The evolutionary history of a gammaretrovirus currently colonizing the mule deer genome is marked by extensive recombination

Lei Yang[1,2], Raunaq Malhotra[3], Rayan Chikhi[2,3,4], Daniel Elleder[1,5], Theodora Kaiser[1], Jesse Rong[3], Paul Medvedev[2,3,4] and Mary Poss[1,2*]

[1] Department of Biology,
[2] Center for Comparative Genomics and Bioinformatics,
[3] Department of Computer Science and Engineering,
[4] Department of Biochemistry and Molecular Biology,
The Pennsylvania State University, University Park, PA 16802, USA
[5] Institute of Molecular Genetics; The Czech Academy of Sciences; Prague; Czech Republic

*Corresponding author: Mary Poss (maryposs@gmail.com)
Current address: Department of Hematology and Oncology, University of Virginia, Charlottesville, VA 22903, USA

Abstract

Background: All vertebrate genomes have been colonized by retroviruses along their evolutionary trajectory. Although it is clear that endogenous retroviruses (ERVs) can contribute important physiological functions to contemporary hosts, such benefits are attributed to long-term co-evolution of ERV and host. Newly colonized ERVs are thought unlikely to contribute to host genome evolution because germline infections are rare and because the host effectively silences them. The genomes of several outbred species including mule deer (*Odocoileus hemionus*) are currently being colonized by ERVs, which provides an opportunity to study ERV dynamics at a time when few are fixed. Here we investigate the history of cervid endogenous retrovirus (CrERV) acquisition and expansion in the mule deer genome to determine the potential impact of endogenizing retroviruses on host genomic diversity.

Methods: A mule deer genome was de novo assembled from short and long insert mate pair reads. Scaffolds were further assembled using reference assisted chromosome assembly (RACA) to provide spatial orientation of CrERV insertion sites and to facilitate assembly of CrERV sequences. We applied phylogenetic and coalescent approaches to non-recombinant genomes to determine CrERV evolutionary history, augmenting ancestral divergence estimates with the prevalence of each CrERV locus in a population of mule deer. Recombination history was investigated on partial genome alignments.

Results: The CrERV composition and diversity in the mule deer genome has recently measurably increased by horizontal acquisition of a new retroviruses lineage and because of recombination with existing CrERV. Resulting interlineage recombinants also endogenized and subsequently retrotransposed. CrERV loci are significantly closer to genes than expected if integration were random and gene proximity might explain the recent expansion by retrotransposition of one recombinant CrERV lineage.

Conclusions: There has been a burst of CrERV integrations during a recent retrovirus epizootic that increased genomic CrERV burden and has resulted in extensive insertional polymorphism in contemporary mule deer genomes. Recombination is a defining feature of CrERV evolutionary dynamics driven by this colonization, increasing CrERV burden and CrERV genetic diversity. These data support that retroviral colonization during an epizootic provides a burst of genomic diversity to the host population.

Keywords: endogenous retrovirus, ERV, recombination, genome diversity, mule deer, Odocoileus hemionus, CrERV

Background

Retroviruses are unique among viruses in adopting life history strategies that enable them to exist independently as an infectious RNA virus (exogenous retrovirus, XRV) [1] or as an integral component of their host germline (endogenous retrovirus, ERV) [2,3].

An ERV is the result of a rare infection of a germ cell by an XRV and is maintained in the population by vertical transmission. Germline colonization has been a successful strategy for retroviruses as they comprise up to 10% of most contemporary vertebrate genomes [4]. Over the evolutionary history of the species, ERV composition increases by acquisition of new germ line XRV infections, and through retrotransposition or reinfection of existing ERVs [5–8], which results in clusters of related ERVs. The ERV profile in extant species therefore reflects both the history of retrovirus epizootics and the fate of individual ERVs. Because the acquisition of retroviral DNA in a host genome has the potential to affect host phenotype [9–11], the dynamic interactions among ERVs and host could shape both retrovirus and host biology. However, the evolutionary processes in play near the time of colonization are difficult to discern based on an ERV colonization event that occurred in an ancestral species. A better understanding of both host and virus responses to recent germ line invasion might inform homeostatic changes in ERV-host regulation that are relevant to the pathogenesis of diseases in which ERV involvement has been implicated [12–17]. Fortunately, there is now evidence that retrovirus colonization is occurring in contemporary, albeit often non-model, species [18–20], allowing for investigation of ERV dynamics near the time of colonization. Our goal in this research is to investigate the evolutionary dynamics of the phylogenetically distinct ERV lineages that have sequentially colonized mule deer over the approximate million-year history of this species using the complete genome sequence of a majority of coding ERVs in the context of a draft assembly of a newly sequenced mule deer genome.

The life history strategy adopted by retroviruses indicates why this virus family has been so successful in colonizing host germline. Retroviral replication requires that the viral RNA genome be converted to DNA and then integrated into the genome of an infected cell [21]. As with many RNA viruses, the virus polymerase enzyme, reverse transcriptase (RT), is error prone, which contributes to a high mutation rate and enables rapid host adaptation. In addition, RT moves between the two RNA copies that comprise a retroviral genome [22]; this process can repair small genomic defects and increases evolutionary rates via recombination if the two strands are not identical. Retroviral DNA is called a provirus and is transported to the nucleus where it integrates into host genomic DNA using a viral integrase enzyme. The provirus represents a newly acquired gene that persists for the life of the cell and is passed to daughter cells, which for XRV are often hematopoietic cells. For a retrovirus infecting a germ cell, all cells in an organism will contain the new retroviral DNA if reproduction of the infected host is successful.

The retroviral life cycle also demonstrates how ERVs can affect host biology [10,23]. ERVs require host transcription factors and RNA polymerases to bind to the retrovirus promoter, called long terminal repeats (LTRs), to produce viral transcripts and the RNA genome. Thus, the viral LTRs compete with host genes for transcription factors and polymerases [24]. A retrovirus encodes at a minimum, genes for the capsid, viral enzymes, and an envelope gene needed for cell entry, which is produced by a sub-genomic mRNA. Hence an ERV also utilizes host-splicing machinery and can alter host gene expression pattern if the site of integration is intronic [25,26]. While XRVs are expressed from small numbers of somatic cells, ERVs are present in all cells and ERV transcripts and proteins can be expressed in any cell type at any stage of host development. Hosts actively silence the expression of full or partial ERV sequences by epigenetic methods [27,28] and by genes called viral restriction factors [29–33]. Because there will be no record of an ERV that causes reproductive failure of the newly colonized host, ERVs in contemporary vertebrates are either effectively controlled by host actions, are nearly neutral in effects on host fitness, or potentially contribute to the overall fitness of the host [34–37].

The coding portion of a new ERV can be eliminated from the genome through non-allelic homologous recombination (NAHR) between the LTRs, which are identical regions that flank the viral coding portion. A single LTR is left at the site of integration as a consequence of the recombination event and serves as a marker of the original retrovirus integration site [38]. Most ERV integration sites in humans are solo LTRs [39,40]. Because the efficiency of NAHR is highest between identical sequences [41], conversion of a full-length ERV to a solo LTR likely arises early during ERV residency in the genome before sequence identity of the LTR is lost as mutations accrue [42]. Because mutations are reported to arise in ERVs at the neutral mutation rate of the host [43], sequence differences between the 5’ and 3’ LTR of an ERV have been used to approximate the date of integration [44,45].

Although in humans most ERV colonization events occurred in ancestral species, acquisition of new retroviral elements is an ongoing [46,47] or contemporary [18] event in several animal species. The consequences of a recent ERV acquisition are important to the host species because it creates an insertionally polymorphic site; the site is occupied in some individuals but not in others. All ERVs are insertionally polymorphic during the trajectory from initial acquisition to fixation or loss in the genome. Indeed, the HERV-K (human endogenous retrovirus type K) family is insertionally polymorphic in humans [48–51] and HERV-K prevalence at polymorphic sites differ among global populations [52]. Phylogenetic analyses of the ERV population in a genome can inform on the origins of ERV lineages to determine which are actively expanding in the genome and the mutational processes that drive evolution. These data indicate if expansion is related to the site of integration or a feature of the virus, or both and coupled with information of ERV prevalence at insertionally polymorphic sites, can inform ERV effects on host phenotype.

