Epigenetic motifs distinguishing endogenous from exogenous retroviral integrants.

Authors: Sarah LaMere,[1]* Hanbei Xiong,[2,3] Wei Wang,[2,3] Brian LaMere,[4] Niema Moshiri[5]

[1]Department of Medicine, University of California San Diego, La Jolla, California, USA

*Corresponding author: salamere@health.ucsd.edu

[2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, USA

[3]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California, USA

[4]Escondido, CA

[5]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA

Corresponding Author:

Sarah LaMere, D.V.M., Ph.D.
Department of Medicine
University of California San Diego
9500 Gilman Dr.
La Jolla, CA 92093
Phone: (858) 947-5055
E-mail: salamere@health.ucsd.edu

Running Title: Comparative endogenous and exogenous retroviral epigenetic motifs

Keywords: DNA methylation, endogenous retrovirus, epigenetics

Acknowledgements: This work was supported by the James B. Pendleton Charitable Trust, UC San Diego faculty research funds, the California HIV Research Program CHRP H24BD7869 (LaMere), and grants from the National Institutes of Health: 1K01OD026565 (LaMere), R01NS137852 (LaMere), and R01HG009626 (Wang). We would also like to acknowledge Dr. Diako Ebrahimi for the provision of the Python code used for the Kmer analysis in this manuscript.

Abstract (231 words)

Retroviruses are subject to epigenetic regulation by the host genome after integrating, similar to vertebrate genes. However, their patterns of integration, and therefore their likely epigenetic regulation, differ between classes. Beta- and gammaretroviruses are two classes of simple retroviruses known to integrate near promoters, and these viruses have a strong tendency to infect germ cells and endogenize. While ancient endogenous retroviruses are often easy to spot due to mutations rendering them non-functional, more recent integrants can maintain the capacity for full viral production, making it sometimes difficult to discern which integrants are exogenous and likely more clinically relevant. Because endogenous retroviruses generally spend a longer time integrated and subject to host epigenetic regulation as proviral DNA, we hypothesized we could show these integrants exhibit sequence differences from their exogenous counterparts, likely resulting from DNA methylation and histone modifications, and that endogenous retroviruses would generally show habituation to host promoters. Therefore, we have used statistical analyses of publicly available sequence data to demonstrate that endogenous retroviral variants exhibit decreased CpG dinucleotide and trinucleotide frequencies over time, and that they will show

evidence for loss of histone modification motifs associated with actively transcribed genes. Close examination of these patterns provides further clues for distinguishing endogenous and exogenous retroviral variants, potentially aiding in the study of retroviruses in less well-characterized wildlife species where retroviral outbreaks are responsible for significant disease and decline in populations.

Introduction

Retroviruses are a unique class of viruses that are capable of integrating into the genomes of their hosts. Because these viruses carry their own distinctive polymerase capable of reverse transcribing the viral RNA genome into DNA, followed by integration into the host DNA, they are capable of permanently altering the genome in both the cell they infect and its progeny. When retroviruses infect germ cells, they can become incorporated into every cell of the resultant offspring, a process known as endogenization. Following endogenization, retroviruses will tend to acquire mutations rendering their promoter sequences and/or gene products deficient over the course of generations. Therefore, the degree of mutation of an endogenous retroviral sequence can give a window into how long it has been endogenized in a species.[1] Over 40% of mammalian genomes are reported to be products of reverse transcription and endogenization. In stark contrast to the 1.5% of human DNA known to be protein coding, 8-10% of the human genome is comprised of endogenous retroviruses.[2]

Retroviruses are divided into seven main classes based upon their genome organization and structures.[3] Different classes of retroviruses exhibit different propensities for the genomic

regions into which they tend to integrate. For example, lentiviruses, which include human immunodeficiency virus (HIV) and feline immunodeficiency virus (FIV), have a tendency to integrate into transcriptionally active gene bodies. On the other hand, gammaretroviruses, which include murine leukemia virus (MuLV) and feline leukemia virus (FeLV), tend to target gene promoters.[4] Betaretroviruses, such as mouse mammary tumor virus (MMTV) and Jaagsiekte sheep retrovirus (JRSV), also tend to target transcriptional start sites.[5] Gammaretroviruses and betaretroviruses are much more prone than lentiviruses to infecting germ cells and endogenizing, as evidenced by their retroviral remnants being present in every major mammalian genome examined, including HERV-K (betaretrovirus) and HERV-W (gammaretrovirus) in humans. Interestingly, endogenous beta and gammaretroviruses can frequently have exogenous counterparts that are fully infectious and replication-competent. Their sequences are often only slightly divergent from each other, usually in the long terminal repeat (LTR) and/or envelope (env), impacting their expression and tropism, respectively.[6]

Endogenous and exogenous retroviral variants are well-characterized for some species, such as MuLV and FeLV. However, the distinction between recent integrants with transcriptionally intact genomes in more poorly characterized species can be difficult. A case in point is the koala retrovirus, KoRV, which is a gammaretrovirus in the process of endogenizing. The virus has been associated with leukemias and lymphomas, as well as with secondary infections such as chlamydia.[7] Thus far, ten different subtypes have been described, varying with geography and exhibiting different associations with disease. KoRV-A appears to be the endogenous variant that has arisen from this virus, while the other subtypes exhibit more overt associations with disease and are believed to be exogenous.[8]

To make matters more confusing, endogenous and exogenous retroviruses will recombine. The endogenous variant of FeLV can recombine with the exogenous variant FeLV-A to yield a new virus with different tropism, and recombination events can even occur with other types of less related endogenous retroviruses, giving rise to diversity and evolution within the genera.[9] Most of these recombination events result from viruses within a single or closely related species, but there are examples of recombination between gammaretroviruses from more divergent species. For example, RD114, another endogenous gammaretrovirus found in domestic cats, appears to be the product of a recombination event with Baboon Endogenous Retrovirus (BaEV) and is believed to have resulted from a cross-species viral transfer from an old-world monkey to a domestic cat.[10]

Multiple host epigenetic mechanisms used to suppress endogenous retroviral infections have been described, including DNA methylation, histone modifications, and non-coding RNAs.[11] Exogenous retroviruses can also be subject to these same mechanisms as a molecular arm of the host antiviral defense, making them an active area of study in HIV.[12] However, endogenous retroviruses that have co-evolved with their host species are frequently more highly transcriptionally suppressed than their exogenous counterparts,[11] prompting the hypothesis that sequence-specific evidence of epigenetic suppression will generally be more apparent in endogenous variants.

