Recombination marks the evolutionary dynamics of a recently endogenized retrovirus

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

All vertebrate genomes have been colonized by retroviruses along their evolutionary trajectory. While endogenous retroviruses (ERVs) can contribute important physiological functions to contemporary hosts, such benefits are attributed to long-term co-evolution of ERV and host because germline infections are rare and expansion is slow, 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. Because we have locus-specific data on the distribution of cervid endogenous retrovirus (CrERV) in populations of mule deer, in this study we determine the molecular evolutionary processes acting on CrERV at each locus in the context of phylogenetic origin, genome location, and population prevalence. A mule deer genome was de novo assembled from short and long insert mate pair reads and CrERV sequence generated at each locus. CrERV composition and diversity have recently measurably increased by horizontal acquisition of a new retrovirus lineage. This new lineage has further expanded CrERV burden and CrERV genomic diversity by activating and recombining with existing CrERV. Resulting inter-lineage recombinants 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. Thus, in mule deer, retroviral colonization is a dynamic period in the molecular evolution of CrERV that also provides a burst of genomic diversity to the host population.

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

Introduction

Retroviruses are unique among viruses in adopting life history strategies that enable them to exist independently as an infectious RNA virus (exogenous retrovirus, XRV) (Coffin 1996) or as an integral component of their host germline (endogenous retrovirus, ERV) (Löwer et al. 1996; Weiss 2006). 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 (Stoye 2012). 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 (Boeke and Stoye 1997; Belshaw et al. 2004; Belshaw, Katzourakis, et al. 2005; Johnson 2015), 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 (Jern and Coffin 2008; Kurth and Bannert 2010; Feschotte and Gilbert 2012), 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 (Antony et al. 2011; Magiorkinis et al. 2013; Wildschutte et al. 2014; Li et al. 2015; Li, Yang, et al. 2019; Xue et al. 2020). Fortunately, there is now evidence that retrovirus colonization is occurring in contemporary, albeit often non-model, species (Arnaud et al. 2007; Elleder et al. 2012; Roca et al. 2017), 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 (Coffin et al. 1997). 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 (Luo and Taylor 1990); 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 (Jern and Coffin 2008; Bolinger and Boris-Lawrie 2009). ERVs require host transcription factors and RNA

103 polymerases to bind to the retrovirus promoter, called long terminal repeats (LTRs), to produce
104 viral transcripts and the RNA genome. Thus, the viral LTRs compete with host genes for
105 transcription factors and polymerases (Sofuku and Honda 2018). A retrovirus encodes at a
106 minimum, genes for the capsid, viral enzymes, and an envelope gene needed for cell entry,
107 which is produced by a sub-genomic mRNA. Hence an ERV also utilizes host-splicing
108 machinery and can alter host gene expression pattern if the site of integration is intronic (Isbel
109 and Whitelaw 2012; Kim 2012). While XRVs are expressed from small numbers of somatic
110 cells, ERVs are present in all cells and ERV transcripts and proteins can be expressed in any
111 cell type at any stage of host development. Hosts actively silence the expression of full or
112 partial ERV sequences by epigenetic methods (Yao et al. 2004; Hurst and Magiorkinis 2017)
113 and by genes called viral restriction factors (Lavie et al. 2005; Matsui et al. 2010; Sze et al.
114 2013; Bruno et al. 2019; Geis and Goff 2020). Because there will be no record of an ERV that
115 causes reproductive failure of the newly colonized host, ERVs in contemporary vertebrates are
116 either effectively controlled by host actions, are nearly neutral in effects on host fitness, or
117 potentially contribute to the overall fitness of the host (Haig 2012; Göke and Ng 2016; Blanco-
118 Melo et al. 2017; Fu et al. 2019).
119
120 The coding portion of a new ERV can be eliminated from the genome through non-allelic
121 homologous recombination (NAHR) between the LTRs, which are identical regions that flank the
122 viral coding portion. A single LTR is left at the site of integration as a consequence of the
123 recombination event and serves as a marker of the original retrovirus integration site (Hughes
124 and Coffin 2004). Most ERV integration sites in humans are solo LTRs (Belshaw, Dawson, et al.
125 2005; Subramanian et al. 2011). Because the efficiency of NAHR is highest between identical
126 sequences (Hoang et al. 2010), conversion of a full-length ERV to a solo LTR likely arises early
127 during ERV residency in the genome before sequence identity of the LTR is lost as mutations
128 accrue (Belshaw et al. 2007). Because mutations are reported to arise in ERVs at the neutral
129 mutation rate of the host (Kijima and Innan 2010), sequence differences between the 5' and 3'
130 LTR of an ERV have been used to approximate the date of integration (Johnson and Coffin
131 1999; Zhuo et al. 2013).
132
133 Although in humans most ERV colonization events occurred in ancestral species, acquisition of
134 new retroviral elements is an ongoing (Stocking and Kozak 2008; Anai et al. 2012) or
135 contemporary (Roca et al. 2017) event in several animal species. The consequences of a recent
136 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 (Soriano et al. 1987; Turner et al. 2001; Moyes et al. 2007; Wildschutte et al. 2016) and HERV-K prevalence at polymorphic sites differ among global populations (Li, Lin, et al. 2019). 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 (Bao et al. 2014) and preliminary data on CrERV sequence variation and colonization history (Elleder et al. 2012; Kamath et al. 2013). 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 (Bao et al. 2014). 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*) (Elleder et al. 2012), 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 (Kamath et al. 2013). 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 (Fábryová et al. 2015). 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

171 We developed a draft assembly of a mule deer genome from animal MT273 in order to
172 determine the sequence at each CrERV locus for phylogenetic analyses and to investigate the
173 effect of CrERV lineage or age on integration site preference. ERV sequences are available in
174 any genome sequencing data because a retrovirus integrates a DNA copy into the host
175 genome. However, there is extensive homology among the most recently integrated ERVs
176 making them difficult to assemble and causing scaffolds to break at the site of an ERV insertion
177 (Chaisson et al. 2015). We assembled scaffolds using a combination of high coverage Illumina
178 short read whole genome sequencing (WGS) and long insert mate pair sequencing. Our *de*
179 *novo* assembly yielded an ~3.31 Gbp draft genome with an N50 of 156 Kbp (Table S1), which is
180 comparable to the 3.33 Gbp (c-value of 3.41 pg) experimentally-determined genome size of
181 reindeer (*Rangifer tarandus*) (Vinogradov 1998; Gregory 2019).
182
183 Approximately half of CrERV loci are located at the ends of scaffolds based on mapping our
184 previously published junction fragment sequences (Bao et al. 2014), which is consistent with the
185 fact that repetitive elements such as ERVs break scaffolds. To determine the sequence of these
186 CrERVs and the genome context in which they are found, we developed a higher order
187 assembly using reference assisted chromosome assembly (RACA) (J. Kim et al. 2013). RACA
188 further scaffolds our *de novo* mule deer assembly into 'chromosome fragments' by identifying
189 synteny blocks among the mule deer scaffolds, the reference species genome (cow), and the
190 outgroup genome (human) (Figure 1A). We created a series of RACA assemblies based on
191 scaffold length to make efficient use of all data (Table S1). RACA150K takes all scaffolds
192 greater than 150,000 bp as input and yielded 41 chromosome fragments, 35 of which are
193 greater than 1.5 Mbp; this is consistent with the known mule deer karyotype of 2n=70
194 (Gallagher et al. 1994). However, RACA150K only incorporates 48% of the total assembled
195 sequences (1.59 Gbp) because of the scaffold size constraint. In contrast, RACA10K uses all
196 scaffolds 10,000 bp or longer and increases the assembly size to 2.37 Gbp (~72% of total
197 assembly) but contains 658 chromosome fragments (Table S1). The majority of scaffolds that
198 cannot be incorporated into a RACA assembly are close to the ends of alignment chains (File
199 S1, section 1a). Most sequences not represented in any assemblies were repeats based on *k-*
200 *mer* analyses (File S1, section 1a and Figure S1).
201
202 Some scaffolds were excluded from the RACA assemblies, presumably because there is no
203 synteny between cow and human for these sequences. We oriented these scaffolds using the
204 cow-mule deer and sheep-mule deer alignments (RACA+, Table S2). Approximately 124 Mbp of