To this end, we explored the evolutionary history of the mule deer (*Odocoileus hemionus*) ERV (Cervid endogenous retrovirus, CrERV) because we have extensive data for prevalence of CrERV loci in northern US mule deer populations [53] and preliminary data on CrERV sequence variation and colonization history [19,54]. A majority of CrERV loci is insertionally polymorphic in mule deer; 90% of animals shared fewer than 10 of approximately 250 CrERV integrations per genome in one study [53]. Further, mule deer appear to have experienced several recent retrovirus epizootics with phylogenetically distinct CrERV and, because none of the CrERV loci occupied in mule deer are found in the sister species, white-tailed deer (*Odocoileus virginianus*) [19], all endogenization events have likely occurred since the split of these sister taxa. Based on the phylogeny of several CrERV identified in the mule deer genome, at least four distinct epizootics resulted in germ line colonization [54]. A full-length retrovirus representing the youngest of the CrERV lineages was recovered by co-culture on human cells, indicating that some of these CrERV are still capable of infection [55]. In this study, we expand on these preliminary data by sequencing a mule deer genome and conducting phylogenetic analyses on a majority of reconstructed CrERV genomes. Our results demonstrate that expression and recombination of recently acquired CrERV with older CrERV have increased CrERV burden and diversity and consequently have increased contemporary mule deer genome diversity.

Results

Establishing a draft mule deer reference genome to study CrERV evolution and integration site preference

We developed a draft assembly of a mule deer genome from animal MT273 in order to determine the sequence at each CrERV locus for phylogenetic analyses and to investigate the effect of CrERV lineage or age on integration site preference. ERV sequences are available in any genome sequencing data because a retrovirus integrates a DNA copy into the host genome. However, there is extensive homology among the most recently integrated ERVs making them difficult to assemble and causing scaffolds to break at the site of an ERV insertion [56]. We assembled scaffolds using a combination of high coverage Illumina short read whole genome sequencing (WGS) and long insert mate pair sequencing. Our *de novo* assembly yielded an ~3.31 Gbp draft genome with an N50 of 156 Kbp (Table S1), which is comparable to the 3.33

Gbp (c-value of 3.41 pg) experimentally-determined genome size of reindeer (*Rangifer tarandus*) [57,58].

Approximately half of CrERV loci are located at the ends of scaffolds based on mapping our previously published junction fragment sequences [53], which is consistent with the fact that repetitive elements such as ERVs break scaffolds. To determine the sequence of these CrERVs and the genome context in which they are found, we developed a higher order assembly using reference assisted chromosome assembly (RACA) [59]. RACA further scaffolds our *de novo* mule deer assembly into ‘chromosome fragments’ by identifying synteny blocks among the mule deer scaffolds, the reference species genome (cow), and the outgroup genome (human) (Figure 1A). We created a series of RACA assemblies based on scaffold length to make efficient use of all data (Table S1). RACA150K takes all scaffolds greater than 150,000 bp as input and yielded 41 chromosome fragments, 35 of which are greater than 1.5 Mbp; this is consistent with the known mule deer karyotype of 2n=70 [60]. However, RACA150K only incorporates 48% of the total assembled sequences (1.59 Gbp) because of the scaffold size constraint. In contrast, RACA10K uses all scaffolds 10,000 bp or longer and increases the assembly size to 2.37 Gbp (~72% of total assembly) but contains 658 chromosome fragments (Table S1). The majority of scaffolds that cannot be incorporated into a RACA assembly are close to the ends of alignment chains (File S1, section 1a). Most sequences not represented in any assemblies were repeats based on *k-mer* analyses (File S1, section 1a and Figure S1).

Some scaffolds were excluded from the RACA assemblies, presumably because there is no synteny between cow and human for these sequences. We oriented these scaffolds using the cow-mule deer and sheep-mule deer alignments (RACA+, Table S2). Approximately 124 Mbp of sequence (~4% of total assembly) is in scaffolds larger than 10kb but cannot be placed in RACA10K, nearly all of which can be found on the mule deer-cow alignment chain and the mule deer-sheep (oviAri3) alignment chain (123 Mbp in each chain). Because there is overlap between these alignments, only ~1 Mbp is specific to cow and ~1 Mbp is specific to sheep. Therefore, RACA+ incorporated all but 69 scaffolds that are greater than 10 kbp, which consisted of 1.17 Mbp of sequence (~0.04% of total scaffold size of the assembly) and yields an assembly size of 2.49 Gbp (Table S1).

To enable the investigation of CrERV integration site preference relative to host genes, we annotated the mule deer scaffolds. We used Maker2 [61,62] for the annotation, which detects candidate genes based on RNA sequencing data and protein homology to any of the three reference genomes: human, cow and sheep. After four Maker iterations, 21,598 genes with an AED (annotation edit distance) [61] of less than 0.8 were annotated (Table S3). Approximately 92% of genes are found on RACA150K scaffolds and 95% of genes are represented in RACA10K scaffolds.

Establishing the location and sequence at CrERV loci

Several lines of evidence suggest that most CrERVs are missing from the assemblies. Only three CrERVs with coding potential were assembled by the *de novo* assembly.

The *k-mer* based analysis shows that less than 9.62% of all LTR repeat elements are in the assemblies (Table S4). The CrERV-host junction fragments previously sequenced [53] support that CrERV loci are near scaffold ends or in long stretches of 'N's. Therefore, we took advantage of the different chromosome fragments generated by RACA10K, RACA150K and RACA+ and the long insert mate pair sequencing data to reconstruct CrERVs at each locus (Figure 1B). We identified 252 CrERV loci in the MT273 genome, which is consistent with our estimates of an average of 240 CrERV loci per mule deer by quantitative PCR [19] and 262 CrERV loci in animal MT273 by junction fragment analysis [53]. The majority of CrERV loci (206/252) contains CrERVs with some coding capacity and 46 are solo LTRs. Of the 206 CrERVs containing genes, 164 were sufficiently complete to allow phylogenetic analysis on the entire genome or, if a deletion was present, on a subset of viral genes; at 42 loci we were unable to obtain sufficient lengths of high-quality data for further analyses.

Evolutionary history of CrERV

We previously showed that mule deer genomes have been colonized multiple times since the ancestral split with white tailed deer approximately one million years ago (MYA) [54] because none of the CrERV integration sites are found in white-tailed deer. To better resolve the colonization history, we conducted a coalescent analysis based on an alignment spanning position 1,477-8,633 bp (omitting a portion of *env*) of the CrERV genome (GenBank: JN592050) using 34 reconstructed CrERV sequences with high quality data that had no signature of recombination and that were representative of the phylogeny in a larger data set (Figure 2). The majority of the *env* gene, which has distinct variable and conserved region [63], was manually blocked because of alignment difficulties (6,923-7,503 bp by JN592050 coordinates; see Figure 2, right panel for diagram of *env* variable regions and Table S5, column C for *env* structure of each CrERV). This tree shows four well-supported CrERV lineages, each diverged from a common ancestor at several points since the split of mule deer and white-tailed deer. Although *env* sequence is not included in the phylogenetic analysis, CrERV assigned to each of the four identified lineages share the same distinct *env* variable region structure of insertions and deletions, which define the receptor-binding domain of the envelope protein (Figure 2, right panel).

Lineage A CrERVs are the youngest ERV family in mule deer. Our estimates indicate that Lineage A colonization has occurred over the last 300 thousand years to the present (Figure 2; Table S6, node f, 95% high posterior density (HPD) interval 110-470 thousand years ago (KYA)) and is represented by three well-supported CrERV subgroups evolving over this time frame. All have a complete open reading frame (ORF) in *env* and likely represent a recent retrovirus epizootic. An infectious virus recovered by co-culture belongs to this lineage [19]. Lineage A represents 30% of all CrERV sampled from MT273 (Table S5). Our age estimates for each subgroup of Lineage A CrERV are consistent with their prevalence in populations of mule deer in the Northern Rocky Mountain ecosystem (Figure 2); [64]. For example, S29996 and S10113 are estimated to derive from an older Lineage A CrERV subgroup and occur in our sampled mule deer at higher prevalence than those estimated to have entered the genome more recently (see S22897 and S111665, Figure 2).

Lineage B CrERV shared a common ancestor with Lineage A approximately 300 KYA (node i, Figure 2). Lineage B CrERVs have a short insertion in the 5' portion of *env* followed by a deletion that removes most of the *env* surface unit (SU) relative to Lineage A *env*. Because our coalescent analysis does not include *env* sequence, these results suggest that two phylogenetically distinct XRV with different envelope proteins were circulating about the same time in mule deer populations. Lineage B CrERV represent 32% of sampled viruses from our sequenced genome (Table S5). Like Lineage A, the prevalence of CrERV from Lineage B among mule deer in the northern Rockies region is low, reflecting their more recent colonization of the mule deer genome. Indeed, six Lineage B CrERVs were identified only in MT273, while only one Lineage A CrERV is found only in MT273 (Table S5), which could be indicative of a recent expansion of some Lineage B CrERV. Of note, there are two related groups of CrERV affiliated with Lineage B (Lineage B1 and B2, Table S5). One shares the short 5' insertion in *env* but has a full-length *env* with an additional short insertion relative to the *env* of Lineage A CrERV (Lineage B1, Figure 2). CrERV with this *env* configuration represent 9% of coding CrERV in MT273. Because the prevalence of Lineage B1 is high in mule deer, this group could represent the ancestral state for Lineage B CrERVs. The second group has a unique *env* not found in any other CrERV lineages (Lineage B2, Figure 2, node k; S16113 and S6404). We are unable to estimate the prevalence of this unusual *env* containing CrERV in mule deer because the host junction fragments are not represented in our draft mule deer assembly. It is possible that these viruses represent a cross-species infection and it would be interesting to determine if representatives of Lineage B2 are found in the genomes of other species that occupied the ecosystem in the past.