Previously, HIV was reported to mirror the human genome with its low CpG content, exhibiting evidence of deamination following CpG methylation.[13] Because of their propensity

121 to integrate into promoter regions, we hypothesized that beta and gammaretroviruses would
122 mirror the CpG content of the promoters in their respective species, and that endogenous
123 variants would exhibit decreased CpG content relative to exogenous variants, suggesting
124 more CpG methylation and deamination following endogenization, as well as altered CpG
125 trinucleotide motifs consistent with the methylation patterns in their host species. We also
126 hypothesized that motifs consistent with active histone modifications would be more apparent
127 in exogenous proximal proviral sequences, while those more consistent with repressive
128 motifs (e.g. H3K9me3) would be more evident in known endogenous sequences. To test
129 these hypotheses, we performed statistical analyses on CpG frequencies and CpG
130 trinucleotide motif frequencies from various retroviruses and compared endogenous to
131 exogenous variants, as well as all those from promoter sequences of the species in which
132 they are present. We also performed bisulfite mapping to examine methylation patterns in
133 endogenous retroviruses versus promoters. Finally, we examined differences in motifs
134 associated with histone modifications using the Simple Enrichment Analysis tool from the
135 MEME suite.[14]
136
137 **Results**
138 *Gammaretroviruses evolve to reflect CpG content of eutherian host promoters, with variability*
139 *attributable to DNA methylation*
140
141 To evaluate how betaretroviruses and gammaretroviruses adapt to the genomes of their
142 respective species, we examined CpG content of full-length endogenous and exogenous
143 retroviruses from cats, mice, pigs, sheep, and koalas from **Tables 1 and 2** (accession

144 numbers in **Supplementary Table S1**) and compared them to host promoter CpG content
145 (defined as 1000 bp upstream of all transcriptional start sites). In three cases, retroviral CpG
146 content was reflective of the median promoter CpG content **(Figure 1a-c)**. Sheep and koalas,
147 on the other hand, had a lower promoter CpG content that has not previously been reported
148 **(Figure 1d-e)**. Marsupials might be a unique case, in that most DNA methylation occurs more
149 commonly in gene bodies instead of promoters,[15] and differences in promoter CpG content
150 reflect potential differences in epigenetic regulation.
151
152 *Endogenization impacts CpG content of beta and gammaretroviruses due to CpG*
153 *methylation*
154
155 When comparing CpG content of endogenous vs. exogenous retroviruses, endogenous
156 gammaretroviruses exhibited a significant decrease in CpG content compared to their
157 exogenous counterparts. However, the betaretroviruses analyzed did not have as robust a
158 difference with the sequences available **(Figure 2a)**. To examine the role of DNA methylation
159 in the CpG content of all retroviruses examined, we calculated the correlation between CpG,
160 TpG, and CpA observed to expected ratios (O/E ratio) to assess for deamination of
161 methylated cytosines. Deamination of methylated cytosines occurs over time, changing the
162 CpG landscape to TpG and CpA on the complementary strand, while non-methylated
163 cytosines are not subject to these same forces.[16] Strikingly, both the endogenous and
164 exogenous betaretroviruses, especially Jaagsiekte Sheep Retrovirus (JSRV), showed a
165 strong significant symmetrical negative correlation **(Figure 2b-d)**. The correlation for
166 exogenous JSRV was surprisingly more profound than that found in the endogenous variant.

167 Exogenous JSRV is believed to have arisen from a previous endogenous retrovirus,
168 suggesting it has likely been subject to selection pressure from host DNA methylation more
169 than other exogenous retroviruses.[17] However, endogenous JSRV showed heavy methylation
170 across the provirus, with a preference for CGC and CGG motifs **(Figure 2e-f)**. Consistent
171 with the predilection for CGC methylation, CGC motifs had the strongest correlation with
172 TGC/CAC **(Figure 2g)**, while endogenous JSRV showed a significant decrease in CGC O/E
173 ratios compared to both exogenous JSRV subtypes **(Figure 2h)**. MMTV had few full-length
174 replicates, three of which had identical sequences, calling into question the statistical power
175 of the analysis for this virus. However, Bismark bisulfite mapping demonstrated heavy CpG
176 methylation across the body of the MMTV provirus **(Figure 2i)**, with CGC trinucleotides
177 exhibiting a small but significantly higher percent methylation **(Figure 2j)**. CGC O/E ratios did
178 not exhibit significant differences between endogenous and exogenous MMTV (data not
179 shown), but few replicates were available for this analysis, making it difficult to achieve
180 significance.
181
182 Correlations between CpG and TpG/CpA in gammaretroviruses were less pronounced; FeLV,
183 which only included 8 replicate endogenous and exogenous full-length sequences each that
184 were publicly available, showed no significant negative correlation, and differences in TpG
185 and CpA O/E ratios were not significant between endogenous and exogenous variants.
186 However, O/E ratios of the trinucleotide motif CGC showed a significant decrease in enFeLV
187 compared to exFeLV, while TGC and CAC were significantly increased **(Figure 3a)**,
188 suggesting this trinucleotide motif in particular is prone to DNA methylation in the context of
189 enFeLV. Bisulfite mapping to cat promoters (defined as 1kb upstream of all transcriptional

190 start sites) showed significantly lower CGC and CGG methylation **(Figure 3b)**, suggesting
191 this pattern is inconsistent with that seen in cat promoters. Unfortunately, mapping the
192 bisulfite converted cat genome to enFeLV failed to produce robust results, preventing
193 definitive conclusions about the methylation patterns in enFeLV and their relationship to that
194 of promoters in the domestic cat.
195
196 More replicate sequences of endogenous and exogenous MuLV were available for analysis
197 from [18], allowing a comprehensive examination of this gammaretrovirus. We analyzed
198 promoters from three endogenous variants: polytropic (PMV), modified polytropic (MPMV),
199 and xenotropic (XMV) MuLV. In addition to a significant decrease in CpG O/E ratios between
200 enMuLV and exMuLV **(Figure 2a)**, enMuLV also exhibited a significant increase in TpG and
201 CpA O/E compared to exMuLV, suggesting a likely role by DNA methylation **(Figure 3c)**. A
202 significant correlation was also observed between CpG O/E and TpG/CpA O/E ratios in
203 enMuLV, providing further support for this conclusion **(Figure 3d)**. Bisulfite analysis of mouse
204 promoters showed no significant difference in frequency of trinucleotide motifs methylated in
205 mouse promoters, and enMuLV also showed no difference **(Supplementary Figure S1)**.
206 CGG trinucleotides showed no significant difference in frequency between enMuLV and
207 exMuLV, and while CCG trinucleotides did show a difference, CCG frequency negatively
208 correlated only with CCA and not CTG **(Supplementary Figure S2a-e)**, suggesting
209 APOBEC3G was a more likely cause of the asymmetric deamination.[19] Analysis of all
210 enMuLV sequences using Hypermut 3[20] demonstrated significant hypermutation due to
211 APOBEC3G for all but three of the 49 sequences examined **(Supplementary Table S2)**,
212 providing further support for this. Of note, the paper from which these enMuLV sequences

