian endogenous retroviruses. Microbiol Spectr. 2015;3: MDNA3–0009–2014.

27. Mager DL, Freeman JD. Novel mouse type D endogenous proviruses and ETn elements share long terminal repeat and internal sequences. J Virol. 2000;74: 7221–7229.

28. Ribet D, Dewannieux M, Heidmann T. An active murine transposon family pair: retrotransposition of "master" MusD copies and ETn trans-mobilization. Genome Res. 2004;14: 2261–2267.

29. Erickson IK, Cantrell MA, Scott L, Wichman HA. Retrofitting the genome: L1 extinction follows endogenous retroviral expansion in a group of muroid rodents. J Virol. 2011;85: 12315–12323.

30. Lee RN, Jaskula JC, van den Bussche RA, Baker RJ, Wichman HA. Retrotransposon Mys was active during evolution of the Peromyscus leucopus-maniculatus complex. J Mol Evol. 1996;42: 44–51.

31. Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. The genomic ecosystem of transposable elements in maize. PLoS Genet. 2021;17: e1009768.

32. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39: W29–37.

33. Cui J, Tachedjian G, Wang L-F. Bats and rodents shape mammalian retroviral phylogeny. Sci Rep. 2015;5: 16561.

34. Baillie GJ, van de Lagemaat LN, Baust C, Mager DL. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. J Virol. 2004;78: 5784–5798.

35. León-Paniagua L, Navarro-Sigüenza AG, Hernández-Baños BE, Morales JC. Diversification of the arboreal mice of the genus Habromys (Rodentia: Cricetidae: Neotominae) in the Mesoamerican highlands. Mol Phylogenet Evol. 2007;42: 653–664.

36. Leite RN, Kolokotronis S-O, Almeida FC, Werneck FP, Rogers DS, Weksler M. In the wake of invasion: tracing the historical biogeography of the South American cricetid radiation (Rodentia, Sigmodontinae). PLoS One. 2014;9: e100687.

37. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017;34: 1812–1819.

38. Abramson NI, Lebedev VS, Bannikova AA, Tesakov AS. Radiation events in the subfamily Arvicolinae (Rodentia): evidence from nuclear genes. Dokl Biol Sci. 2009;428: 458–461.

39. Hoover EA, Mullins JI. Feline leukemia virus infection and diseases. J Am Vet Med Assoc. 1991;199: 1287–1297.

40. Polat M, Takeshima S-N, Aida Y. Epidemiology and genetic diversity of bovine leukemia

virus. Virol J. 2017;14: 209.

41. Ribet D, Harper F, Dupressoir A, Dewannieux M, Pierron G, Heidmann T. An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res. 2008;18: 597–609.

42. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9: e1003470.

43. Dechaud C, Volff J-N, Schartl M, Naville M. Sex and the TEs: transposable elements in sexual development and function in animals. Mob DNA. 2019;10: 42.

44. Kent TV, Uzunović J, Wright SI. Coevolution between transposable elements and recombination. Philos Trans R Soc Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0458

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

46. Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. Genome Res. 2017;27: 1384–1394.

47. Sakashita A, Maezawa S, Takahashi K, Alavattam KG, Yukawa M, Hu Y-C, et al. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. Nat Struct Mol Biol. 2020;27: 967–977.

48. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9: 397–405.

49. Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. Nat Rev Mol Cell Biol. 2022; 1–17.

50. Thomas JH, Schneider S. Coevolution of retroelements and tandem zinc finger genes. Genome Res. 2011;21: 1800–1812.

51. Yang P, Wang Y, Macfarlan TS. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. Trends Genet. 2017;33: 871–881.

52. Wolf G, de Iaco A, Sun M-A, Bruno M, Tinkham M, Hoang D, et al. KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. Elife. 2020;9. doi:10.7554/eLife.56337

53. Emerson RO, Thomas JH. Adaptive evolution in zinc finger transcription factors. PLoS Genet. 2009;5: e1000325.

54. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat Biotechnol. 2015;33:

555–562.