Our coalescent estimates indicate that Lineage C CrERV emerged about 500 KYA (Table S6). Several members of this lineage are found in all mule deer sampled (Figure 2; Table S5), consistent with a longer residence in the genome. There is a 59 bp insertion (C) and 362 bp deletion (E) in *env* (Figure 2; Table S5) compared to the full length *env* of Lineage A; none have an intact *env* ORF. Despite evidence that Lineage C is an older CrERV, the approximately 13% of identified CrERV in MT273 belonging to this lineage share a common ancestor ~50 KYA (95% HPD: 16-116 KYA, Table S6). These data are consistent with a recent expansion of a long-term resident CrERV.

The first representatives of the CrERV family still identifiable in mule deer colonized shortly after their split from white-tailed deer, approximately one MYA [19]. Lineage D CrERVs comprise 12% of reconstructed CrERV in MT273 and appear to be near fixation. Indeed, all mule deer in a larger survey of over 250 deer had CrERV S26536, which is not found in white-tailed deer [54]. This lineage shares an *env* insertion with Lineage C but lacks the deletion, which removes the transmembrane region of *env*.

These data expand our previous findings that over the approximately one million year history of mule deer, the mule deer genome has been colonized at least four times by phylogenetically distinct CrERVs; this likely reflects several retroviral epizootics each characterized by a unique *env* structure. The two lineages responsible for most recent endogenization events comprise 62% of sampled CrERV. In addition, these data capture the evolutionary processes acting on the *env* gene of exogenous retroviruses, which are characterized by gain or loss of variable regions of this important viral protein.

Recombination among CrERV lineages

Our coalescent estimates (Figure 2) indicate that two phylogenetically distinct CrERV lineages have been expanding in contemporary mule deer genomes over the last 100,000 years. Both lineages have been actively colonizing contemporary mule deer genomes based on divergence estimates, which include zero. While CrERVs represented by Lineage A are capable of infection [19], all Lineage B CrERVs have an identical deletion of the SU portion of *env* and should not be able to spread by reinfecting germ cells. However, the mule deer genome is comprised of approximately equal percentages of Lineage B and Lineage A CrERVs so we considered two modes by which defective Lineage B CrERVs could expand in the genome at a similar rate with Lineage A. Firstly, ERVs that have lost *env* are proposed to preferentially expand by retrotransposition [65] because a functional envelope is not necessary for intracellular replication. Secondly, we consider that Lineage B CrERVs could increase in the genome by infection if the co-circulating Lineage A group provided a functional envelope protein, a process called complementation [5,66]. This latter mechanism requires that a member of each CrERV lineage be transcriptionally active at the same time in the same cell, and that intact proteins from the ‘helper’ genome be used to assemble a particle with a functional envelope for reinfection. If two different CrERV loci are expressed in the same cell, both genomes could be co-packaged in the particle.

Because the reverse transcriptase moves between the two RNA genomes as first strand DNA synthesis proceeds, evidence of inter-lineage recombination would support that the molecular components necessary for complementation were in place. We assessed Lineage B CrERV for recombination with Lineage A to determine if coincident expression of the RNA genomes of these two lineages could explain the expansion by infection through complementation of the *env*-less Lineage B CrERV.

There is good support for recombination between Lineage A and B in a region spanning a portion of *pol* to the beginning of the variable region in *env* (4,422-7,076 based on coordinates of JN592050). In this region, several CrERV that we provisionally classified as Lineage B because they carried the prototypical *env* deletion of SU form a monophyletic group that is affiliated with Lineage A CrERV (Figure 3, upper collapsed clade containing orange diamonds). These Lineage B recombinants all share the same recombination breakpoint just 5' of the characteristic short insertion for these viruses (Figure S2, indicated by "**"; Table S7). In addition, several other CrERVs with Lineage B *env* branch between lineages A and B, indicating that the recombination breakpoints fall within the region assessed (Figure S2). Indeed, the breakpoint in a group of three CrERV is at position 6630 based on coordinates of JN592050, which is near the predicted splice site for *env* at position 6591 [19]; this confers an additional 500 bp of the Lineage B *env* on these viruses (Figure S2) resulting in their observed phylogenetic placement. Because recombination between the two retroviral RNA genomes occurs during reverse transcription, our data indicate that both Lineage A and B CrERVs were expressed and assembled in a particle containing a copy of each genome. A functional envelope from Lineage A would therefore have been available for infection. These data support our premise that complementation with a replication competent Lineage A CrERV or CrXRV (cervid exogenous gammaretrovirus, an exogenous version of CrERV) contributes to the 32% prevalence of *env*-deleted Lineage B CrERV in the genome. It is notable that recent retrotransposition of the lineage A-B recombinant CrERVs likely occurred because they are nearly identical and the branches supporting them are short (Figure 3, orange diamonds in the Lineage A type *env* cluster).

There is additional data to support the transcriptional activity of a Lineage B CrERV, which is requisite for recombination with an infectious Lineage A CrERV or for retrotransposition. We identified a non-recombinant Lineage B CrERV (S24870 in Table S5) with extensive G to A changes (184 changes) compared to other members of this monophyletic group. These data are indicative of a cytidine deaminase acting on the single stranded DNA produced during reverse transcription [67].

Lineage C CrERV are enigmatic because based on full length sequences lacking a signature of recombination it diverged around 500KYA (Figure 2) but all extant members of this group diverged recently. From Figure 3, it is evident that over the region of *pol* assessed, CrERVs containing the Lineage C *env* cluster with an older Lineage A subgroup. Given that the *env* of Lineage C CrERV shares sequence homology and an insertion with that of the oldest Lineage D, it is likely that Lineage C is in fact the result of recombination between an early member of Lineage A and a relative of a Lineage D CrERV. Many, but not all, Lineage C CrERVs are found at high prevalence in the mule deer population (Figure 2; Table S5), supporting that the initial recombination event occurred early during the Lineage A colonization. Our identification of Lineage C as derived from a non-recombinant CrXRV is therefore incorrect. Instead, Lineage C CrERVs are derived from a CrERV or CrXRV that is not currently represented in mule deer genomes either because it was lost or it never endogenized. Fourteen of the twenty-two CrERV in Lineage C have multiple signatures of recombination predominantly with Lineage A CrERV. The expansion of a subset of Lineage C as a monophyletic group approximately 50 KYA (Figure 2; Table S6) suggests that like some members of Lineage B, CrERVs generated by recombination with Lineage A have recently retrotransposed.

Genomic distribution of CrERV lineages

Of the 164 CrERV that we reconstructed from MT273, only 12 can be detected in all mule deer that we have sampled [53,54] (Table S5). This means that the majority of CrERV loci in mule deer are insertionally polymorphic; not all animals will have a CrERV occupying a given locus. ERVs can impact genome function in multiple ways but the best documented is by altering host gene regulation, which occurs if the integration site is near a host gene [68]. Thus, we investigated the spatial distribution of CrERV loci relative to host genes to determine the potential of either fixed or polymorphic CrERV to impact gene expression, which could affect host phenotype.

The actual distance between genes is likely to be unreliable in our assembly because most high copy number repeats are missing in the mule deer assembly (Figure S1, Table S4, section 1a of File S1). To investigate potential problems determining the spatial distribution of CrERV insertions imposed by using a draft assembly, we simulated the distribution of retrovirus insertions (File S1, section 2l) in mule deer (scaffold N50=156 Kbp) and the genomes of cow (Btau7, scaffold N50=2.60 Mbp) and human (hg19, scaffold N50=46.4 Mbp). The mean distance between insertion and the closest gene for all simulation replicates (Figure S3) is significantly higher in the cow and human (Mann-Whitney U test $p < 2.2\text{x}10^{-16}$ for any pair of comparison among the three species). Therefore, we determined if the number of CrERV loci observed to be within 20Kbp of a gene differed from that expected if the distribution was random. There are significantly more observed insertions that fall within 20 Kbp of the translation start site of a gene than occur randomly (Figure 4A). In contrast, intronic CrERV insertions are significantly less than expected based on our simulations (Figure 4B). Among Lineage A CrERVs, only a single sub-lineage (CrERVs that are associated to node 'a' in Figure 2) are found in closer proximity to genes (bold font in Column G of Table S5) than expected if integrations are random (Fisher's exact test p = 0.002891). We also investigated whether any of the recombinant CrERV with a signature of recent expansion was integrated within 20 Kbp of a gene. Two of the three recombinant clusters (Figure 3) contain members that are close to a gene (Table S5, bold font in column G). In particular, Lineage A/B recombinant CrERV S10 is 494 bp from the start of a gene. Remarkably, four Lineage C CrERVs with the typical *env* sequence are within 20 Kbp of a gene (Table S5, bold font in column G). Our data indicate that integration site preference overall favors proximity to genes but that this is not reflected in all lineages. In particular, the history of Lineage C CrERV suggests they could have acquired a different integration site preference through recombination that facilitated recent genome expansion.