213 were retrieved discusses the role of APOBEC3 in genetic diversity among enMuLVs.[18]
214 Interestingly, as was true for FeLV, a significant reduction was found in CGC trinucleotide
215 motifs in the endogenous MuLVs compared to exMuLVs, while a significant increase in
216 corresponding TGC and CAC frequency was also observed **(Figure 3e)**. A modest but
217 significant negative correlation between CGC with TGC and CAC was also found in enMuLV
218 **(Figure 3f)**, supporting the CGC motif as a more frequent past target for DNA
219 methyltransferases in enMuLV. Unlike the methylation patterns in the other endogenous
220 retroviruses, MuLV did not exhibit a significant preference for CGC methylation, nor did
221 mouse promoters **(Supplementary Figure S1)**, calling into question whether CpG
222 methylation is responsible in this case. However, these CpG frequency patterns might reflect
223 initial methylation following integration rather than endogenization. In all cases comparing
224 endogenous and exogenous variants, there was evidence of a predilection for CGC
225 methylation following endogenization, either via available methylation data, CGC O/E data, or
226 both.
227
228 *CpG content of endogenous retroviruses is variably impacted by DNA methylation and*
229 *corresponds with transcriptional activity*
230
231 Pigs have three highly related endogenous gammaretroviruses known as Porcine
232 Endogenous Retrovirus A, B, and C. The viral sequences are primarily different in their
233 envelope regions, as well as some difference in the LTRs that appear to confer varying levels
234 of transcriptional activity. PERV-A is the oldest, having endogenized approximately 6.6 million
235 years ago. PERV-B is close in age, having endogenized approximately 6.4 million years ago.

236 Finally, PERV-C is the most divergent of the three and is thought to have endogenized 3.2
237 million years ago.[21] Of the three, PERV-B is the most transcriptionally active, while PERV-C
238 is the least, and the CpG contents of the three proviruses reflect this, with PERV-B exhibiting
239 the highest CpG O/E ratio, and PERV-C exhibiting the lowest **(Figure 4a)**. Of the three
240 proviruses, only PERV-C demonstrates a correlation between CpG and TpG/CpA O/E ratios
241 **(Figure 4b)**, suggesting this provirus is the most heavily impacted by DNA methylation. To
242 attempt to confirm this finding, we mapped a bisulfite converted pig genome to full-length
243 PERV-A, PERV-B, and PERV-C. The PERV-C envelope had no reads mapping, suggesting
244 this variant was not present in the publicly available bisulfite converted pig genome we
245 accessed. However, methylation was heavy across both PERV-A and PERV-B **(Figure 4c)**.
246 Because sequences are identical for the two variants across the LTR, gag, and pol, we
247 compared CpG methylation across the env regions of the two subtypes, demonstrating that
248 CpG methylation was comparable between the two at approximately 80% **(Figure 4d)**.
249
250 The koala gammaretrovirus, KoRV, is the youngest known endogenous retrovirus, as
251 evidence points to it being in the midst of endogenizing. KoRV-A was the first subtype
252 discovered and appears to be the most commonly endogenized variant, showing less
253 association with disease than other subtypes. It is not entirely clear whether the other
254 subtypes are endogenous or exogenous, but subtypes B-F are believed to be primarily
255 exogenous.[8] The CpG patterns of the other subtypes suggest they are less impacted by
256 DNA methylation and potentially more transcriptionally active. For example, KoRV-A exhibits
257 a significantly lower frequency of CpG dinucleotides than the other KoRV subtypes **(Figure**
258 **4e)**. Additionally, KoRV-A shows a significant correlation between CpG and TpG/CpA, once

259 again suggesting DNA methylation plays a more significant role for this subtype **(Figure 4f)**,
260 while no significant correlation was found between CpG O/E ratios and TpG/CpA ratios of the
261 other subtypes (data not shown). Mapping of bisulfite converted koala DNA against the
262 KoRV-A sequence showed heavy DNA methylation across the provirus, which was also
263 recently reported by [22] **(Figure 4g)**, reinforcing the likely impact of DNA methylation upon the
264 CpG composition of this provirus. Interestingly, the endogenous koala betaretrovirus, phaCin-
265 β,[23] also exhibited evidence for past DNA methylation, with a significant correlation between
266 CpG O/E ratios with TpG/CpA ratios **(Figure 4h)**. Calculation of median %CpG methylation
267 from mapping of bisulfite converted genomes for all endogenous retroviruses examined
268 showed a significant correlation to median CpG O/E ratios, with KoRV-A proving to be a
269 major outlier **(Figure 4i)**, perhaps reflecting its recent integration into the koala genome, as
270 CpG frequencies have not yet decreased to the same extent as older endogenous
271 retroviruses.
272
273 *Exogenous retroviruses exhibit more histone modification motifs associated with active*
274 *transcription than endogenous variants*
275
276 To evaluate differences in histone modifications across endogenous and exogenous
277 retroviruses, we used Simple Enrichment Analysis (SEA) from the MEME suite[14] to compare
278 endogenous to exogenous proviral sequences and discover the most commonly enriched
279 motifs in each. For this analysis we compared each of two types of endogenous MuLV
280 sequences (modified polytropic (MPMV) and polytropic (PMV)) to exogenous MuLV
281 sequences. We also compared endogenous FeLV to exogenous FeLV-A, and endogenous

282 JSRV to the two types of exogenous JSRV (Enzootic Nasal Tumor Virus, or ENTV, and
283 Ovine Progressive Pleuropneumonia, or OPP).
284
285 Because mice have considerably more genome-wide histone modifications publicly available,
286 a more thorough characterization of histone modification motifs across MuLVs was possible.
287 We performed CpG island (CGI) analysis with Emboss Cpgplot[24] to refine results for histone
288 modifications typically confined to promoters and CGIs. CGI analysis showed that exogenous
289 MuLV exhibited a CGI not present in the endogenous variants **(Figure 5a)**. Three statistically
290 significant motifs for H3K27ac and H3K9ac, which are histone modifications associated with
291 active transcription, were found in this CGI that were not located in the endogenous MuLVs
292 **(Figure 5a, Supplementary Table S3)**. However, in PMV, which is more transcriptionally
293 active than MPMV,[25] a considerable number of motifs associated with both active (H3K9ac,
294 H3K27ac, H3K4me3) and repressive (H3K27me3, H3K9me3) modifications were
295 concentrated in the CGI traversing the U5 region of the 5' LTR and the leader region **(Figure**
296 **5b)**. Interestingly, in addition to some of the 'active' histone modification motifs, MPMV had
297 multiple repressive H3K27me3 motifs throughout the provirus, including in the LTR, the CGI
298 in the leader region, located in *gag*, and also in the CGI located at the *pol-env* junction
299 **(Figure 5c)**.
300
301 To attempt to validate these findings, we examined enMuLVs aligned with mouse genome
302 assembly MM10 in the UCSC Genome Browser with ENCODE histone modification tracks.
303 Surprisingly, 73% of enMuLV's (57/78) showed no histone modifications at all, but a few did,
304 summarized in **Table 3**. Repressive H3K9me3 and H3K27me3 were the most commonly