55. Kauzlaric A, Ecco G, Cassano M, Duc J, Imbeault M, Trono D. The mouse genome displays highly dynamic populations of KRAB-zinc finger protein genes and related genetic units. PLoS One. 2017;12: e0173746.

56. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 2003;100: 5280–5285.

57. Zemojtel T, Penzkofer T, Schultz J, Dandekar T, Badge R, Vingron M. Exonization of active mouse L1s: a driver of transcriptome evolution? BMC Genomics. 2007;8: 392.

58. Deininger PL, Batzer MA, Hutchison CA 3rd, Edgell MH. Master genes in mammalian repetitive DNA amplification. Trends Genet. 1992;8: 307–311.

59. Fanning TG. Size and structure of the highly repetitive BAM HI element in mice. Nucleic Acids Res. 1983;11: 5073–5091.

60. Grimaldi G, Skowronski J, Singer MF. Defining the beginning and end of KpnI family segments. EMBO J. 1984;3: 1753–1759.

61. Wolf G, Greenberg D, Macfarlan TS. Spotting the enemy within: Targeted silencing of foreign DNA in mammalian genomes by the Krüppel-associated box zinc finger protein family. Mob DNA. 2015;6: 17.

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

63. Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. PLoS Comput Biol. 2016;12: e1004956.

64. Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, et al. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 2012;13: R45.

65. Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 2002;12: 1483–1495.

66. Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH Jr. L1 integration in a transgenic mouse model. Genome Res. 2006;16: 240–250.

67. Gonçalves PR, Christoff AU, Machado LF, Bonvicino CR, Peters FB, Percequillo AR. Unraveling deep branches of the Sigmodontinae tree (Rodentia: Cricetidae) in Eastern South America. J Mamm Evol. 2020;27: 139–160.

68. Yang L, Scott L, Wichman HA. Tracing the history of LINE and SINE extinction in sigmodontine rodents. Mob DNA. 2019;10: 22.

69. Harringmeyer OS, Hoekstra HE. Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. Nature Ecology & Evolution. 2022; 1–15.

70. Hager ER, Harringmeyer OS, Wooldridge TB, Theingi S, Gable JT, McFadden S, et al. A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. Science. 2022;377: 399–405.

71. Baril T, Hayward A. Migrators within migrators: exploring transposable element dynamics in the monarch butterfly, Danaus plexippus. Mob DNA. 2022;13: 5.

72. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117: 9451–9457.

73. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12: 1269–1276.

74. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1: i351–8.

75. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

76. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780.

77. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973.

78. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277.

79. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30: 3276–3278.

80. Storer JM, Hubley R, Rosen J, Smit AFA. Curation guidelines for de novo generated transposable element families. Curr Protoc. 2021;1: e154.

81. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.

82. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nature reviews. Genetics. 2008. pp. 411–2; author reply 414.

83. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR

retrotransposons. Nucleic Acids Res. 2007;35: W265–8.

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

85. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20: 275.

86. Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Bioinformatics. 2019;35: 1051–1052.

87. Rubino F, Creevey CJ. MGkit: Metagenomic framework for the study of microbial communities. 2014. doi:10.6084/m9.figshare.1269288.v1

88. Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 2011;39: D70–4.

89. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49: D412–D419.

90. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45: D200–D203.

91. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43: D222–6.

92. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017;114: E1460–E1469.

93. Bray N, Dubchak I, Pachter L. AVID: A global alignment program. Genome Res. 2003;13: 97–102.

94. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32: W273–9.

95. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37: 1530–1534.

96. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 1994;39: 306–314.

97. Miura RM. Some mathematical questions in biology: DNA sequence analysis. American Mathematical Soc.; 1986.

98. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33: 1635–1638.

99. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49: W293–W296.

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

101. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: a python library for gene ontology analyses. Sci Rep. 2018;8: 10872.

102. Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 2020;6: e251.

103. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7: 539.

104. Castiglione S, Tesone G, Piccolo M, Melchionna M, Mondanaro A, Serio C, et al. A new method for testing evolutionary rate variation and shifts in phenotypic evolution. Methods Ecol Evol. 2018;9: 974–983.