205 sequence (~4% of total assembly) is in scaffolds larger than 10kb but cannot be placed in
206 RACA10K, nearly all of which can be found on the mule deer-cow alignment chain and the mule
207 deer-sheep (oviAri3) alignment chain (123 Mbp in each chain). Because there is overlap
208 between these alignments, only ~1 Mbp is specific to cow and ~1 Mbp is specific to sheep.
209 Therefore, RACA+ incorporated all but 69 scaffolds that are greater than 10 kbp, which
210 consisted of 1.17 Mbp of sequence (~0.04% of total scaffold size of the assembly) and yields an
211 assembly size of 2.49 Gbp (Table S1).
212
213 To enable the investigation of CrERV integration site preference relative to host genes, we
214 annotated the mule deer scaffolds. We used Maker2 (Cantarel et al. 2008; Holt and Yandell
215 2011) for the annotation, which detects candidate genes based on RNA sequencing data and
216 protein homology to any of the three reference genomes: human, cow and sheep. After four
217 Maker iterations, 21,598 genes with an AED (annotation edit distance) (Cantarel et al. 2008) of
218 less than 0.8 were annotated (Table S3). Approximately 92% of genes are found on RACA150K
219 scaffolds and 95% of genes are represented in RACA10K scaffolds.
220
221 Establishing the location and sequence at CrERV loci
222 Several lines of evidence suggest that most CrERVs are missing from the assemblies. Only
223 three CrERVs with coding potential were assembled by the *de novo* assembly. The *k-mer* based
224 analysis shows that less than 9.62% of all LTR repeat elements are in the assemblies (Table
225 S4). The CrERV-host junction fragments previously sequenced (Bao et al. 2014) support that
226 CrERV loci are near scaffold ends or in long stretches of 'N's. Therefore, we took advantage of
227 the different chromosome fragments generated by RACA10K, RACA150K and RACA+ and the
228 long insert mate pair sequencing data to reconstruct CrERVs at each locus (Figure 1B). We
229 identified 252 CrERV loci in the MT273 genome, which is consistent with our estimates of an
230 average of 240 CrERV loci per mule deer by quantitative PCR (Elleder et al. 2012) and 262
231 CrERV loci in animal MT273 by junction fragment analysis (Bao et al. 2014). The majority of
232 CrERV loci (206/252) contains CrERVs with some coding capacity and 46 are solo LTRs. Of the
233 206 CrERVs containing genes, 164 were sufficiently complete to allow phylogenetic analysis on
234 the entire genome or, if a deletion was present, on a subset of viral genes; at 42 loci we were
235 unable to obtain sufficient lengths of high-quality data for further analyses.
236
237 Evolutionary history of CrERV

238 We previously showed that mule deer genomes have been colonized multiple times since the
239 ancestral split with white tailed deer approximately one million years ago (MYA) (Kamath et al.
240 2013) because none of the CrERV integration sites are found in white-tailed deer. To better
241 resolve the colonization history, we conducted a coalescent analysis based on an alignment
242 spanning position 1,477-8,633 bp (omitting a portion of *env*) of the CrERV genome (GenBank:
243 JN592050) using 34 reconstructed CrERV sequences with high quality data that had no
244 signature of recombination and that were representative of the phylogeny in a larger data set
245 (Figure 2). The majority of the *env* gene, which has distinct variable and conserved region (Benit
246 et al. 2001), was manually blocked because of alignment difficulties (6,923-7,503 bp by
247 JN592050 coordinates; see Figure 2, right panel for diagram of *env* variable regions and Table
248 S5, column C for *env* structure of each CrERV). This tree shows four well-supported CrERV
249 lineages, each diverged from a common ancestor at several points since the split of mule deer
250 and white-tailed deer. Although *env* sequence is not included in the phylogenetic analysis,
251 CrERV assigned to each of the four identified lineages share the same distinct *env* variable
252 region structure of insertions and deletions, which define the receptor-binding domain of the
253 envelope protein (Figure 2, right panel).
254
255 Lineage A CrERVs are the youngest ERV family in mule deer. Our estimates indicate that
256 Lineage A colonization has occurred over the last 300 thousand years to the present (Figure 2;
257 Table S6, node f, 95% high posterior density (HPD) interval 110-470 thousand years ago (KYA))
258 and is represented by three well-supported CrERV subgroups evolving over this time frame. All
259 have a complete open reading frame (ORF) in *env* and likely represent a recent retrovirus
260 epizootic. An infectious virus recovered by co-culture belongs to this lineage (Elleder et al.
261 2012). Lineage A represents 30% of all CrERV sampled from MT273 (Table S5). Our age
262 estimates for each subgroup of Lineage A CrERV are consistent with their prevalence in
263 populations of mule deer in the Northern Rocky Mountain ecosystem (Figure 2); (Hunter et al.
264 2017). For example, S29996 and S10113 are estimated to derive from an older Lineage A
265 CrERV subgroup and occur in our sampled mule deer at higher prevalence than those
266 estimated to have entered the genome more recently (see S22897 and S111665, Figure 2).
267
268 Lineage B CrERV shared a common ancestor with Lineage A approximately 300 KYA (node i,
269 Figure 2). Lineage B CrERVs have a short insertion in the 5' portion of *env* followed by a
270 deletion that removes most of the *env* surface unit (SU) relative to Lineage A *env*. Because our
271 coalescent analysis does not include *env* sequence, these results suggest that two

272 phylogenetically distinct XRV with different envelope proteins were circulating about the same
273 time in mule deer populations. Lineage B CrERV represent 32% of sampled viruses from our
274 sequenced genome (Table S5). Like Lineage A, the prevalence of CrERV from Lineage B
275 among mule deer in the northern Rockies region is low, reflecting their more recent colonization
276 of the mule deer genome. Indeed, six Lineage B CrERVs were identified only in MT273, while
277 only one Lineage A CrERV is found only in MT273 (Table S5), which could be indicative of a
278 recent expansion of some Lineage B CrERV. Of note, there are two related groups of CrERV
279 affiliated with Lineage B (Lineage B1 and B2, Table S5). One shares the short 5' insertion in *env*
280 but has a full-length *env* with an additional short insertion relative to the *env* of Lineage A
281 CrERV (Lineage B1, Figure 2). CrERV with this *env* configuration represent 9% of coding
282 CrERV in MT273. Because the prevalence of Lineage B1 is high in mule deer, this group could
283 represent the ancestral state for Lineage B CrERVs. The second group has a unique *env* not
284 found in any other CrERV lineages (Lineage B2, Figure 2, node k; S16113 and S6404). We are
285 unable to estimate the prevalence of this unusual *env* containing CrERV in mule deer because
286 the host junction fragments are not represented in our draft mule deer assembly. It is possible
287 that these viruses represent a cross-species infection and it would be interesting to determine if
288 representatives of Lineage B2 are found in the genomes of other species that occupied the
289 ecosystem in the past.
290
291 Our coalescent estimates indicate that Lineage C CrERV emerged about 500 KYA (Table S6).
292 Several members of this lineage are found in all mule deer sampled (Figure 2; Table S5),
293 consistent with a longer residence in the genome. There is a 59 bp insertion (C) and 362 bp
294 deletion (E) in *env* (Figure 2; Table S5) compared to the full length *env* of Lineage A; none have
295 an intact *env* ORF. Despite evidence that Lineage C is an older CrERV, the approximately 13%
296 of identified CrERV in MT273 belonging to this lineage share a common ancestor ~50 KYA
297 (95% HPD: 16-116 KYA, Table S6). These data are consistent with a recent expansion of a
298 long-term resident CrERV.
299
300 The first representatives of the CrERV family still identifiable in mule deer colonized shortly after
301 their split from white-tailed deer, approximately one MYA (Elleder et al. 2012). Lineage D
302 CrERVs comprise 12% of reconstructed CrERV in MT273 and appear to be near fixation.
303 Indeed, all mule deer in a larger survey of over 250 deer had CrERV S26536, which is not found
304 in white-tailed deer (Kamath et al. 2013). This lineage shares an *env* insertion with Lineage C
305 but lacks the deletion, which removes the transmembrane region of *env*.