305 found modifications, with some modifications associated with active transcription, including
306 H3K27ac, H3K4me3, and H3K9ac **(Figure 5d)**. H3K27me3 and H3K27ac were most
307 commonly located in the proximal LTR, while H3K4me3 and H3K9ac were distributed among
308 CGIs and CpG-rich regions. H3K9me3 was broadly distributed across gene bodies, while
309 H3K36me3 was located in the gene body of one provirus, with H3K9me3 marking the same
310 spot. CpG methylation in embryonic stem cells and neural progenitors was usually found in
311 sparse or moderate amounts in all proviruses examined and frequently overlapped histone
312 modifications when they were present **(Table 3, Supplementary Figure S3a-b)**. However,
313 mapping of bisulfite converted DNA from liver tissue showed heavy average CpG methylation
314 across the provirus, suggesting CpG methylation might be more critical for proviral epigenetic
315 regulation after differentiation **(Supplementary Figure S3c)**, although these data represent
316 an average of proviral CpG methylation from different integrants at different sites vs. the
317 individual sites examined for histone modifications.
318
319 To examine another gammaretrovirus, we generated H3K4me3 and H3K27ac motifs for the
320 domestic cat and used SEA to examine enrichment throughout the endogenous and
321 exogenous FeLVs. As both of these marks are associated with active transcription, there was
322 a marked absence of any of these motifs meeting the relatively high significance threshold in
323 enFeLVs (e-value < 0.05). However, exFeLV-A had multiple CGIs with which these
324 modifications were associated **(Figure 6a, Supplementary Table S4)**.
325
326 We also examined betaretroviruses for significantly enriched histone motifs. SEA analysis
327 was overall unsuccessful for MMTV because too few full-length sequences were publicly

328 available. However, JSRV analysis yielded some interesting findings. Both JSRV and MMTV
329 exhibit considerably fewer conventional CGIs than the gammaretroviruses, so we narrowed
330 the minimum length to 100 bp to pick up CpG-rich regions where histone modifying enzymes
331 would be more likely to bind. All JSRV variants had short CpG stretches in the U5 region of
332 the LTR and the CGI located in the pol-env junction. Endogenous JSRV had a CpG-rich
333 region located in the *gag-pro-pol* junction, while the exogenous JSRVs had one more
334 proximally located in *gag* **(Figure 6b-d)**. A few statistically significant H3K4me3 sites were
335 found in the endogenous JSRV, but overall they were sparse **(Figure 6b, Supplementary**
336 **Table S5)**. Both exogenous JSRVs had H3K27ac and H3K4me3 motifs in the proximal LTR
337 **(Figure 6c-d)**. However, the ENTV JSRV variant had more of these motifs in the CpG-rich
338 leader region and the *pol-env* junction, while the OPP variant had significant motifs in *gag,*
339 suggesting these two exogenous variants are potentially regulated differently.
340

341 **Discussion**

342 We have shown that beta and gammaretroviruses frequently reflect the CpG content of their
343 host promoters, but this is not always true in species where promoter CpG content is
344 naturally lower. As shown in **Figure 1**, both beta and gammaretroviruses can reflect promoter
345 CpG content, where in most species, ~67% of promoters will contain CpG islands.[26]
346 Interestingly, sheep and koalas, where the retroviral CpG content did not reflect that of
347 promoters, exhibited significantly lower CpG content than other vertebrate species examined
348 (i.e. 0.3 vs 0.45-0.5) **(Figure 1)**. This could suggest that these retroviruses reflect evolution
349 within other species first. KoRV in particular is hypothesized to have arisen in either rodents
350 or bats based on phylogenomic studies.[27, 28] JSRV is an older virus thought to have

351 integrated around 5-7 million years ago, but the exogenous virus is believed to have arisen
352 recently from endogenous variants and continues to produce more endogenous integrants.[17]
353 The strong negative correlation between CpG and TpA/CpG motifs appears to reflect this
354 evolutionary history, suggesting a long history of host CpG methylation and therefore
355 integration within the species. Overall, sheep promoters exhibit low CpG content compared to
356 other species, resulting in a disparity between CpG content of promoters and JSRV.
357 Betaretroviruses have a predilection for integration near transcriptional start sites,[5] but they
358 do not generally show a preference for CGIs as gammaretroviruses do,[4] suggesting their
359 integration requirements might be less dependent on CGIs. The viral p12 protein of
360 gammaretroviruses is known to tether the preintegration complex to nucleosomes via
361 bromodomain and extraterminal domain (BET) proteins, playing a role in both integration and
362 in nucleosome loading onto the provirus while targeting the PIC to CpG promoters and
363 enhancers.[29, 30] However, a tethering mechanism still has not been elucidated for
364 betaretroviruses,[31] so it is more difficult to hypothesize how their evolution might be impacted
365 by host CpG content.
366 We have also shown that endogenization clearly impacts CpG content and CpG trinucleotide
367 motif patterns, and that these results are likely due to CpG methylation. Every endogenous
368 retrovirus to which we mapped bisulfite sequencing exhibited >80% methylation across the
369 entire provirus **(Figures 2 and 4, Supplementary Figure S3c)**. CpG methylation has long
370 been known to silence endogenous retroviruses during embryonic development.[32] However,
371 CpG methylation within gene bodies is generally not silencing and instead associated with
372 active transcription.[33] From an epigenetic standpoint, this could suggest that the entire
373 provirus is treated similarly to a silenced host promoter. That said, even highly transcribed

374 endogenous retroviruses (e.g. PERV[21]) showed heavy CpG methylation in somatic cells
375 **(Figure 4c)**, calling into question what this mechanism achieves beyond transcriptional
376 silencing.
377 Almost across the board, endogenous retroviruses exhibited significantly lower CpG content
378 than their exogenous counterparts **(Figures 2-4)**, with some showing evidence more refined
379 to methylation of particular CpG trinucleotide motifs. In general, these did not appear to
380 reflect the methylation patterns of specific trinucleotide motifs in the host species. Different
381 DNA methyltransferases will target different CpG trinucleotide motifs more heavily,[34, 35] and it
382 is unknown whether these might change between species. The varying promoter methylation
383 patterns of CpG trinucleotides between the different species seem to suggest this is the case.
384 Retroviral methylation is often initiated during embryonic development by *de novo*
385 methyltransferases DNMT3a and DNMT3b.[34-36] This may not be reflective of the overall
386 methylation pattern in somatic cells, which is normally maintained by DNMT1.[37]
387
388 Using enMuLV data, we found that histone modifications were generally sparse among
389 endogenous retroviruses, suggesting that the seemingly ubiquitous DNA methylation we
390 observed **(Figures 2 and 4, Supplementary Figure S3c)** was more likely important for
391 endogenous retroviral epigenetic regulation. Of proviruses that did have modifications, those
392 associated with both active (H3K27ac, H3K9ac, H3K4me3, H3K36me3) and repressive
393 (H3K27me3, H3K9me3) chromatin could be found **(Figure 5d, Table 3)**. It is probable the
394 proviral histone modification pattern is dictated by the integration site environment and less
395 by the endogenous provirus itself, which requires further exploration. However, there did
396 appear to be significant differences in motifs associated with active and repressive histone