Discussion

The wealth of data on human ERVs (HERVs) provides the contemporary status of events that initiated early in hominid evolution. Potential impacts of an ERV near the time of colonization on a host population is thought to be minimal because infection of host germ line by an XRV is a rare event and ERVs that affect host fitness are quickly lost. Potentially deleterious ERVs that are not lost due to reproductive failure can be removed by recombination leaving a solo LTR at the integration site or can suffer degradation presumably because there is no benefit to retain function at these loci; most HERVs are represented by these two states. In addition, humans and other vertebrate hosts have invested extensive genomic resources [4,9,69] to control the expression of ERVs that are maintained. The dynamics between host and ERV are described as an evolutionary arms race [70,71]. This narrative may underrepresent any contributions of ERVs to fitness as they were establishing in a newly colonized host population. Because there are now several species identified to be at different points along the evolutionary scale initiated by the horizontal acquisition of retroviral DNA it is possible to investigate dynamics of ERV that are not yet fixed in a contemporary species. Considering the numerous mechanisms by which newly integrated retroviral DNA affect host biology, such as by introducing new hotspots for recombination [72], altering host gene regulation [68,73,74], and providing retroviral transcripts and proteins for host exaptation [75–78], colonizing ERVs could make a substantive contribution to species' evolution. Our research on the evolutionary dynamics of mule deer CrERV demonstrates that genomic CrERV content and diversity increased significantly during a recent retroviral epizootic due to acquisition of new XRV and from endogenization and retrotransposition of recombinants generated between recent and older CrERVs. These data suggest that CrERV provide a pulse of genetic diversity, which could impact this species' evolutionary trajectory.

Our analyses of CrERV dynamics in mule deer are based on the sequence of a majority of coding CrERVs in MT273. Of the 252 CrERV loci identified in the MT273 assembly, we were able to reconstruct CrERV sequences from long insert mate pair and Sanger sequencing to use for phylogenetic analysis at 164 sites; 46 sites were solo LTR and 42 were occupied by CrERV retaining some coding capacity. We complimented phylogenetic analyses with our previous data on the frequency of each CrERV locus identified in MT273 in a population of mule deer in the northern Rocky Mountain ecosystem [53,64]. In addition, we incorporated information on the variable structure of the retroviral envelope gene, *env*, which is characteristic of retrovirus lineages but was excluded from phylogenetic analyses. The variable regions of retroviral *env* result from balancing its role in receptor-mediated, cell specific infection while evading host adaptive immune response [79,80]. Despite excluding most of *env* from our phylogenic analysis because of alignment problems, each of the lineages we identified has a similar distinct *env* structure, as is well known for infectious retroviruses. By integrating population frequency, coalescent estimation, and the unique structural features of *env* we provide an integrated approach to explore the evolutionary dynamics of an endogenizing ERV.

The most recent CrERV epizootic recorded by germline infection was coincident with the last glacial period, which ended about 12 KYA. The retroviruses that endogenized during this epizootic belong to Lineage A, have open reading frames for all genes and have been recovered by co-culture as infectious viruses [55]. There are several sub-lineages within Lineage A, which likely reflect the evolutionary history of CrXRV contributing to germline infections over this time period. Lineage A retroviruses constitute approximately one third of all retroviral integrations in the genome. Only four of the fifty Lineage A CrERV that we were able to reconstruct did not have a full length *env*. An important implication of this result is that over the most recent approximately 100,000 years of the evolution of this species, the mule deer genome acquired up to half a megabase of new DNA, which introduced new regulatory elements with promoter and enhancer capability, new splice sites, and sites for genome rearrangements. Thus, there is a potential to impact host fitness through altered host gene regulation even if host control mechanisms suppress retroviral gene expression. None of the Lineage A CrERV is fixed in mule deer populations (Table S5, column F) so any effect of CrERV on the host will not be experienced equally in all animals. However, none of the Lineage A CrERV is found only in M273 indicating that the burst of new CrERV DNA acquired during the most recent epizootic has not caused reproductive failure among mule deer. These data demonstrate that in mule deer, a substantial accrual of retroviral DNA in the genome can occur over short time spans in an epizootic and could impose differential fitness in the newly colonized population.

Lineage A CrERV has an open reading frame for *env* but Lineages B-D do not. Lineage B CrERVs are intriguing in this regard because they also constitute approximately a third of the CrERV in the genome. Yet all have identical deletions of the extracellular portion of *env*, which should render them incapable of genome expansion by reinfection. ERV that have deleted *env* are reportedly better able to expand by retrotransposition [65], which could account for the prevalence of Lineage B. However, because we have evidence for recent expansion of Lineage A and B recombinants, we considered an alternative explanation; that *env*-deficient Lineage B CrERV was complemented with an intact Lineage A CrERV envelope glycoprotein allowing for germline infection. Complementation is not uncommon between XRV and ERV [81,82], is well established for murine Intracisternal A-type Particle (IAP) [83] and has been reported for ERV expansion in canids [84]. Complementation requires that two different retroviruses are co-expressed in the same cell [85]. During viral assembly functional genes supplied by either virus are incorporated into the virus particle and either or both retroviral genomes can be packaged. Because the retroviral polymerase uses both strands of RNA during reverse transcription to yield proviral DNA, a recombinant can arise if the two co-packaged RNA strands are not identical. We investigated the possibility of complementation by searching for Lineage A-B recombinants. Our data show that Linage A and B recombination has occurred several times. A group of CrERV that encode a Lineage B *env* cluster with Lineage A CrERV in a phylogeny based on a partial genome alignment (JN592050: 4422-7076bp). The recombinant breakpoint within this monophyletic group is identical, suggesting that the inter-lineage recombinant subsequently expanded by retrotransposition. Notably, two of the CrERV in this recombinant cluster were only found in M273, indicating that retrotransposition was a recent event. There are other clusters of CrERV with Lineage B *env* affiliated with Lineage A CrERV that have different breakpoints in this partial phylogeny. Recombination between an XRV and ERV is also a well-documented property of retroviruses [86–88]. However, the recombinant retroviruses that result are typically identified because they are XRV and often associated with disease or a host switch. Our data indicate that multiple recombination events between Lineage A and B CrERV have been recorded in germline; this in itself is remarkable given that endogenization is a rare event. Thus both the burden of CrERV integrations and the sequence diversity of CrERV in the mule deer genome increase concomitant with a retrovirus epizootic by CrERV inter-lineage recombination.

Recombination is a dominant feature of CrERV dynamics and is also displayed in the evolutionary history of Lineage C CrERV. Our phylogenetic analysis places the ancestor of Lineage C CrERV at 500 KYA and indeed, Lineage C and Lineage D, which is estimated to be the first CrERV to colonize mule deer after splitting from white-tailed deer [19,54], share many features in *env* that are distinct from those of Lineage A and B. Consistent with a long-term residency in the genome, many Lineage C CrERV are found in most or all mule deer surveyed. A recent expansion of a CrERV that has been quiescent in the genome since endogenizing could explain the estimated 50 KYA time to most recent common ancestor of extant members of this lineage. Although this scenario is consistent with the paradigm that a single XRV colonized the genome and recently expanded by retrotransposition, our analysis shows that all Lineage C CrERV are recombinants of a Lineage A CrERV and a CrERV not recorded in or lost from contemporary mule deer genomes. Hence the resulting monophyletic lineage does not arise from retrotransposition of an ancient colonizing XRV. Rather, as is the case with Lineage B CrERV, recombination between an older CrERV and either a Lineage A CrXRV or CrERV occurred, infected germline, and recently expanded by retrotransposition. It is noteworthy that all retrotransposition events detectable in our data involve recombinant CrERV. Further, recombination often leads to duplications and deletions in the retroviral genome, therefore some of the deletions we document in Lineages B-D are not a consequence of slow degradation in the genome but rather are due to reverse transcription and as was recently reported for Koala retrovirus [88]**.**

These data highlight that expansion of CrERV diversity and genomic burden has occurred in the recent evolutionary history of mule deer by new acquisitions, complementation, and pulses of retrotransposition of inter-lineage recombinants. Indeed, several of the recombinant Lineage C CrERVs that have expanded by retrotransposition are within 20kbp of a gene raising the question as to whether there is a fitness effect at these loci that is in balance with continued expression of the retrovirus. It is remarkable that so many of the events marking the dynamics of retrovirus endogenization are preserved in contemporary mule deer genomes. Given that germline infection is a rare event, it is likely that the dynamics we describe here also resulted in infection of somatic cells. It is worthwhile to consider the potential for ERVs in other species, in particular in humans where several HERVs are expressed, to generate novel antigens through recombination or disruptive somatic integrations that could contribute to disease states.