306
307 These data expand our previous findings that over the approximately one million year history of
308 mule deer, the mule deer genome has been colonized at least four times by phylogenetically
309 distinct CrERVs; this likely reflects several retroviral epizootics each characterized by a unique
310 *env* structure. The two lineages responsible for most recent endogenization events comprise
311 62% of sampled CrERV. In addition, these data capture the evolutionary processes acting on
312 the *env* gene of exogenous retroviruses, which are characterized by gain or loss of variable
313 regions of this important viral protein.
314
315 Recombination among CrERV lineages
316 Our coalescent estimates (Figure 2) indicate that two phylogenetically distinct CrERV lineages
317 have been expanding in contemporary mule deer genomes over the last 100,000 years. Both
318 lineages have been actively colonizing contemporary mule deer genomes based on divergence
319 estimates, which include zero. While CrERVs represented by Lineage A are capable of infection
320 (Elleder et al. 2012), all Lineage B CrERVs have an identical deletion of the SU portion of *env*
321 and should not be able to spread by reinfecting germ cells. However, the mule deer genome is
322 comprised of approximately equal percentages of Lineage B and Lineage A CrERVs so we
323 considered two modes by which defective Lineage B CrERVs could expand in the genome at a
324 similar rate with Lineage A. Firstly, ERVs that have lost *env* are proposed to preferentially
325 expand by retrotransposition (Gifford et al. 2012) because a functional envelope is not
326 necessary for intracellular replication. Secondly, we consider that Lineage B CrERVs could
327 increase in the genome by infection if the co-circulating Lineage A group provided a functional
328 envelope protein, a process called complementation (MAGER and FREEMAN 1995; Belshaw,
329 Katzourakis, et al. 2005). This latter mechanism requires that a member of each CrERV lineage
330 be transcriptionally active at the same time in the same cell, and that intact proteins from the
331 'helper' genome be used to assemble a particle with a functional envelope for reinfection. If two
332 different CrERV loci are expressed in the same cell, both genomes could be co-packaged in the
333 particle. Because the reverse transcriptase moves between the two RNA genomes as first
334 strand DNA synthesis proceeds, evidence of inter-lineage recombination would support that the
335 molecular components necessary for complementation were in place. We assessed Lineage B
336 CrERV for recombination with Lineage A to determine if coincident expression of the RNA
337 genomes of these two lineages could explain the expansion by infection through
338 complementation of the *env*-less Lineage B CrERV.
339

340 There is good support for recombination between Lineage A and B in a region spanning a
341 portion of *pol* to the beginning of the variable region in *env* (4,422-7,076 based on coordinates
342 of JN592050). In this region, several CrERV that we provisionally classified as Lineage B
343 because they carried the prototypical *env* deletion of SU form a monophyletic group that is
344 affiliated with Lineage A CrERV (Figure 3, upper collapsed clade containing orange diamonds).
345 These Lineage B recombinants all share the same recombination breakpoint just 5' of the
346 characteristic short insertion for these viruses (Figure S2, indicated by "**"; Table S7). In
347 addition, several other CrERVs with Lineage B *env* branch between lineages A and B, indicating
348 that the recombination breakpoints fall within the region assessed (Figure S2). Indeed, the
349 breakpoint in a group of three CrERV is at position 6630 based on coordinates of JN592050,
350 which is near the predicted splice site for *env* at position 6591 (Elleder et al. 2012); this confers
351 an additional 500 bp of the Lineage B *env* on these viruses (Figure S2) resulting in their
352 observed phylogenetic placement. Because recombination between the two retroviral RNA
353 genomes occurs during reverse transcription, our data indicate that both Lineage A and B
354 CrERVs were expressed and assembled in a particle containing a copy of each genome. A
355 functional envelope from Lineage A would therefore have been available for infection. These
356 data support our premise that complementation with a replication competent Lineage A CrERV
357 or CrXRV (cervid exogenous gammaretrovirus, an exogenous version of CrERV) contributes to
358 the 32% prevalence of *env*-deleted Lineage B CrERV in the genome. It is notable that recent
359 retrotransposition of the lineage A-B recombinant CrERVs likely occurred because they are
360 nearly identical and the branches supporting them are short (Figure 3, orange diamonds in the
361 Lineage A type *env* cluster).
362
363 There is additional data to support the transcriptional activity of a Lineage B CrERV, which is
364 requisite for recombination with an infectious Lineage A CrERV or for retrotransposition. We
365 identified a non-recombinant Lineage B CrERV (S24870 in Table S5) with extensive G to A
366 changes (184 changes) compared to other members of this monophyletic group. These data are
367 indicative of a cytidine deaminase acting on the single stranded DNA produced during reverse
368 transcription (Suspène et al. 2004).
369
370 Lineage C CrERV are enigmatic because based on full length sequences lacking a signature of
371 recombination it diverged around 500KYA (Figure 2) but all extant members of this group
372 diverged recently. From Figure 3, it is evident that over the region of *pol* assessed, CrERVs
373 containing the Lineage C *env* cluster with an older Lineage A subgroup. Given that the *env* of

374 Lineage C CrERV shares sequence homology and an insertion with that of the oldest Lineage
375 D, it is likely that Lineage C is in fact the result of recombination between an early member of
376 Lineage A and a relative of a Lineage D CrERV. Many, but not all, Lineage C CrERVs are found
377 at high prevalence in the mule deer population (Figure 2; Table S5), supporting that the initial
378 recombination event occurred early during the Lineage A colonization. Our identification of
379 Lineage C as derived from a non-recombinant CrXRV is therefore incorrect. Instead, Lineage C
380 CrERVs are derived from a CrERV or CrXRV that is not currently represented in mule deer
381 genomes either because it was lost or it never endogenized. Fourteen of the twenty-two CrERV
382 in Lineage C have multiple signatures of recombination predominantly with Lineage A CrERV.
383 The expansion of a subset of Lineage C as a monophyletic group approximately 50 KYA (Figure
384 2; Table S6) suggests that like some members of Lineage B, CrERVs generated by
385 recombination with Lineage A have recently retrotransposed.
386
387 Genomic distribution of CrERV lineages
388 Of the 164 CrERV that we reconstructed from MT273, only 12 can be detected in all mule deer
389 that we have sampled (Kamath et al. 2013; Bao et al. 2014) (Table S5). This means that the
390 majority of CrERV loci in mule deer are insertionally polymorphic; not all animals will have a
391 CrERV occupying a given locus. ERVs can impact genome function in multiple ways but the
392 best documented is by altering host gene regulation, which occurs if the integration site is near
393 a host gene (Rebollo et al. 2012). Thus, we investigated the spatial distribution of CrERV loci
394 relative to host genes to determine the potential of either fixed or polymorphic CrERV to impact
395 gene expression, which could affect host phenotype.
396
397 The actual distance between genes is likely to be unreliable in our assembly because most high
398 copy number repeats are missing in the mule deer assembly (Figure S1, Table S4, section 1a of
399 File S1). To investigate potential problems determining the spatial distribution of CrERV
400 insertions imposed by using a draft assembly, we simulated the distribution of retrovirus
401 insertions (File S1, section 2l) in mule deer (scaffold N50=156 Kbp) and the genomes of cow
402 (Btau7, scaffold N50=2.60 Mbp) and human (hg19, scaffold N50=46.4 Mbp). The mean distance
403 between insertion and the closest gene for all simulation replicates (Figure S3) is significantly
404 higher in the cow and human (Mann-Whitney U test $p < 2.2\text{x}10^{-16}$ for any pair of comparison
405 among the three species). Therefore, we determined if the number of CrERV loci observed to
406 be within 20Kbp of a gene differed from that expected if the distribution was random. There are
407 significantly more observed insertions that fall within 20 Kbp of the translation start site of a

408 gene than occur randomly (Figure 4A). In contrast, intronic CrERV insertions are significantly
409 less than expected based on our simulations (Figure 4B). Among Lineage A CrERVs, only a
410 single sub-lineage (CrERVs that are associated to node 'a' in Figure 2) are found in closer
411 proximity to genes (bold font in Column G of Table S5) than expected if integrations are random
412 (Fisher's exact test $p = 0.002891$). We also investigated whether any of the recombinant CrERV
413 with a signature of recent expansion was integrated within 20 Kbp of a gene. Two of the three
414 recombinant clusters (Figure 3) contain members that are close to a gene (Table S5, bold font
415 in column G). In particular, Lineage A/B recombinant CrERV S10 is 494 bp from the start of a
416 gene. Remarkably, four Lineage C CrERVs with the typical *env* sequence are within 20 Kbp of a
417 gene (Table S5, bold font in column G). Our data indicate that integration site preference overall
418 favors proximity to genes but that this is not reflected in all lineages. In particular, the history of
419 Lineage C CrERV suggests they could have acquired a different integration site preference
420 through recombination that facilitated recent genome expansion.
421
422 **Discussion**
423 The wealth of data on human ERVs (HERVs) provides the contemporary status of events that
424 initiated early in hominid evolution. Potential impacts of an ERV near the time of colonization on
425 a host population is thought to be minimal because infection of host germ line by an XRV is a
426 rare event and ERVs that affect host fitness are quickly lost. Potentially deleterious ERVs that
427 are not lost due to reproductive failure can be removed by recombination leaving a solo LTR at
428 the integration site or can suffer degradation presumably because there is no benefit to retain
429 function at these loci; most HERVs are represented by these two states. In addition, humans
430 and other vertebrate hosts have invested extensive genomic resources (Feschotte and Gilbert
431 2012; Stoye 2012; Zheng et al. 2012) to control the expression of ERVs that are maintained.
432 The dynamics between host and ERV are described as an evolutionary arms race (Daugherty
433 and Malik 2012; Duggal and Emerman 2012). This narrative may underrepresent any
434 contributions of ERVs to fitness as they were establishing in a newly colonized host population.
435 Because there are now several species identified to be at different points along the evolutionary
436 scale initiated by the horizontal acquisition of retroviral DNA it is possible to investigate
437 dynamics of ERV that are not yet fixed in a contemporary species. Considering the numerous
438 mechanisms by which newly integrated retroviral DNA affect host biology, such as by
439 introducing new hotspots for recombination (Campbell et al. 2014), altering host gene regulation
440 (Maksakova et al. 2006; Cohen et al. 2009; Rebollo et al. 2012), and providing retroviral
441 transcripts and proteins for host exaptation (Bénit et al. 1997; Finnerty et al. 2002; Lu et al.