397 modifications when comparing endogenous and exogenous retroviruses **(Figures 5 and 6)**,
398 suggesting histone modifications could be playing an important role early after
399 endogenization, but future work will need to evaluate this hypothesis. Alterations to histone
400 modifications frequently precede DNA methylation for vertebrate gene regulation,[38] so newly
401 endogenized proviruses might be treated similarly. However, previous work on the more
402 ancient human endogenous retroviruses suggested that DNA methylation was an earlier
403 modification, while histone modifications were found in older and more degenerate LTRs.[39]
404 The profound (~93%) DNA methylation across the provirus of the recently integrated KoRV-A
405 supports this hypothesis **(Figure 4g)**. Additionally, KoRV-A defied the clear relationship
406 between proviral DNA methylation and CpG frequency observed for older endogenous
407 proviruses **(Figure 4i)**, suggesting changes to the viral sequence associated with methylated
408 cytosine deamination have thus far been minimal, and supporting the idea that DNA
409 methylation is a common initial epigenetic method for regulation of recent integrants.
410 Interestingly, with the exception of MMTV, the amount of CpG methylation across
411 endogenous proviruses appeared to correlate roughly to their ages, with lower amounts of
412 methylation corresponding to older integrants **(Table 1, Figure 4i)**, further supporting this
413 hypothesis.
414
415 We confined our analysis of mouse enMuLVs to sequences over 4000 bp, selecting for more
416 intact sequences, and the majority of these sequences (73%) contained no histone
417 modifications. Of those that did, CpG methylation was frequently present overlapping both
418 activation-associated and repressive histone marks in neural progenitors **(Table 3,**
419 **Supplementary Figure S3a-b)**. CpG methylation in embryonic stem cells and neural

420 progenitors was almost ubiquitously present, but generally either in sparse or moderate
421 amounts. However, heavy average CpG methylation is present across the enMuLV provirus
422 in mouse liver **(Supplementary Figure S3c)**, suggesting DNA methylation is dynamic across
423 enMuLV during development and also suggesting marked regulation via CpG methylation in
424 somatic tissue.
425
426 In conclusion, retroviral endogenization shapes the sequence landscape of the provirus due
427 to host epigenetic factors. CpG methylation clearly impacts proviral CpG content over time,
428 while histone modification motifs appear to reflect the transcriptional state of the provirus,
429 whether endogenous or exogenous. Generally speaking, of the three retroviruses we
430 evaluated for histone modification motifs, exogenous variants appeared to have more
431 prominent activation-associated motifs (i.e. H3K27ac and H3K4me3) than their endogenous
432 counterparts relative to repression-associated motifs **(Figures 5 and 6)**. Endogenous MuLVs,
433 for which we had more well-defined motifs to examine, exhibited more significant motifs
434 associated with repression than exogenous variants, although those classified as PMV also
435 harbored a number of activation-associated motifs, consistent with their known elevated
436 transcriptional activity compared to other enMuLVs.[40] While no one clear pattern emerged to
437 distinguish endogenous from exogenous variants, we are still able to show that
438 endogenization impacts the proviral sequence landscape in generally predictable ways,
439 despite our analyses traversing multiple endogenous integrants in what are likely a wide
440 variety of chromatin environments from different species. These changes due to evolutionary
441 pressure from the host epigenetic machinery provide a starting point to aid in deciphering
442 when beta- and gammaretroviruses are clinically significant and warrant further investigation.

443

444 **Methods**

445

446 *CG trinucleotide motif O/E calculation*:

447

448 To calculate trinucleotide D-ratios (Observed/Expected or O/E ratios), a first order Markov

449 model of conditional probabilities was used.[13, 41, 42] The expected frequency of each motif is

450 estimated using the individual base components with the following example equation:

451

452 $$P_{exp}(CpGpT) = \frac{P_{obs}(CpG) \times P_{obs}(GpT)}{P_{obs}(G)}$$

453

454

455 The D-ratio was calculated by dividing the P_{obs} of each trinucleotide by the P_{exp}.

456

457 CG trinucleotide D-ratios were calculated for all proviral sequences using the provided Kmer

458 tool, as well as promoter sequences retrieved from each respective species using the

459 *Ensembl* Biomart Tool,[43] where sequences 1000 bp upstream of each transcriptional start

460 site of all Refseq annotated genes were extracted.

461

462 *Bisulfite sequencing alignment and cytosine methylation calculation*

463 Bisulfite converted genomes used in these analyses are reported in **Supplementary Table**

464 **S6**. SRA files obtained from NCBI were converted to fastq using the NCBI prefetch utility.

465 Fastq files were subjected to Trim Galore!.[44] Alignment of retrieved promoter and retroviral

466 sequences (accessions in **Supplementary Table S1**) to bisulfite converted genomes was
467 performed using Bismark software.[45] Deduplication and methylation extraction were also
468 performed using Bismark. CpG methylation reports generated from Bismark were used to
469 calculate CpG methylation for CGA, CGC, CGG, and CGT trinucleotides.
470
471 *Histone motif generation*
472 Motifs for all mouse histone modifications were retrieved from[46]. Epigram was used as
473 described in [46] to generate motifs for H3K4me3 and H3K27ac in sheep and domestic cats
474 using the following data sets: H3K4me3 and H3K27ac from sheep liver GSE206736[47] aligned
475 with the Ovis aries genome assembly Oar_rambouillet_v1.0, and H3K4me3 and H3K27ac
476 from cat heart GSE182952 aligned with the Felis catus genome assembly
477 F.catus_Fca126_mat1.0.
478
479 *CpG Island Identification*
480 For CpG Island identification, consensus sequences for each provirus were generated using
481 MacVector™ software. Sequences were entered into Emboss Cpgplot[24] using a window size
482 of 100, a minimum length of 200, a minimum observed CpG frequency of 0.6, and a minimum
483 CpG percentage of 50. For JSRV sequences, CGI minimum lengths were dropped to 100.
484
485 *SEA analysis*
486 For SEA analysis, histone modification motif analyses for exogenous retrovirus variants were
487 submitted for enrichment compared to their corresponding endogenous variants as negative
488 controls, and vice versa with histone motifs generated from Epigram analyzed for enrichment