Methods

Sequencing

Whole genome sequencing (WGS) was performed for a male mule deer, MT273, at ~30x depth using the library of ~260 bp insert size, ~10x using the library of ~1,400-5,000 bp insert size and ~30x using the library of ~6,600 bp insert size. 3' CrERV-host junction fragment sequencing was performed as described by Bao *et al.* [53]. 5' CrERV-host junction fragment sequencing was performed on the Roche 454 platform, with a target size of ~500bp containing up to 380 bp of CrERV LTR.

Assembly and mapping

The draft assembly of MT273 was generated using SOAPdenovo2 [89] (File S1, section 2a). WGS data were then mapped back to the assembly using the default setting of bwa mem [90] for further use in RACA and CrERV reconstruction. RNA-seq data was mapped to the WGS scaffolds using the default setting of tophat [91,92]. 3' junction fragments were clustered as described in Bao *et al.* [53]. 3' junction fragment clusters and 5' junction fragment reads were mapped to the WGS assembly using the default setting of blat [93]. A perl script was used to filter for the clusters or reads whose host side of the fragment maps to the host at its full length and high identity. 5' junction fragments were then clustered using the default setting of bedtools merge.

RACA

Synteny based scaffolding using RACA was performed based on the genome alignment between the mule deer WGS assembly, a reference genome (cow, bosTau7 or Btau7), and an outgroup genome (hg19). Genome alignments were performed with lastz [94] under the setting of '--notransition --step=20', and then processed using the UCSC axtChain and chainNet tools. The mule deer-cow-human phylogeny was derived from Bininda-Emonds *et al*. [95] using the 'ape' package of R.

CrERV sequence reconstruction

CrERV locations and sequences were retrieved based on junction fragment and long insert mate pair WGS data. The long insert mate pair WGS reads were mapped to the reference CrERV (GenBank: JN592050) using bwa mem. Mates of reads that mapped to the reference CrERV were extracted and then mapped to the WGS assembly using bwa mem. Mates mapped to the WGS assembly were then clustered using the 'cluster' function of bedtools. Anchoring mate pair clusters on both sides of the insertion site were complemented by junction fragments to localize CrERVs. Based on the RACA data, CrERVs that sit between scaffolds were also retrieved in this manner. CrERV reads were then assigned to their corresponding cluster and were assembled using SeqMan (DNASTAR). Sanger sequencing was performed to complement key regions used in CrERV evolutionary analyses. All reconstructed CrERV sequences used in the phylogenetic analyses are included in File S2 in fasta format.

CrERV evolution analyses

CrERV sequences of interest were initially aligned using the default setting of muscle [96], manually trimmed for the region of interest, and then re-aligned using the default setting of Prank [97]. Lineage-specific regions are manually curated to form lineage-specific blocks. Models for phylogeny were selected by AICc (Akaike Information Criterion with correction) using jModelTest [98]. Coalescent analysis and associated phylogeny (Figure 2) was generated using BEAST2 [99]. In the coalescent analysis, we used GTR substitution matrix, four Gamma categories, estimated among-site variation, Calibrated Yule tree prior with ucldMean ucldStddev from exponential distribution, relaxed lognormal molecular clock, shared common ancestor of all CrERVs 0.47-1 MYA as a prior [19,54]. Maximum likelihood phylogeny in Figure 3 was generated using PhyML [100] using the models selected by AICc and the setting of '-o tlr -s BEST' according to the selected model.

CrERV spatial distribution

We simulated 274 insertions per genome to approximate the average number of CrERVs in a mule deer [53]. The simulation was performed 10,000 times on three genomes: the mule deer WGS scaffolds, cow (Btau7) and human (hg19). Distance between simulated insertions and the closest start of the coding sequence of a gene was calculated using the 'closest' function of bedtools, and the simulated insertions that overlap with a gene were marked with the 'intersect' function of bedtools. Number of simulated simulations that are within 20 Kbp or intronic to a gene was counted for each of the 10,000 replicates. Counts were then normalized by the total number of insertions and plotted using the 'hist' function of R.

Supplementary methods

Methods with extended details are available in File S1.

Availability of supporting data

The raw sequencing data was deposited in SRR9121136. Other data generated are included in supplementary file and figures.

List of abbreviations

ERV: endogenous retrovirus

XRV: exogenous retrovirus or infectious retrovirus

LTR: long terminal repeat

CrERV: cervid endogenous gammaretrovirus

CrXRV: cervid exogenous gammaretrovirus

HERV-K: human endogenous retrovirus type K

RACA: reference-assisted chromosome assembly

WGS: whole genome sequencing

NAHR: non-allelic homologous recombination

KYA: thousand years ago

MYA: million years ago

RT: reverse transcriptase

ORF: open reading frame

IAP: intracisternal A-type particle

Author's contributions

LY, RM, RC, TK, JR, PM, and MP conducted analyses; LY, RM, DE, MP interpreted data; LY and MP wrote the manuscript.

Acknowledgements

The authors would like to thank David Chen for contributions to genome analysis.

Declarations

Funding

This work was funded in part by USGS 06HQAG0131. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors claim no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Coffin JM. Retroviridae and their replication. Fields Virol. Lippincott-Raven; 1996. p. 1767–1848.

2. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006;3:67.

3. Löwer R, Löwer J, Kurth R. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A. 1996;93:5177–84.

4. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 2012;10:395–406.

5. Belshaw R, Katzourakis A, Pac□es J, Burt A, Tristem M. High Copy Number in Human Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to Reinfection. Mol Biol Evol. 2005;22:814–7.

6. Johnson WE. Endogenous Retroviruses in the Genomics Era. Annu Rev Virol. 2015;2:135–59.

7. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, et al. Long-term

reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A. 2004;101:4894–9.

8. Boeke JD, Stoye JP. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY); 1997;343–435.

9. Feschotte C, Gilbert C. Endogenous viruses: Insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13:283–96.

10. Jern P, Coffin JM. Effects of retroviruses on host genome function. Annu Rev Genet. 2008;42:709–32.

11. Kurth R, Bannert N. Beneficial and detrimental effects of human endogenous retroviruses. Int. J. Cancer. 2010. p. 306–14.

12. Antony JM, DesLauriers AM, Bhat RK, Ellestad KK, Power C. Human endogenous retroviruses and multiple sclerosis: Innocent bystanders or disease determinants? Biochim Biophys Acta - Mol Basis Dis. 2011;1812:162–76.

13. Magiorkinis G, Belshaw R, Katzourakis A. “There and back again”: revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. Philos Trans R Soc Lond B Biol Sci. 2013;368:20120504.

14. Wildschutte JH, Ram D, Subramanian R, Stevens VL, Coffin JM. The distribution of insertionally polymorphic endogenous retroviruses in breast cancer patients and cancer-free controls. Retrovirology. 2014;11:62.

15. Xue B, Sechi LA, Kelvin DJ. Human Endogenous Retrovirus K (HML-2) in Health and Disease. Front Microbiol. 2020;11.

16. Li W, Lee M, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human

endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med. 2015;7:307ra153.

17. Li W, Yang L, Harris RS, Lin L, Olson TL, Hamele CE, et al. Retrovirus insertion site analysis of LGL leukemia patient genomes. BMC Med Genomics. BMC Medical Genomics; 2019;12:88.

18. Roca AL, O'Brien SP, Greenwood AD, Eiden M V., Ishida Y. Transmission, Evolution, and Endogenization: Lessons Learned from Recent Retroviral Invasions. Microbiol Mol Biol Rev. 2017;82:1–41.

19. Elleder D, Kim O, Padhi A, Bankert JG, Simeonov I, Schuster SC, et al. Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. J Virol. 2012;86:2787–96.

20. Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, et al. A Paradigm for Virus–Host Coevolution: Sequential Counter-Adaptations between Endogenous and Exogenous Retroviruses. PLoS Pathog. 2007;3:e170.

21. Coffin JM, Hughes SH, Varmus HE. Retroviral Virions and Genomes--Retroviruses. Cold Spring Harbor Laboratory Press; 1997.