442 2014; Kawasaki and Nishigaki 2018), colonizing ERVs could make a substantive contribution to
443 species' evolution. Our research on the evolutionary dynamics of mule deer CrERV
444 demonstrates that genomic CrERV content and diversity increased significantly during a recent
445 retroviral epizootic due to acquisition of new XRV and from endogenization and
446 retrotransposition of recombinants generated between recent and older CrERVs. These data
447 suggest that CrERV provide a pulse of genetic diversity, which could impact this species'
448 evolutionary trajectory.
449
450 Our analyses of CrERV dynamics in mule deer are based on the sequence of a majority of
451 coding CrERVs in MT273. Of the 252 CrERV loci identified in the MT273 assembly, we were
452 able to reconstruct CrERV sequences from long insert mate pair and Sanger sequencing to use
453 for phylogenetic analysis at 164 sites; 46 sites were solo LTR and 42 were occupied by CrERV
454 retaining some coding capacity. We complimented phylogenetic analyses with our previous data
455 on the frequency of each CrERV locus identified in MT273 in a population of mule deer in the
456 northern Rocky Mountain ecosystem (Bao et al. 2014; Hunter et al. 2017). In addition, we
457 incorporated information on the variable structure of the retroviral envelope gene, *env*, which is
458 characteristic of retrovirus lineages but was excluded from phylogenetic analyses. The variable
459 regions of retroviral *env* result from balancing its role in receptor-mediated, cell specific infection
460 while evading host adaptive immune response (Stamatatos et al. 2009; Murin et al. 2019).
461 Despite excluding most of *env* from our phylogenic analysis because of alignment problems,
462 each of the lineages we identified has a similar distinct *env* structure, as is well known for
463 infectious retroviruses. By integrating population frequency, coalescent estimation, and the
464 unique structural features of *env* we provide an integrated approach to explore the evolutionary
465 dynamics of an endogenizing ERV.
466
467 The most recent CrERV epizootic recorded by germline infection was coincident with the last
468 glacial period, which ended about 12 KYA. The retroviruses that endogenized during this
469 epizootic belong to Lineage A, have open reading frames for all genes and have been
470 recovered by co-culture as infectious viruses (Fábryová et al. 2015). There are several sub-
471 lineages within Lineage A, which likely reflect the evolutionary history of CrXRV contributing to
472 germline infections over this time period. Lineage A retroviruses constitute approximately one
473 third of all retroviral integrations in the genome. Only four of the fifty Lineage A CrERV that we
474 were able to reconstruct did not have a full length *env*. An important implication of this result is
475 that over the most recent approximately 100,000 years of the evolution of this species, the mule

476 deer genome acquired up to half a megabase of new DNA, which introduced new regulatory
477 elements with promoter and enhancer capability, new splice sites, and sites for genome
478 rearrangements. Thus, there is a potential to impact host fitness through altered host gene
479 regulation even if host control mechanisms suppress retroviral gene expression. None of the
480 Lineage A CrERV is fixed in mule deer populations (Table S5, column F) so any effect of CrERV
481 on the host will not be experienced equally in all animals. However, none of the Lineage A
482 CrERV is found only in M273 indicating that the burst of new CrERV DNA acquired during the
483 most recent epizootic has not caused reproductive failure among mule deer. These data
484 demonstrate that in mule deer, a substantial accrual of retroviral DNA in the genome can occur
485 over short time spans in an epizootic and could impose differential fitness in the newly colonized
486 population.
487
488 Lineage A CrERV has an open reading frame for *env* but Lineages B-D do not. Lineage B
489 CrERVs are intriguing in this regard because they also constitute approximately a third of the
490 CrERV in the genome. Yet all have identical deletions of the extracellular portion of *env*, which
491 should render them incapable of genome expansion by reinfection. ERV that have deleted *env*
492 are reportedly better able to expand by retrotransposition (Gifford et al. 2012), which could
493 account for the prevalence of Lineage B. However, because we have evidence for recent
494 expansion of Lineage A and B recombinants, we considered an alternative explanation; that
495 *env*-deficient Lineage B CrERV was complemented with an intact Lineage A CrERV envelope
496 glycoprotein allowing for germline infection. Complementation is not uncommon between XRV
497 and ERV (Hanafusa 1965; Evans et al. 2009), is well established for murine Intracisternal A-
498 type Particle (IAP) (Dewannieux et al. 2004) and has been reported for ERV expansion in
499 canids (Halo et al. 2019). Complementation requires that two different retroviruses are co-
500 expressed in the same cell (Ali et al. 2016). During viral assembly functional genes supplied by
501 either virus are incorporated into the virus particle and either or both retroviral genomes can be
502 packaged. Because the retroviral polymerase uses both strands of RNA during reverse
503 transcription to yield proviral DNA, a recombinant can arise if the two co-packaged RNA strands
504 are not identical. We investigated the possibility of complementation by searching for Lineage A-
505 B recombinants. Our data show that Linage A and B recombination has occurred several times.
506 A group of CrERV that encode a Lineage B *env* cluster with Lineage A CrERV in a phylogeny
507 based on a partial genome alignment (JN592050: 4422-7076bp). The recombinant breakpoint
508 within this monophyletic group is identical, suggesting that the inter-lineage recombinant
509 subsequently expanded by retrotransposition. Notably, two of the CrERV in this recombinant

510 cluster were only found in M273, indicating that retrotransposition was a recent event. There are
511 other clusters of CrERV with Lineage B *env* affiliated with Lineage A CrERV that have different
512 breakpoints in this partial phylogeny. Recombination between an XRV and ERV is also a well-
513 documented property of retroviruses (Kozak 2014; Bamunusinghe et al. 2016; Löber et al.
514 2018). However, the recombinant retroviruses that result are typically identified because they
515 are XRV and often associated with disease or a host switch. Our data indicate that multiple
516 recombination events between Lineage A and B CrERV have been recorded in germline; this in
517 itself is remarkable given that endogenization is a rare event. Thus both the burden of CrERV
518 integrations and the sequence diversity of CrERV in the mule deer genome increase
519 concomitant with a retrovirus epizootic by CrERV inter-lineage recombination.
520
521 Recombination is a dominant feature of CrERV dynamics and is also displayed in the
522 evolutionary history of Lineage C CrERV. Our phylogenetic analysis places the ancestor of
523 Lineage C CrERV at 500 KYA and indeed, Lineage C and Lineage D, which is estimated to be
524 the first CrERV to colonize mule deer after splitting from white-tailed deer (Elleder et al. 2012;
525 Kamath et al. 2013), share many features in *env* that are distinct from those of Lineage A and B.
526 Consistent with a long-term residency in the genome, many Lineage C CrERV are found in most
527 or all mule deer surveyed. A recent expansion of a CrERV that has been quiescent in the
528 genome since endogenizing could explain the estimated 50 KYA time to most recent common
529 ancestor of extant members of this lineage. Although this scenario is consistent with the
530 paradigm that a single XRV colonized the genome and recently expanded by retrotransposition,
531 our analysis shows that all Lineage C CrERV are recombinants of a Lineage A CrERV and a
532 CrERV not recorded in or lost from contemporary mule deer genomes. Hence the resulting
533 monophyletic lineage does not arise from retrotransposition of an ancient colonizing XRV.
534 Rather, as is the case with Lineage B CrERV, recombination between an older CrERV and
535 either a Lineage A CrXRV or CrERV occurred, infected germline, and recently expanded by
536 retrotransposition. It is noteworthy that all retrotransposition events detectable in our data
537 involve recombinant CrERV. Further, recombination often leads to duplications and deletions in
538 the retroviral genome, therefore some of the deletions we document in Lineages B-D are not a
539 consequence of slow degradation in the genome but rather are due to reverse transcription and
540 as was recently reported for Koala retrovirus (Löber et al. 2018).
541
542 These data highlight that expansion of CrERV diversity and genomic burden has occurred in the
543 recent evolutionary history of mule deer by new acquisitions, complementation, and pulses of