489 as described above with an E-value cutoff of 0.05. Motifs for H3K4me3, H3K27ac, and
490 H3K27me3 were retrieved only if they aligned with CpG islands from the corresponding
491 proviruses being analyzed. In all cases, only sequence motifs enriched in all inputted
492 sequences with no more than one of the negative control sequences, or one less than all
493 inputted sequences with no negative control sequences, were considered significant, even
494 when E-value and CGI criteria were met.
495
496 *enMuLV histone modification identification*
497 Consensus sequences for endogenous MuLV variants MPMV, PMV, and XMV were all
498 entered into the UCSC Genome Browser BLAT tool.[48, 49] ENCODE histone tracks for
499 H3K27ac, H3K27me3, H3K9ac, H3K9me3, H3K4me3, H3K36me3, and H3K4me1 from
500 forebrain, liver, kidney, and intestine from day 16 embryonic mice aligning ≥ 4000 bp of
501 enMuLV sequence were visualized to identify sites of modification throughout the provirus.
502 ENCODE tracks for CpG methylation in mouse embryonic stem cells and neural progenitor
503 cells were also included.
504
505 *Statistical analysis*
506 Statistical analyses and data visualization were performed using GraphPad Prism software
507 version 10.4.2 for macOS. Multiple comparisons for CpG trinucleotide motifs between
508 proviruses and promoters were performed using the Kruskal-Wallis test. Two-sided
509 independent t-tests or Mann-Whitney tests were used for individual comparisons where
510 appropriate. All correlation analyses were performed using Spearman's rank correlation test
511 with a 95% confidence interval. Pearson correlation was used in one case (Figure 4i) after

512 the normality assumption was confirmed by Shapiro-Wilk and Komogorov-Smirnov tests. In

513 all analyses, statistical significance was defined as p-value < 0.05.

514

515

516 **Tables:**

Table 1: Endogenous with counterpart exogenous retroviruses used for analysis.

Gammaretroviruses						
	Exogenous vs endogenous % identity	**Integration time**	**Exogenous CpG O/E**	**Endogenous CpG O/E**	**Exog. N**	**Endog. N**
FeLV	86%	<2 MYA[50]	0.53	0.49	8	8
MuLV	90-100%	1 MYA[25]	0.53	0.52	16	49
KoRV	90%	22,000-49,000 YA[51]	0.51	0.5	8	16
Betaretroviruses						
MMTV	95-99%	20 MYA[52]	0.45	0.46	6	5
JSRV	90-98%	5-7 MYA[17]	0.58	0.57	20	19

517

Table 2: Further endogenous retroviruses used for analysis.

Gammaretroviruses			
	Integration time	**CpG O/E**	**Endog. N**
PERV-A	6.6 MYA[47]	0.52	8
PERV-B	6.4 MYA[47]	0.53	14
PERV-C	3.2 MYA[47]	0.48	14
Betaretroviruses			
phaCin-b	0.7-1.5 MYA[53]	0.52	18

518

519

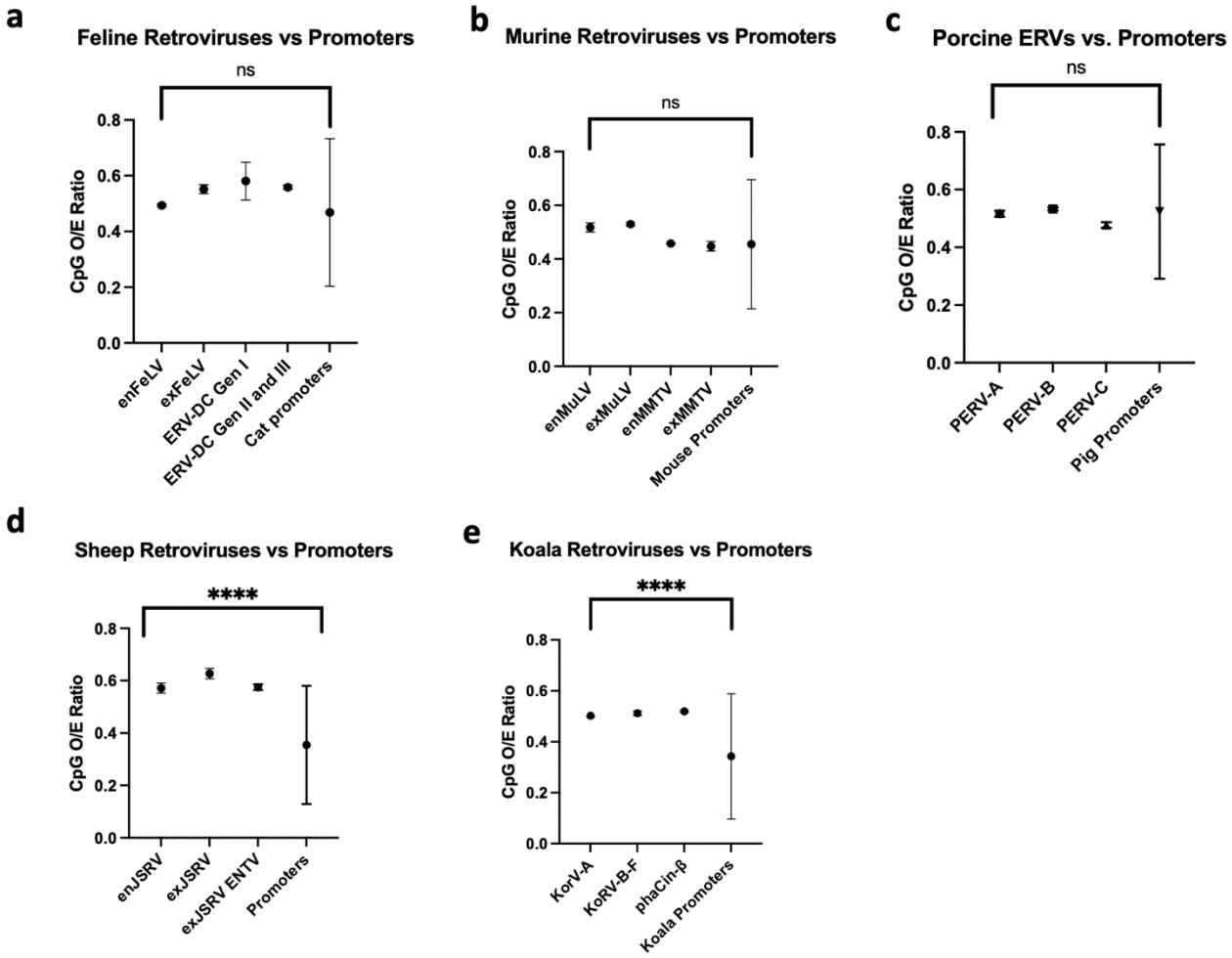
520

Table 3: Histone modifications found in enMuLV using UCSC Genome Browser.