22. Luo GX, Taylor J. Template switching by reverse transcriptase during DNA synthesis. J Virol. 1990;64:4321–8.

23. Bolinger C, Boris-Lawrie K. Mechanisms employed by retroviruses to exploit host factors for translational control of a complicated proteome. Retrovirology. 2009;6:8.

24. Sofuku K, Honda T. Influence of Endogenous Viral Sequences on Gene Expression. Gene Expr Regul Mamm Cells - Transcr From Gen Asp. InTech; 2018.

25. Kim H-S. Genomic impact, chromosomal distribution and transcriptional regulation

of HERV elements. Mol Cells. 2012;33:539–44.

26. Isbel L, Whitelaw E. Endogenous retroviruses in mammals: An emerging picture of how ERVs modify expression of adjacent genes. BioEssays. 2012;34:734–8.

27. Yao S, Sukonnik T, Kean T, Bharadwaj RR, Pasceri P, Ellis J. Retrovirus Silencing, Variegation, Extinction, and Memory Are Controlled by a Dynamic Interplay of Multiple Epigenetic Modifications. Mol Ther. 2004;10:27–36.

28. Hurst TP, Magiorkinis G. Epigenetic control of human endogenous retrovirus expression: Focus on regulation of long-terminal repeats (LTRs). Viruses. 2017;9:1–13.

29. Geis FK, Goff SP. Silencing and Transcriptional Regulation of Endogenous Retroviruses: An Overview. Viruses. 2020;12:884.

30. Lavie L, Kitova M, Maldener E, Meese E, Mayer J. CpG Methylation Directly Regulates Transcriptional Activity of the Human Endogenous Retrovirus Family HERV-K(HML-2). J Virol. 2005;79:876–83.

31. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature. 2010;464:927–31.

32. Bruno M, Mahgoub M, Macfarlan TS. The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. Annu Rev Genet. 2019;53:annurev-genet-112618-043717.

33. Sze A, Olagnier D, Lin R, van Grevenynghe J, Hiscott J. SAMHD1 Host Restriction Factor: A Link with Innate Immune Sensing of Retrovirus Infection. J Mol Biol. 2013;425:4981–94.

34. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus

envelope for host defense in hominid ancestors. Elife. 2017;6.

35. Haig D. Retroviruses and the Placenta. Curr Biol. 2012;22:R609–13.

36. Fu B, Ma H, Liu D. Endogenous Retroviruses Function as Gene Expression Regulatory Elements During Mammalian Pre-implantation Embryo Development. Int J Mol Sci. 2019;20:790.

37. Göke J, Ng HH. CTRL + INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. EMBO Rep. 2016;17:1131–44.

38. Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc Natl Acad Sci U S A. 2004;101:1668–72.

39. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol. 2005;79:12507–14.

40. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. BioMed Central Ltd; 2011;8:90.

41. Hoang ML, Tan FJ, Lai DC, Celniker SE, Hoskins RA, Dunham MJ, et al. Competitive repair by naturally dispersed repetitive DNA during non-allelic homologous recombination. PLoS Genet. 2010;6:e1001228.

42. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, et al. Rate of recombinational deletion among human endogenous retroviruses. J Virol. 2007;81:9437–42.

43. Kijima TE, Innan H. On the Estimation of the Insertion Time of LTR Retrotransposable Elements. Mol Biol Evol. 2010;27:896–904.

44. Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci. 1999;96:10254–60.

45. Zhuo X, Rho M, Feschotte C. Genome-Wide Characterization of Endogenous Retroviruses in the Bat Myotis lucifugus Reveals Recent and Diverse Infections. J Virol. 2013;87:8493–501.

46. Stocking C, Kozak CA. Murine endogenous retroviruses. Cell Mol Life Sci. 2008;65:3383–98.

47. Anai Y, Ochi H, Watanabe S, Nakagawa S, Kawamura M, Gojobori T, et al. Infectious endogenous retroviruses in cats and emergence of recombinant viruses. J Virol. 2012;86:8634–44.

48. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A. 2016;113:E2326-34.

49. Soriano P, Gridley T, Jaenisch R. Retroviruses and insertional mutagenesis in mice: proviral integration at the Mov 34 locus leads to early embryonic death. Genes Dev. 1987;1:366–75.

50. Moyes D, Griffiths DJ, Venables PJ. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet. 2007;23:326–33.

51. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol. 2001;11:1531–5.

52. Li W, Lin L, Malhotra R, Yang L, Acharya R, Poss M. A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. Wilke CO, editor. PLOS Comput Biol. 2019;15:e1006564.

53. Bao L, Elleder D, Malhotra R, DeGiorgio M, Maravegias T, Horvath L, et al. Computational and Statistical Analyses of Insertional Polymorphic Endogenous Retroviruses in a Non-Model Organism. Computation. 2014;2:221–45.

54. Kamath PL, Poss M, Elleder D, Powell JH, Bao L, Cross PC. The Population History of Endogenous Retroviruses in Mule Deer (Odocoileus hemionus) . J Hered. 2013;105:173–87.

55. Fábryová H, Hron T, Kabíčková H, Poss M, Elleder D. Induction and characterization of a replication competent cervid endogenous gammaretrovirus (CrERV) from mule deer cells. Virology. 2015;485:96–103.

56. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. Nature Publishing Group; 2015;517:608–11.

57. Gregory TR. Animal Genome Size Database. 2019.

58. Vinogradov AE. Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. Cytometry. 1998;31:100–9.

59. Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, et al. Reference-assisted chromosome assembly. Proc Natl Acad Sci U S A. 2013;110:1785–90.

60. Gallagher DS, Derr JN, Womack JE. Chromosome conservation among the advanced pecorans and determination of the primitive bovid karyotype. J Hered. 1994;85:204–10.

61. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.

62. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. BioMed Central Ltd; 2011;12:491.

63. Benit L, Dessen P, Heidmann T. Identification, Phylogeny, and Evolution of Retroviral Elements Based on Their Envelope Genes. J Virol. 2001;75:11709–19.

64. Hunter DR, Bao L, Poss M. Assignment of endogenous retrovirus integration sites using a mixture model. Ann Appl Stat. 2017;11:751–70.

65. Gifford RJ, Katzourakis A, De Ranter J, Magiorkinis G, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci. 2012;109:7385–90.

66. MAGER DL, FREEMAN JD. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. Virology. 1995;213:395–404.

67. Suspène R, Sommer P, Henry M, Ferris S, Guétard D, Pochet S, et al. APOBEC3G is a single-stranded DNA cytidine deaminase and functions independently of HIV reverse transcriptase. Nucleic Acids Res. 2004;32:2421–9.

68. Rebollo R, Romanish MT, Mager DL. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu Rev Genet. 2012;46:21–42.

69. Zheng Y-H, Jeang K-T, Tokunaga K. Host restriction factors in retroviral infection:

promises in virus-host interaction. Retrovirology. 2012;9:112.

70. Duggal NK, Emerman M. Evolutionary conflicts between viruses and restriction factors shape immunity. Nat Rev Immunol. 2012;12:687–95.

71. Daugherty MD, Malik HS. Rules of Engagement: Molecular Insights from Host-Virus Arms Races. Annu Rev Genet. 2012;46:677–700.

72. Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, Hixson P, et al. Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. BMC Biol. 2014;12:74.

73. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. Gene. 2009;448:105–14.

74. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. 2006;2:e2.

75. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol. Nature Publishing Group; 2014;21:423–5.

76. Kawasaki J, Nishigaki K. Tracking the Continuous Evolutionary Processes of an Endogenous Retrovirus of the Domestic Cat: ERV-DC. Viruses. 2018;10:179.

77. Bénit L, De Parseval N, Casella JF, Callebaut I, Cordonnier A, Heidmann T. Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J Virol. 1997;71:5652–7.

78. Finnerty H, Mi S, Veldman GM, McCoy JM, LaVallie E, Edouard P, et al. Syncytin is

a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2002;403:785–9.

79. Murin CD, Wilson IA, Ward AB. Antibody responses to viral infections: a structural perspective across three different enveloped viruses. Nat Microbiol. 2019;4:734–47.

80. Stamatatos L, Morris L, Burton DR, Mascola JR. Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? Nat Med. 2009;15:866–70.

81. Evans LH, Alamgir ASM, Owens N, Weber N, Virtaneva K, Barbian K, et al. Mobilization of Endogenous Retroviruses in Mice after Infection with an Exogenous Retrovirus. J Virol. 2009;83:2429–35.

82. Hanafusa H. Analysis of the defectiveness of rous sarcoma virus III. Determining influence of a new helper virus on the host range and susceptibility to interference of RSV. Virology. 1965;25:248–55.

83. Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. Nat Genet. 2004;36:534–9.

84. Halo J V., Pendleton AL, Jarosz AS, Gifford RJ, Day ML, Kidd JM. Origin and recent expansion of an endogenous gammaretroviral lineage in domestic and wild canids. Retrovirology. 2019;16:6.