544 retrotransposition of inter-lineage recombinants. Indeed, several of the recombinant Lineage C
545 CrERVs that have expanded by retrotransposition are within 20kbp of a gene raising the
546 question as to whether there is a fitness effect at these loci that is in balance with continued
547 expression of the retrovirus. It is remarkable that so many of the events marking the dynamics
548 of retrovirus endogenization are preserved in contemporary mule deer genomes. Given that
549 germline infection is a rare event, it is likely that the dynamics we describe here also resulted in
550 infection of somatic cells. It is worthwhile to consider the potential for ERVs in other species, in
551 particular in humans where several HERVs are expressed, to generate novel antigens through
552 recombination or disruptive somatic integrations that could contribute to disease states.
553
554 **Materials and Methods**
555 Sequencing
556 Whole genome sequencing (WGS) was performed for a male mule deer, MT273, at ~30x depth
557 using the library of ~260 bp insert size, ~10x using the library of ~1,400-5,000 bp insert size and
558 ~30x using the library of ~6,600 bp insert size. 3' CrERV-host junction fragment sequencing was
559 performed as described by Bao *et al.* (Bao et al. 2014). 5' CrERV-host junction fragment
560 sequencing was performed on the Roche 454 platform, with a target size of ~500bp containing
561 up to 380 bp of CrERV LTR.
562
563 Assembly and mapping
564 The draft assembly of MT273 was generated using SOAPdenovo2 (Luo et al. 2012) (File S1,
565 section 2a). WGS data were then mapped back to the assembly using the default setting of bwa
566 mem (Li and Durbin 2009) for further use in RACA and CrERV reconstruction. RNA-seq data
567 was mapped to the WGS scaffolds using the default setting of tophat (Trapnell et al. 2009; D.
568 Kim et al. 2013). 3' junction fragments were clustered as described in Bao *et al.* (Bao et al.
569 2014). 3' junction fragment clusters and 5' junction fragment reads were mapped to the WGS
570 assembly using the default setting of blat (Kent 2002). A perl script was used to filter for the
571 clusters or reads whose host side of the fragment maps to the host at its full length and high
572 identity. 5' junction fragments were then clustered using the default setting of bedtools merge.
573
574 RACA
575 Synteny based scaffolding using RACA was performed based on the genome alignment
576 between the mule deer WGS assembly, a reference genome (cow, bosTau7 or Btau7), and an
577 outgroup genome (hg19). Genome alignments were performed with lastz (Harris 2007) under

578 the setting of '--notransition --step=20', and then processed using the UCSC axtChain and
579 chainNet tools. The mule deer-cow-human phylogeny was derived from Bininda-Emonds *et al*.
580 (Bininda-Emonds et al. 2007) using the 'ape' package of R.
581
582 CrERV sequence reconstruction
583 CrERV locations and sequences were retrieved based on junction fragment and long insert
584 mate pair WGS data. The long insert mate pair WGS reads were mapped to the reference
585 CrERV (GenBank: JN592050) using bwa mem. Mates of reads that mapped to the reference
586 CrERV were extracted and then mapped to the WGS assembly using bwa mem. Mates mapped
587 to the WGS assembly were then clustered using the 'cluster' function of bedtools. Anchoring
588 mate pair clusters on both sides of the insertion site were complemented by junction fragments
589 to localize CrERVs. Based on the RACA data, CrERVs that sit between scaffolds were also
590 retrieved in this manner. CrERV reads were then assigned to their corresponding cluster and
591 were assembled using SeqMan (DNASTAR). Sanger sequencing was performed to
592 complement key regions used in CrERV evolutionary analyses. All reconstructed CrERV
593 sequences used in the phylogenetic analyses are included in File S2 in fasta format.
594
595 CrERV evolution analyses
596 CrERV sequences of interest were initially aligned using the default setting of muscle (Edgar
597 2004), manually trimmed for the region of interest, and then re-aligned using the default setting
598 of Prank (Löytynoja and Goldman 2005). Lineage-specific regions are manually curated to form
599 lineage-specific blocks. Models for phylogeny were selected by AICc (Akaike Information
600 Criterion with correction) using jModelTest (Posada 2008). Coalescent analysis and associated
601 phylogeny (Figure 2) was generated using BEAST2 (Bouckaert et al. 2014). In the coalescent
602 analysis, we used GTR substitution matrix, four Gamma categories, estimated among-site
603 variation, Calibrated Yule tree prior with ucldMean ucldStddev from exponential distribution,
604 relaxed lognormal molecular clock, shared common ancestor of all CrERVs 0.47-1 MYA as a
605 prior (Elleder et al. 2012; Kamath et al. 2013). Maximum likelihood phylogeny in Figure 3 was
606 generated using PhyML (Guindon and Gascuel 2003) using the models selected by AICc and
607 the setting of '-o tlr -s BEST' according to the selected model.
608
609 CrERV spatial distribution
610 We simulated 274 insertions per genome to approximate the average number of CrERVs in a
611 mule deer (Bao et al. 2014). The simulation was performed 10,000 times on three genomes: the

612 mule deer WGS scaffolds, cow (Btau7) and human (hg19). Distance between simulated
613 insertions and the closest start of the coding sequence of a gene was calculated using the
614 'closest' function of bedtools, and the simulated insertions that overlap with a gene were marked
615 with the 'intersect' function of bedtools. Number of simulated simulations that are within 20 Kbp
616 or intronic to a gene was counted for each of the 10,000 replicates. Counts were then
617 normalized by the total number of insertions and plotted using the 'hist' function of R.
618
619 Supplementary methods
620 Methods with extended details are available in File S1.
621
622 **Acknowledgements**
623 This work was supported in part by United States Geological Survey 06HQAG0131. The
624 authors would like to thank David Chen for contributions to genome analysis.
625
626 **Availability of Data and Materials**
627 The raw sequencing data was deposited in NCBI BioProject PRJNA705069. Other data
628 generated are included in supplementary file and figures.
629
630 **Author's contributions**
631 LY, RM, RC, TK, JR, PM, and MP conducted analyses; LY, RM, DE, MP interpreted data; LY
632 and MP wrote the manuscript.
633
634 **References**
635 Ali L, Rizvi T, Mustafa F. 2016. Cross- and Co-Packaging of Retroviral RNAs and Their
636 Consequences. Viruses 8:276.
637 Anai Y, Ochi H, Watanabe S, Nakagawa S, Kawamura M, Gojobori T, Nishigaki K. 2012.
638 Infectious endogenous retroviruses in cats and emergence of recombinant viruses. J. Virol.
639 86:8634–8644.
640 Antony JM, DesLauriers AM, Bhat RK, Ellestad KK, Power C. 2011. Human endogenous
641 retroviruses and multiple sclerosis: Innocent bystanders or disease determinants? Biochim.
642 Biophys. Acta - Mol. Basis Dis. 1812:162–176.
643 Arnaud F, Caporale M, Varela M, et al. 2007. A Paradigm for Virus–Host Coevolution:
644 Sequential Counter-Adaptations between Endogenous and Exogenous Retroviruses. PLoS
645 Pathog. 3:e170.