Chromosomal position	Approximate Proviral Position							
	H3K27 ac	H3K27 me3	H3K4 me3	H3K9 me3	H3K9ac	H3K36 me3	CpG Meth ESC	CpG Meth NP
Chr13:21812647	~1-50		~1-50, ~1900		~1900		Sparse	Sparse
Chr15:76559905		~1-50					Moderate	Moderate
Chr5:23700719				~5700		~5700	Sparse	Heavy
Chr1:171481826				~2000			Sparse	Heavy
Chr1:164224022			~1085		~1085		Moderate	Moderate
Chr18:30266726			~1100, 1780		~1100, 1780		Moderate	Moderate
Chr8:123428312				~2830,3370,5650			Sparse	Heavy
Chr4:134159771			~925-1300		~925-1300		Moderate	Moderate
Chr18:82693551	~1-50		~1-50		~1-50		Sparse	Sparse
Chr9:62439174	~1-50				~1-50		Sparse	Sparse
Chr8:119889565		~1-150					Moderate	Moderate
Chr11:88890180		~1-50					Moderate	Moderate
Chr5:109901526				~1730			Sparse	Sparse
Chr19:38375022		~1-50					Moderate	Sparse
Chr5:24215308		~1-50					Sparse	Sparse
Chr1:164223807			~1100		~1100		Moderate	Moderate
Chr7:30688640		~1-50					Moderate	Sparse
Chr10:41436648		~1-50					Moderate	Sparse
Chr1:191933613		~1-50					Sparse	Sparse
Chr8:85125855		~1-50					Moderate	Moderate
Chr14:54292373		~1-50					Sparse	Sparse

521
522

523 **Figures:**



524

525 **Figure 1: CpG content of beta- and gammaretroviruses commonly reflects CpG content**

526 **of promoter sequences.** Endogenous and exogenous retroviral median CpG D-ratios are

527 compared to those of promoters for **a)** domestic cats **b)** mice **c)** pigs **d)** sheep and **e)** koalas.

528 P-values are reported from Kruskal-Wallis tests for multiple comparisons. Error bars

529 represent interquartile ranges. n.s. $p > 0.05$, **** $p < 0.0001$

530

531

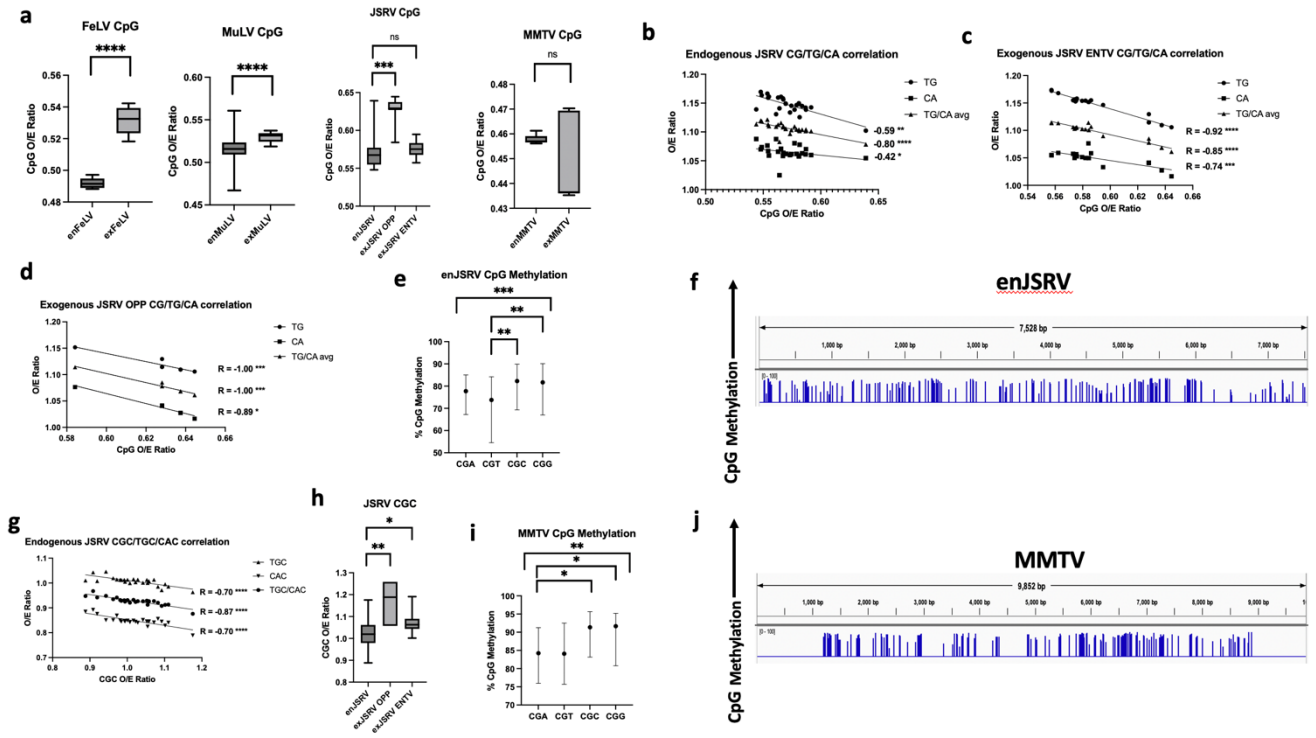


Figure 2: Endogenization of beta- and gammaretroviruses impacts retroviral CpG content: a) Endogenized retroviral variants exhibit lower CpG content than exogenous counterparts for Feline Leukemia Virus (FeLV), Murine Leukemia Virus (MuLV), and Jaagsiekte sheep retrovirus (JSRV). Mouse Mammary Tumor Virus (MMTV) shows no significant change to median CpG content. JSRV shows a strong negative correlation between CpG content and TpG/CpA content in **b)** endogenous **c)** exogenous Enzootic Nasal Tumor Virus (ENTV) and **d)** exogenous Ovine Progressive Pleuropneumonia (OPP) variants. **e)** CpG methylation across endogenous JSRV shows uneven CpG trinucleotide distribution, with a preference for CGC trinucleotides. **f)** Bismark mapping shows heavy CpG methylation across the entire JSRV provirus. **g)** Endogenous JSRV shows a strong negative correlation between CGC content and TGC/CAC. **h)** CGC content is significantly decreased in endogenous JSRV compared to both exogenous JSRV strains. **i)** Trinucleotide analysis of