85. Ali L, Rizvi T, Mustafa F. Cross- and Co-Packaging of Retroviral RNAs and Their Consequences. Viruses. 2016;8:276.

86. Kozak C. Origins of the Endogenous and Infectious Laboratory Mouse Gammaretroviruses. Viruses. 2014;7:1–26.

87. Bamunusinghe D, Naghashfar Z, Buckler-White A, Plishka R, Baliji S, Liu Q, et al. Sequence Diversity, Intersubgroup Relationships, and Origins of the Mouse Leukemia Gammaretroviruses of Laboratory and Wild Mice. Beemon KL, editor. J Virol. 2016;90:4186–98.

88. Löber U, Hobbs M, Dayaram A, Tsangaras K, Jones K, Alquezar-Planas DE, et al. Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. Proc Natl Acad Sci. 2018;115:8609–14.

89. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.

90. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

91. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.

92. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. BioMed Central Ltd; 2013;14:R36.

93. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Res. 2002;12:656–64.

94. Harris RS. Improved pairwise alignment of genomic DNA. The Pennsylvania State University; 2007.

95. Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, et al. The delayed rise of present-day mammals. Nature. 2007;446:507–12.

96. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

97. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 2005;102:10557–62.

98. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008;25:1253–6.

99. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10:e1003537.

100. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52:696–704.

Figure legends

Figure 1. Diagram of CrERV reconstruction and RACA. (A) Mule deer chromosome fragment reconstruction using syntenic fragments. Gray, green and blue boxes correspond to aligned human, cow and mule deer scaffold respectively. Lighter shades represent regions that can only be aligned between two species. Dashed boxes highlight syntenic fragments where the region is conserved among all three species, which yield a chromosome fragment that orients mule deer scaffolds. (B) Reconstruction of CrERV sequences. CrERV and mule deer scaffolds are shown in bold orange and blue boxes, respectively. Long insert mate pair reads are connected by dotted lines and are colored to indicate whether they derive from the mule deer scaffold or CrERV genome. CrERV genomes were assembled by gathering the broken mate pairs surrounding each CrERV loci as described.

Figure 2. Coalescent phylogeny, *env* structural variation and population frequency of representative full-length non-recombinant CrERVs. Nodes with at least 95% posterior probability support are marked by black dots. The high posterior density for each labeled node is shown in Table S6. Boxes next to CrERV names display the frequency of the CrERVs in the mule deer population with a gray scale (annotated at the top-left corner). Diagrams on the right side depict the lineage-specific structural variations in the CrERV envelope gene. White triangles represent insertions (A, B, C), and white rectangles represent deletions (D and E).

Figure 3. Recombination among CrERVs. Shown is a maximum likelihood phylogeny based on a region spanning a portion of *pol* to 5'*env* (JN592050: 4422-7076). Taxa used are a subset of full-length non-recombinant CrERVs representing the four lineages shown in Figure 2 and CrERVs with a recombinant signature containing a Lineage B *env*. Supported nodes (aLRT >= 0.85) are represented by black dots on the backbone of the tree. Lineage designation is assigned to supported branches based on the non-recombinant CrERV. Over this interval, Lineage B CrERVs are found as a sister group to lineage A CrERV but some CrERV containing a prototypical Lineage B *env* are dispersed among Lineage A CrERV. Note that in this interval lineage C CrERVs cluster with Lineage A CrERVs.

Figure 4. CrERV insertions are enriched within 20 kbp of genes and depleted in introns. We simulated the expected number of CrERV insertions by randomly placing

them on the *de novo* assembled MT273 genome. The proportion of insertions expected within 20kb of a gene from the 10,000 replicates is shown in Panel A. The proportion of intronic insertions is in panel B. The distribution of insertions within 20kb of a gene or an intron from the simulation is shown as a histogram. Blue dashed lines indicate the mean of the simulated data. Red dashed lines indicate the observed data in MT273. Black dashed lines indicate the 5th and 95th percentile of the simulated data, which are used to call significant differences.

Additional files

Figure S1. *K-mer* representation of missing data in assemblies. *60-mers* were generated from the raw paired-end WGS, then ranked and classified into 20 bins containing equal number of *60-mers*. Sorted by the *60-mer* frequency range, bars represent genome repeats from high copy number (left) to low copy number (right). Color bars show percent of all *60-mers* present in the scaffolds/contigs (blue), RACA10K (green), and RACA150K (red) assemblies. *K-mer* counts were based on the total number of *k-mers* (*k-mer* of frequency *n* were counted *n* times). *K-mers* in the raw sequencing data were normalized based on sequencing depth, genome ploidy, read length and *k-mer* length, so that the *k-mer* fractions reflect the proportion of *k-mers* that are present in each assembly compared to the raw sequencing data. All bars start from 0% instead of being stacked. Numbers beneath each bar indicates the range of frequency of *60-mers* in the raw paired-end genome sequencing data in that bin. M: million, K: thousand.

Figure S2. Diagram of CrERV recombination breakpoints. Gray lines point at the key recombination breakpoints on the CrERV. Text box connected to the gray lines indicate the coordinate and adjusted p-value of the breakpoint. Solid gray lines indicate breakpoints of recombinant lineages; dashed gray lines indicate additional breakpoints detected by testing the alignment of reference non-recombinant and candidate recombinant CrERVs. All coordinates are relative to GenBank entry JN592050. Double star (**) indicates breakpoints used in the Lineage B recombinant analysis.

Figure S3. Distribution of simulated mean distance to gene per replicate in mule deer, cow and human genome. Distribution of mule deer, cow and human are colored in red, green and blue respectively. Mann-Whitney U test p-values in all three comparisons are less than 2.2×10^{-16}.

Table S1. de novo and RACA assembly statistics. As the resolution increases, scaffolds can be placed into less chromosome fragments at the expense of less scaffolds incorporated. PE: paired-end sequencing. MP: long insert mate pair sequencing. N50: length of the shortest contig at 50% of the total genome length, a measurement of assembly contiguity. RACA: reference-assisted chromosome assembly. RACA assembly size: total size of RACA chromosome fragments at given RACA resolution.

Table S2. Scaffold assignment to RACA chromosome fragments. The tab "RACA10K" corresponds to the RACA at 10 Kbp resolution, and "RACA150K" corresponds to 150 Kbp resolution. "R150K.R10K.CowChain10KSort" contains the chromosome fragment assignment information of both RACA150K (column A-G) and RACA10K (column H-N), sorted by the scaffolds' alignment chain (column O-V) to the cow genome, with column W indicating the genes that are present on the scaffold. Similarly, "R150K.R10K.SheepChain10KSort" represents scaffold assignments to RACA150K and RACA10K chromosome fragments along the sheep genome.

Table S3. Number of gene structures annotated after each maker annotation. Each column represents a Maker iteration by their order.

Table S4. Summary of *k-mers* with > 50 frequency. Numbers outside of parentheses show the cumulative frequency of *k-mers* with >50 frequency belonging to each repeat family. Given the read length of 100 bp and k-mer size of 60, frequency of >50 in the whole genome sequencing library corresponds to >4 copies in the genome. Percentage in parentheses show the fraction of *k-mers* of the repeat family that are present in each assembly denominated by the total of that family in the raw paired-end sequencing library.

Table S5. Inventory of CrERVs used in phylogenetic analysis. Column A: CrERV names, by scaffold number. Multiples CrERVs on the same scaffold are discerned by additional character after scaffold numbers. Column B: Lineage designation, consistent

with assignments mentioned in text and Figure 2. B1 and B2 are the two lineages that are closely related to lineage B but have a different type of env. Bold font of lineage A indicates the sub-lineage that are significantly closer to genes. Column C: Env status, as illustrated in Figure 2. A, B: insertion; D, E: deletion; C: insertion and deletion; M: missing data for the corresponding insertion or deletion. Column D: Inter-lineage recombination status. Text indicates the recombination partner and section. Empty cell means no recombination was detected. Column E: Assignment to RACA10K chromosome fragment (also refer to Table S2). Parentheses mean that they are not on RACA10K and provisionally assigned by being the closest to sheep or cow aligmnt chain. Column F: Frequency in all 63 animals used by Bao et al. 2014, using >=0.95 probability cutoff. The repetitive junction fragments (multiple mapping) are designated 'uncertain' and their frequency in the probability table was listed in the parenthesis. Bold font indicates the singletons. Column G: Distance to the closest gene. 'NA' means that the virus-containing scaffold cannot be assigned to RACAs and no genes can be found on the CrERV-containing scaffold. Negative value means intronic insertion. Bold font indicates recombinant CrERVs (Figure 3) that are close to genes. Column H: Presence in Figure 2, with letter in parenthesis representing the node it corresponds to.