646 Bamunusinghe D, Naghashfar Z, Buckler-White A, et al. 2016. Sequence Diversity,
647 Intersubgroup Relationships, and Origins of the Mouse Leukemia Gammaretroviruses of
648 Laboratory and Wild Mice.Beemon KL, editor. J. Virol. 90:4186–4198.
649 Bao L, Elleder D, Malhotra R, et al. 2014. Computational and Statistical Analyses of Insertional
650 Polymorphic Endogenous Retroviruses in a Non-Model Organism. Computation 2:221–
651 245.
652 Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. 2005. Genomewide
653 screening reveals high levels of insertional polymorphism in the human endogenous
654 retrovirus family HERV-K(HML2): implications for present-day activity. J. Virol. 79:12507–
655 12514.
656 Belshaw R, Katzourakis A, Pačes J, Burt A, Tristem M. 2005. High Copy Number in Human
657 Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to
658 Reinfection. Mol. Biol. Evol. 22:814–817.
659 Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004. Long-term
660 reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U. S.
661 A. 101:4894–4899.
662 Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M. 2007. Rate
663 of recombinational deletion among human endogenous retroviruses. J. Virol. 81:9437–
664 9442.
665 Benit L, Dessen P, Heidmann T. 2001. Identification, Phylogeny, and Evolution of Retroviral
666 Elements Based on Their Envelope Genes. J. Virol. 75:11709–11719.
667 Bénit L, De Parseval N, Casella JF, Callebaut I, Cordonnier A, Heidmann T. 1997. Cloning of a
668 new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L
669 element and with a gag coding sequence closely related to the Fv1 restriction gene. J.
670 Virol. 71:5652–5657.
671 Bininda-Emonds ORP, Cardillo M, Jones KE, et al. 2007. The delayed rise of present-day
672 mammals. Nature 446:507–512.
673 Blanco-Melo D, Gifford RJ, Bieniasz PD. 2017. Co-option of an endogenous retrovirus envelope
674 for host defense in hominid ancestors. Elife 6.
675 Boeke JD, Stoye JP. 1997. Retrotransposons, endogenous retroviruses, and the evolution of
676 retroelements. :343–435.
677 Bolinger C, Boris-Lawrie K. 2009. Mechanisms employed by retroviruses to exploit host factors
678 for translational control of a complicated proteome. Retrovirology 6:8.
679 Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A,

680 Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis.
681 PLoS Comput. Biol. 10:e1003537.
682 Bruno M, Mahgoub M, Macfarlan TS. 2019. The Arms Race Between KRAB–Zinc Finger
683 Proteins and Endogenous Retroelements and Its Impact on Mammals. Annu. Rev. Genet.
684 53:annurev-genet-112618-043717.
685 Campbell IM, Gambin T, Dittwald P, et al. 2014. Human endogenous retroviral elements
686 promote genome instability via non-allelic homologous recombination. BMC Biol. 12:74.
687 Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A,
688 Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging
689 model organism genomes. Genome Res. 18:188–196.
690 Chaisson MJP, Huddleston J, Dennis MY, et al. 2015. Resolving the complexity of the human
691 genome using single-molecule sequencing. Nature 517:608–611.
692 Coffin JM. 1996. Retroviridae and their replication. In: Fields Virology. Fields Virology.
693 Lippincott-Raven. p. 1767–1848.
694 Coffin JM, Hughes SH, Varmus HE. 1997. Retroviral Virions and Genomes--Retroviruses. Cold
695 Spring Harbor Laboratory Press
696 Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human
697 genes: A critical assessment. Gene 448:105–114.
698 Daugherty MD, Malik HS. 2012. Rules of Engagement: Molecular Insights from Host-Virus Arms
699 Races. Annu. Rev. Genet. 46:677–700.
700 Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T. 2004. Identification of
701 autonomous IAP LTR retrotransposons mobile in mammalian cells. Nat. Genet. 36:534–
702 539.
703 Duggal NK, Emerman M. 2012. Evolutionary conflicts between viruses and restriction factors
704 shape immunity. Nat. Rev. Immunol. 12:687–695.
705 Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high
706 throughput. Nucleic Acids Res. 32:1792–1797.
707 Elleder D, Kim O, Padhi A, Bankert JG, Simeonov I, Schuster SC, Wittekindt NE, Motameny S,
708 Poss M. 2012. Polymorphic integrations of an endogenous gammaretrovirus in the mule
709 deer genome. J. Virol. 86:2787–2796.
710 Evans LH, Alamgir ASM, Owens N, Weber N, Virtaneva K, Barbian K, Babar A, Malik F,
711 Rosenke K. 2009. Mobilization of Endogenous Retroviruses in Mice after Infection with an
712 Exogenous Retrovirus. J. Virol. 83:2429–2435.
713 Fábryová H, Hron T, Kabíčková H, Poss M, Elleder D. 2015. Induction and characterization of a

714 replication competent cervid endogenous gammaretrovirus (CrERV) from mule deer cells.
715 Virology 485:96–103.
716 Feschotte C, Gilbert C. 2012. Endogenous viruses: Insights into viral evolution and impact on
717 host biology. Nat. Rev. Genet. 13:283–296.
718 Finnerty H, Mi S, Veldman GM, et al. 2002. Syncytin is a captive retroviral envelope protein
719 involved in human placental morphogenesis. Nature 403:785–789.
720 Fu B, Ma H, Liu D. 2019. Endogenous Retroviruses Function as Gene Expression Regulatory
721 Elements During Mammalian Pre-implantation Embryo Development. Int. J. Mol. Sci.
722 20:790.
723 Gallagher DS, Derr JN, Womack JE. 1994. Chromosome conservation among the advanced
724 pecorans and determination of the primitive bovid karyotype. J. Hered. 85:204–210.
725 Geis FK, Goff SP. 2020. Silencing and Transcriptional Regulation of Endogenous Retroviruses:
726 An Overview. Viruses 12:884.
727 Gifford RJ, Katzourakis A, De Ranter J, Magiorkinis G, Belshaw R. 2012. Env-less endogenous
728 retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. 109:7385–7390.
729 Göke J, Ng HH. 2016. CTRL + INSERT: retrotransposons and their contribution to regulation
730 and innovation of the transcriptome. EMBO Rep. 17:1131–1144.
731 Gregory TR. 2019. Animal Genome Size Database.
732 Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large
733 phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
734 Haig D. 2012. Retroviruses and the Placenta. Curr. Biol. 22:R609–R613.
735 Halo J V., Pendleton AL, Jarosz AS, Gifford RJ, Day ML, Kidd JM. 2019. Origin and recent
736 expansion of an endogenous gammaretroviral lineage in domestic and wild canids.
737 Retrovirology 16:6.
738 Hanafusa H. 1965. Analysis of the defectiveness of rous sarcoma virus III. Determining
739 influence of a new helper virus on the host range and susceptibility to interference of RSV.
740 Virology 25:248–255.
741 Harris RS. 2007. Improved pairwise alignment of genomic DNA.
742 Hoang ML, Tan FJ, Lai DC, Celniker SE, Hoskins RA, Dunham MJ, Zheng Y, Koshland D.
743 2010. Competitive repair by naturally dispersed repetitive DNA during non-allelic
744 homologous recombination. PLoS Genet. 6:e1001228.
745 Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management
746 tool for second-generation genome projects. BMC Bioinformatics 12:491.
747 Hughes JF, Coffin JM. 2004. Human endogenous retrovirus K solo-LTR formation and

748 insertional polymorphisms: implications for human and viral evolution. Proc. Natl. Acad.
749 Sci. U. S. A. 101:1668–1672.
750 Hunter DR, Bao L, Poss M. 2017. Assignment of endogenous retrovirus integration sites using a
751 mixture model. Ann. Appl. Stat. 11:751–770.
752 Hurst TP, Magiorkinis G. 2017. Epigenetic control of human endogenous retrovirus expression:
753 Focus on regulation of long-terminal repeats (LTRs). Viruses 9:1–13.
754 Isbel L, Whitelaw E. 2012. Endogenous retroviruses in mammals: An emerging picture of how
755 ERVs modify expression of adjacent genes. BioEssays 34:734–738.
756 Jern P, Coffin JM. 2008. Effects of retroviruses on host genome function. Annu. Rev. Genet.
757 42:709–732.
758 Johnson WE. 2015. Endogenous Retroviruses in the Genomics Era. Annu. Rev. Virol. 2:135–
759 159.
760 Johnson WE, Coffin JM. 1999. Constructing primate phylogenies from ancient retrovirus
761 sequences. Proc. Natl. Acad. Sci. 96:10254–10260.
762 Kamath PL, Poss M, Elleder D, Powell JH, Bao L, Cross PC. 2013. The Population History of
763 Endogenous Retroviruses in Mule Deer (Odocoileus hemionus) . J. Hered. 105:173–187.
764 Kawasaki J, Nishigaki K. 2018. Tracking the Continuous Evolutionary Processes of an
765 Endogenous Retrovirus of the Domestic Cat: ERV-DC. Viruses 10:179.
766 Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Res. 12:656–664.
767 Kijima TE, Innan H. 2010. On the Estimation of the Insertion Time of LTR Retrotransposable
768 Elements. Mol. Biol. Evol. 27:896–904.
769 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate
770 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
771 Genome Biol. 14:R36.
772 Kim H-S. 2012. Genomic impact, chromosomal distribution and transcriptional regulation of
773 HERV elements. Mol. Cells 33:539–544.
774 Kim J, Larkin DM, Cai Q, et al. 2013. Reference-assisted chromosome assembly. Proc. Natl.
775 Acad. Sci. U. S. A. 110:1785–1790.
776 Kozak C. 2014. Origins of the Endogenous and Infectious Laboratory Mouse
777 Gammaretroviruses. Viruses 7:1–26.
778 Kurth R, Bannert N. 2010. Beneficial and detrimental effects of human endogenous retroviruses.
779 Int. J. Cancer 126:306–314.
780 Lavie L, Kitova M, Maldener E, Meese E, Mayer J. 2005. CpG Methylation Directly Regulates
781 Transcriptional Activity of the Human Endogenous Retrovirus Family HERV-K(HML-2). J.

782 Virol. 79:876–883.
783 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
784 Bioinformatics 25:1754–1760.
785 Li W, Lee M, Henderson L, et al. 2015. Human endogenous retrovirus-K contributes to motor
786 neuron disease. Sci. Transl. Med. 7:307ra153.
787 Li W, Lin L, Malhotra R, Yang L, Acharya R, Poss M. 2019. A computational framework to
788 assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human
789 populations.Wilke CO, editor. PLOS Comput. Biol. 15:e1006564.
790 Li W, Yang L, Harris RS, Lin L, Olson TL, Hamele CE, Feith DJ, Loughran TP, Poss M. 2019.
791 Retrovirus insertion site analysis of LGL leukemia patient genomes. BMC Med. Genomics
792 12:88.
793 Löber U, Hobbs M, Dayaram A, et al. 2018. Degradation and remobilization of endogenous
794 retroviruses by recombination during the earliest stages of a germ-line invasion. Proc. Natl.
795 Acad. Sci. 115:8609–8614.
796 Löwer R, Löwer J, Kurth R. 1996. The viruses in all of us: characteristics and biological
797 significance of human endogenous retrovirus sequences. Proc. Natl. Acad. Sci. U. S. A.
798 93:5177–5184.
799 Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences
800 with insertions. Proc. Natl. Acad. Sci. U. S. A. 102:10557–10562.
801 Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus
802 HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat.
803 Struct. Mol. Biol. 21:423–425.
804 Luo GX, Taylor J. 1990. Template switching by reverse transcriptase during DNA synthesis. J.
805 Virol. 64:4321–4328.
806 Luo R, Liu B, Xie Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-
807 read de novo assembler. Gigascience 1:18.
808 MAGER DL, FREEMAN JD. 1995. HERV-H Endogenous Retroviruses: Presence in the New
809 World Branch but Amplification in the Old World Primate Lineage. Virology 213:395–404.
810 Magiorkinis G, Belshaw R, Katzourakis A. 2013. “There and back again”: revisiting the
811 pathophysiological roles of human endogenous retroviruses in the post-genomic era.
812 Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368:20120504.
813 Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. 2006.
814 Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS
815 Genet. 2:e2.

816 Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz
817 MC, Shinkai Y. 2010. Proviral silencing in embryonic stem cells requires the histone
818 methyltransferase ESET. Nature 464:927–931.
819 Moyes D, Griffiths DJ, Venables PJ. 2007. Insertional polymorphisms: a new lease of life for
820 endogenous retroviruses in human disease. Trends Genet. 23:326–333.
821 Murin CD, Wilson IA, Ward AB. 2019. Antibody responses to viral infections: a structural
822 perspective across three different enveloped viruses. Nat. Microbiol. 4:734–747.
823 Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.
824 Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural
825 Source of Regulatory Sequences for Host Genes. Annu. Rev. Genet. 46:21–42.
826 Roca AL, O'Brien SP, Greenwood AD, Eiden M V., Ishida Y. 2017. Transmission, Evolution, and
827 Endogenization: Lessons Learned from Recent Retroviral Invasions. Microbiol. Mol. Biol.
828 Rev. 82:1–41.
829 Sofuku K, Honda T. 2018. Influence of Endogenous Viral Sequences on Gene Expression. In:
830 Gene Expression and Regulation in Mammalian Cells - Transcription From General
831 Aspects. InTech.
832 Soriano P, Gridley T, Jaenisch R. 1987. Retroviruses and insertional mutagenesis in mice:
833 proviral integration at the Mov 34 locus leads to early embryonic death. Genes Dev. 1:366–
834 375.
835 Stamatatos L, Morris L, Burton DR, Mascola JR. 2009. Neutralizing antibodies generated during
836 natural HIV-1 infection: good news for an HIV-1 vaccine? Nat. Med. 15:866–870.
837 Stocking C, Kozak CA. 2008. Murine endogenous retroviruses. Cell. Mol. Life Sci. 65:3383–
838 3398.
839 Stoye JP. 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat.
840 Rev. Microbiol. 10:395–406.
841 Subramanian RP, Wildschutte JH, Russo C, Coffin JM. 2011. Identification, characterization,
842 and comparative genomic distribution of the HERV-K (HML-2) group of human
843 endogenous retroviruses. Retrovirology 8:90.
844 Suspène R, Sommer P, Henry M, et al. 2004. APOBEC3G is a single-stranded DNA cytidine
845 deaminase and functions independently of HIV reverse transcriptase. Nucleic Acids Res.
846 32:2421–2429.
847 Sze A, Olagnier D, Lin R, van Grevenynghe J, Hiscott J. 2013. SAMHD1 Host Restriction
848 Factor: A Link with Innate Immune Sensing of Retrovirus Infection. J. Mol. Biol. 425:4981–
849 4994.

850 Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq.
851 Bioinformatics 25:1105–1111.
852 Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. 2001. Insertional
853 polymorphisms of full-length endogenous retroviruses in humans. Curr. Biol. 11:1531–
854 1535.
855 Vinogradov AE. 1998. Genome size and GC-percent in vertebrates as determined by flow
856 cytometry: The triangular relationship. Cytometry 31:100–109.
857 Weiss RA. 2006. The discovery of endogenous retroviruses. Retrovirology 3:67.
858 Wildschutte JH, Ram D, Subramanian R, Stevens VL, Coffin JM. 2014. The distribution of
859 insertionally polymorphic endogenous retroviruses in breast cancer patients and cancer-
860 free controls. Retrovirology 11:62.
861 Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. 2016.
862 Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc.
863 Natl. Acad. Sci. U. S. A. 113:E2326-34.
864 Xue B, Sechi LA, Kelvin DJ. 2020. Human Endogenous Retrovirus K (HML-2) in Health and
865 Disease. Front. Microbiol. 11.
866 Yao S, Sukonnik T, Kean T, Bharadwaj RR, Pasceri P, Ellis J. 2004. Retrovirus Silencing,
867 Variegation, Extinction, and Memory Are Controlled by a Dynamic Interplay of Multiple
868 Epigenetic Modifications. Mol. Ther. 10:27–36.
869 Zheng Y-H, Jeang K-T, Tokunaga K. 2012. Host restriction factors in retroviral infection:
870 promises in virus-host interaction. Retrovirology 9:112.
871 Zhuo X, Rho M, Feschotte C. 2013. Genome-Wide Characterization of Endogenous
872 Retroviruses in the Bat Myotis lucifugus Reveals Recent and Diverse Infections. J. Virol.
873 87:8493–8501.
874
875 **Figure Legends**
876 **Figure 1. Diagram of CrERV reconstruction and RACA.** A. Mule deer chromosome fragment
877 reconstruction using syntenic fragments. Gray, green and blue boxes correspond to aligned
878 human, cow and mule deer scaffold respectively. Lighter shades represent regions that can only
879 be aligned between two species. Dashed boxes highlight syntenic fragments where the region is
880 conserved among all three species, which yield a chromosome fragment that orients mule deer
881 scaffolds. B. Reconstruction of CrERV sequences. CrERV and mule deer scaffolds are shown in
882 bold orange and blue boxes, respectively. Long insert mate pair reads are connected by dotted
883 lines and are colored to indicate whether they derive from the mule deer scaffold or CrERV