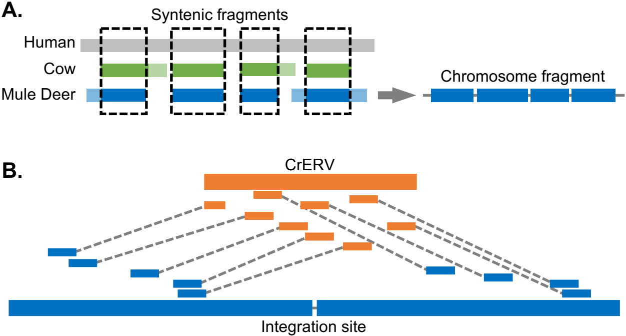
Table S6. Intervals for the 95% highest probability density (HPD) intervals for key nodes in Figure 2.

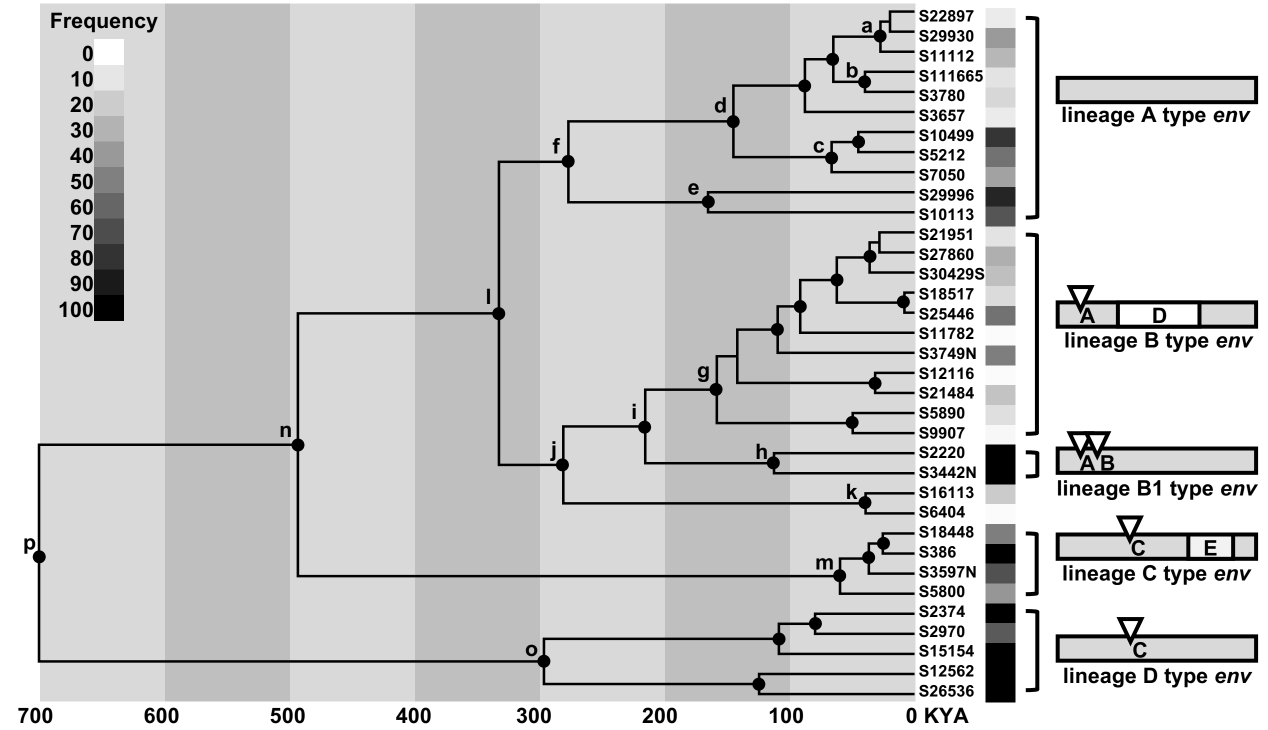
Table S7. Sequences used for recombination breakpoint tests and results of the tests. (A) List of reference non-recombinant CrERVs used for each lineage. (B)

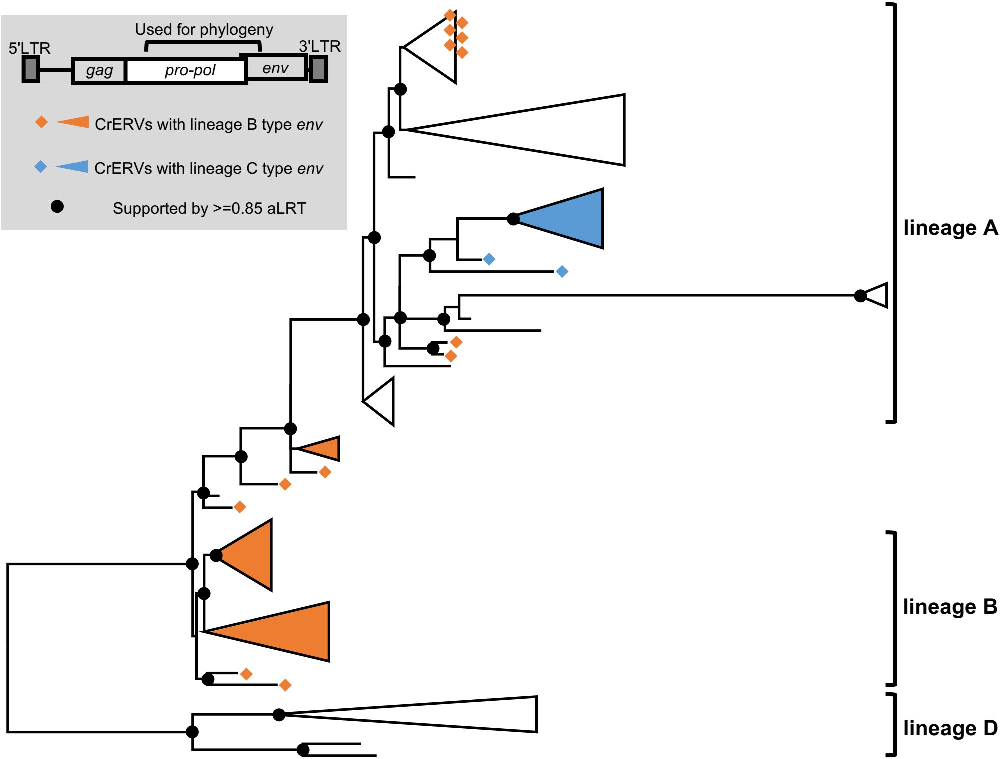
Summary of recombination breakpoint tests. Statistically significant breakpoints (adjusted $p < 0.01$) are shown. Breakpoint coordinates are based on GenBank JN592050.

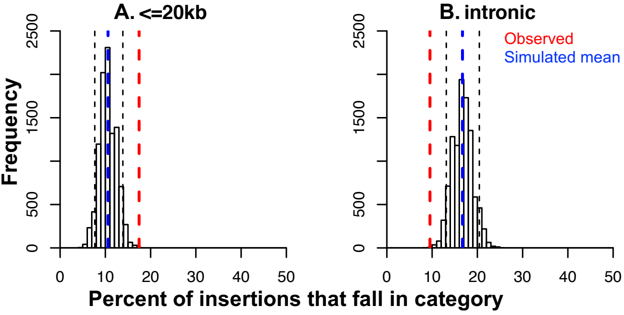
File S1. Supplementary analyses and methods.

File S2. Reconstructed CrERV sequences used in phylogenetic analyses. The file is in fasta format. CrERVs are named after the scaffold they came from, as listed in Table S5.


A.
Syntenic fragments
Human
Cow
Mule Deer
Chromosome fragment
B.
CrERV
Integration site


Frequency
0
10
20
30
40
50
60
70
80
90
100
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
S22897
S29930
S11112
S111665
S3780
S3657
S10499
S5212
S7050
S29996
S10113
S21951
S27860
S30429S
S18517
S25446
S11782
S3749N
S12116
S21484
S5890
S9907
S2220
S3442N
S16113
S6404
S18448
S386
S3597N
S5800
S2374
S2970
S15154
S12562
S26536
700
600
500
400
300
200
100
0 KYA
lineage A type *env*
A
D
lineage B type *env*
A
B
lineage B1 type *env*
C
E
lineage C type *env*
C
lineage D type *env*


Used for phylogeny
5'LTR
3'LTR
gag
pro-pol
env
CrERVs with lineage B type env
CrERVs with lineage C type env
Supported by >=0.85 aLRT
lineage A
lineage B
lineage D


A. <=20kb
B. intronic
Observed
Simulated mean
Frequency
2500
1500
500
0
2500
1500
500
0
0
10
20
30
40
50
0
10
20
30
40
50
Percent of insertions that fall in category