884 genome. CrERV genomes were assembled by gathering the broken mate pairs surrounding
885 each CrERV loci as described.
886
887 **Figure 2. Coalescent phylogeny, *env* structural variation and population frequency of**
888 **representative full-length non-recombinant CrERVs.** Nodes with at least 95% posterior
889 probability support are marked by black dots. The high posterior density for each labeled node
890 is shown in Table S6. Boxes next to CrERV names display the frequency of the CrERVs in the
891 mule deer population with a gray scale (annotated at the top-left corner). Diagrams on the right
892 side depict the lineage-specific structural variations in the CrERV envelope gene. White
893 triangles represent insertions (A, B, C), and white rectangles represent deletions (D and E).
894
895 **Figure 3. Recombination among CrERVs.** Shown is a maximum likelihood phylogeny based
896 on a region spanning a portion of *pol* to 5'*env* (JN592050: 4422-7076). Taxa used are a subset
897 of full-length non-recombinant CrERVs representing the four lineages shown in Figure 2 and
898 CrERVs with a recombinant signature containing a Lineage B *env*. Supported nodes (aLRT >=
899 0.85) are represented by black dots on the backbone of the tree. Lineage designation is
900 assigned to supported branches based on the non-recombinant CrERV. Over this interval,
901 Lineage B CrERVs are found as a sister group to lineage A CrERV but some CrERV containing
902 a prototypical Lineage B *env* are dispersed among Lineage A CrERV. Note that in this interval
903 lineage C CrERVs cluster with Lineage A CrERVs.
904
905 **Figure 4. CrERV insertions are enriched within 20 kbp of genes and depleted in introns.**
906 We simulated the expected number of CrERV insertions by randomly placing them on the *de*
907 *novo* assembled MT273 genome. The proportion of insertions expected within 20kb of a gene
908 from the 10,000 replicates is shown in Panel A. The proportion of intronic insertions is in panel
909 B. The distribution of insertions within 20kb of a gene or an intron from the simulation is shown
910 as a histogram. Blue dashed lines indicate the mean of the simulated data. Red dashed lines
911 indicate the observed data in MT273. Black dashed lines indicate the 5th and 95th percentile of
912 the simulated data, which are used to call significant differences.
913
914
915
916
917

Figures

Figure 1

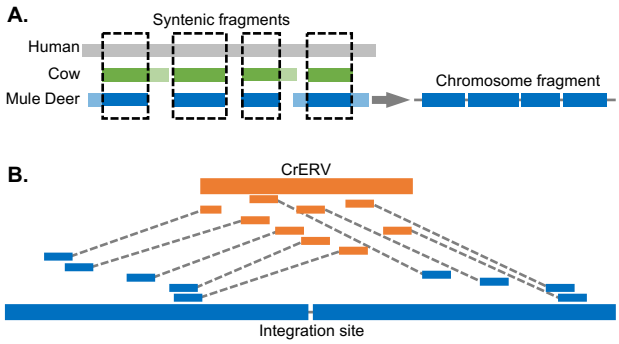

Figure 2

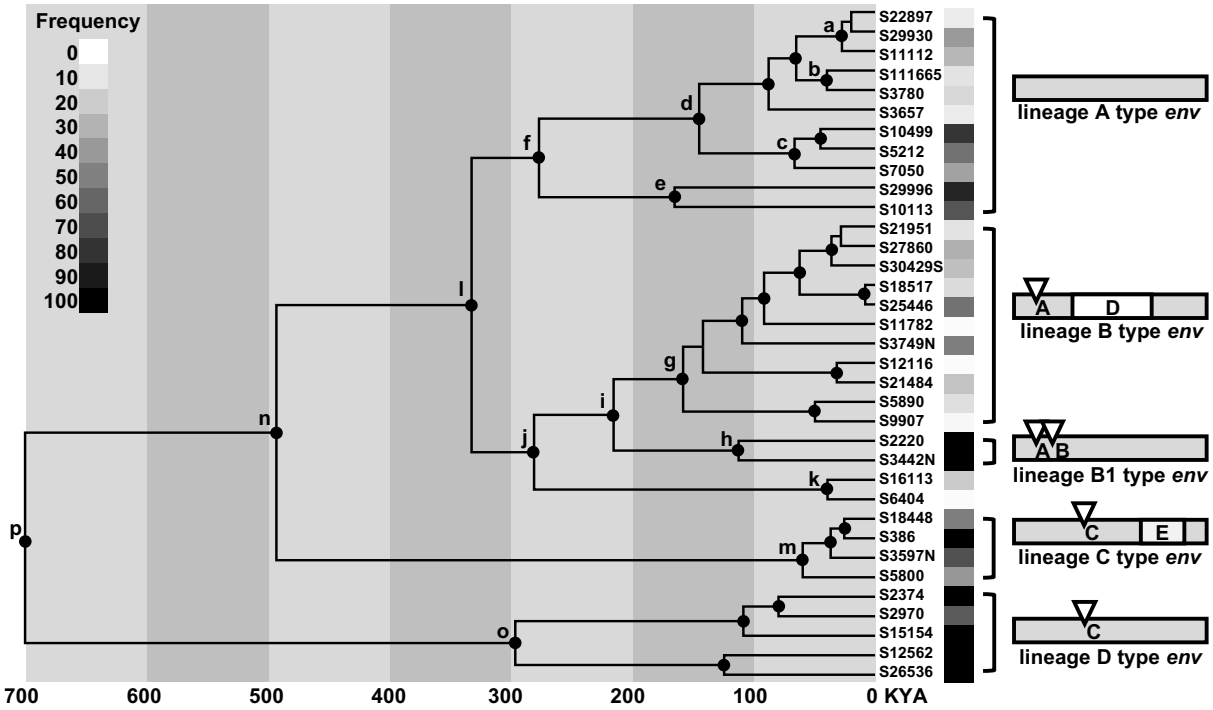

Figure 3

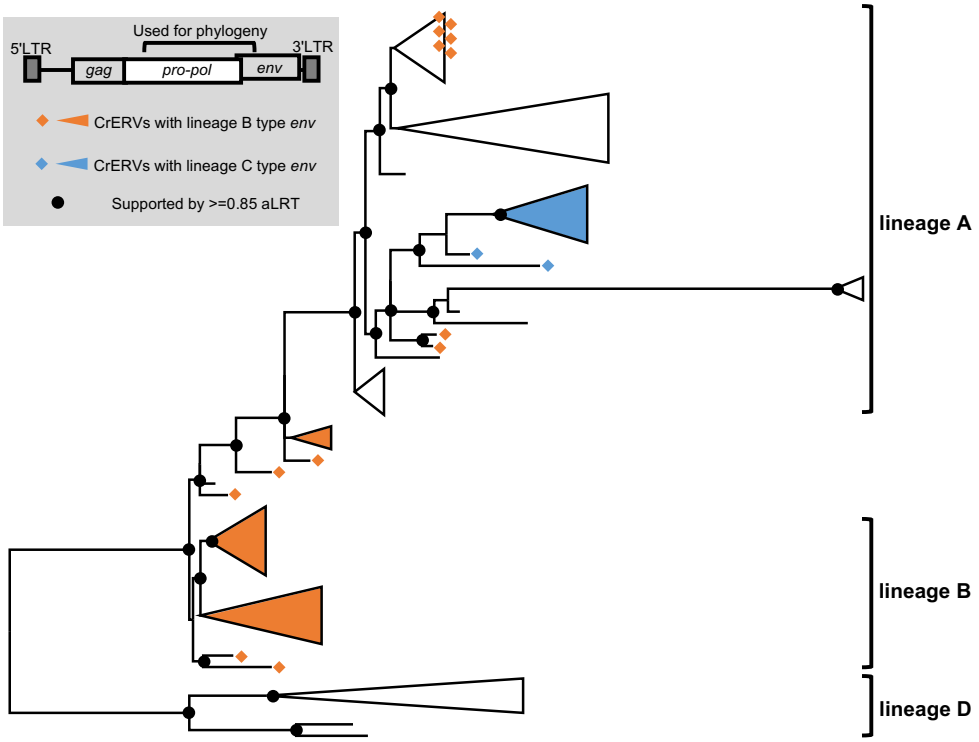
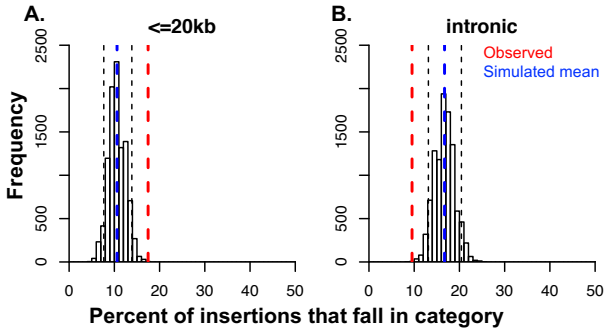
Figure 4