545 MMTV methylation shows a significant preference for CGC motifs. **j)** Bismark mapping shows
546 heavy CpG methylation across the body of the MMTV provirus.
547 P-values for individual comparisons are from Mann-Whitney tests **(a)** and Kruskal-Wallis test
548 for multiple comparisons with Dunn's multiple comparisons **(e,i)**, while R and p-values for
549 correlation analyses are from Spearman correlations **(b-d,g)**. Error bars in **e)** and **i)** represent
550 interquartile ranges. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$
551
552

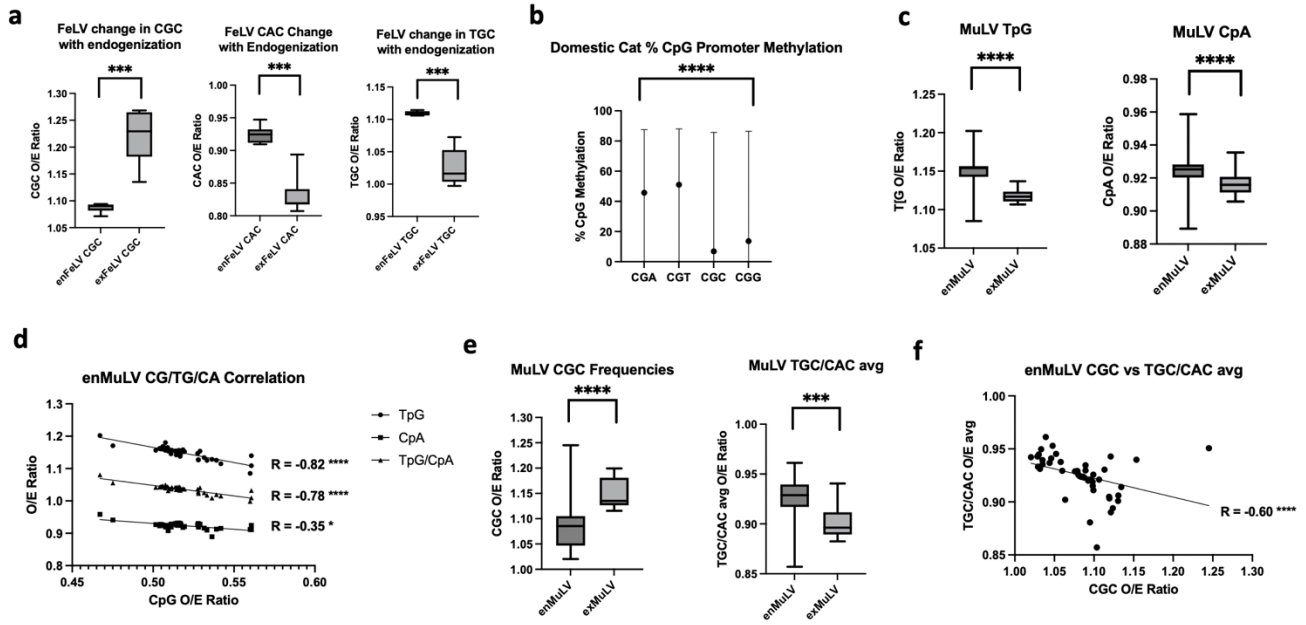


Figure 3: Endogenization of gammaretroviruses impacts retroviral CpG content: a) FeLV endogenization leads to a significant decrease in CGC content, with corresponding increases in CAC and TGC content. **b)** Domestic cat promoters exhibit uneven median CpG methylation across trinucleotides, with CGC being the least methylated. **c)** Endogenous MuLV exhibits significantly increased TpG and CpA content compared to exogenous variants. **d)** enMuLV shows a strong negative correlation between CpG, TpG, and CpA, suggesting methylated cytosine deamination of enMuLV. **e)** EnMuLV CGC frequencies are significantly decreased compared to exMuLV, with corresponding increases in TGC/CAC frequencies. **f)** enMuLV shows a strong negative correlation between CGC and TGC/CAC, suggesting methylated cytosine deamination of enMuLV at CGC trinucleotides. P-values are reported for Mann-Whitney tests **(a, c, e)**, Kruskal-Wallis test for multiple comparisons **(b)**, and Spearman correlation **(d, f)**. Error bars for **b)** represent interquartile ranges. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$

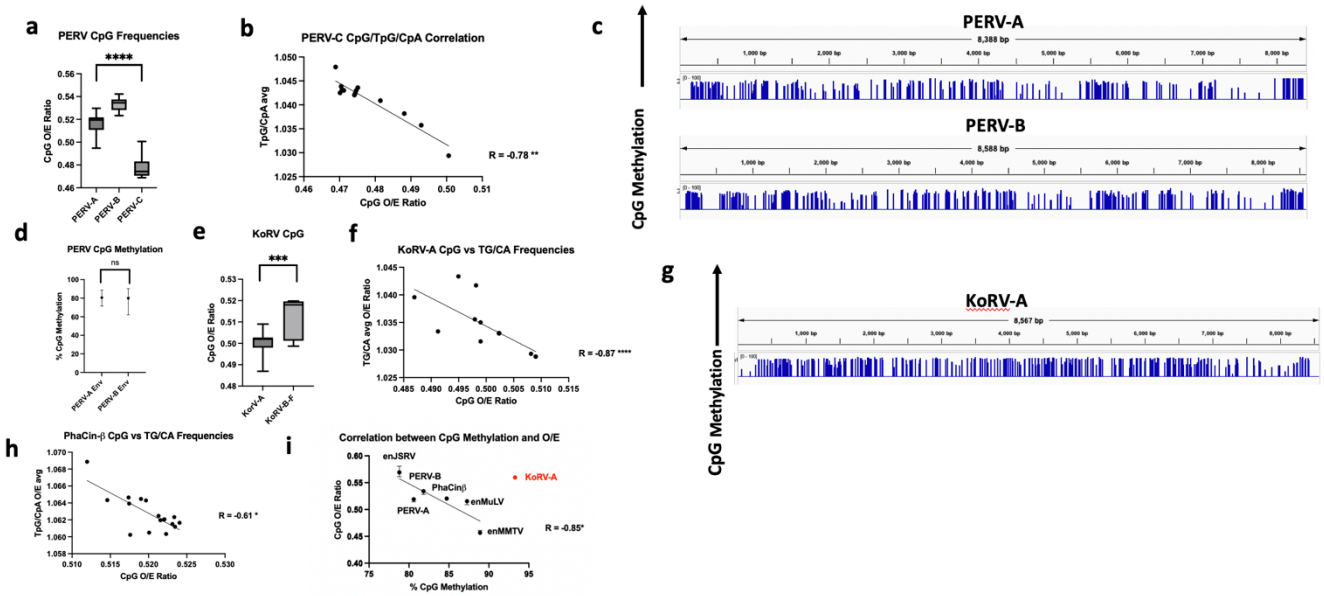


Figure 4: CpG content of endogenous retroviruses corresponds with prior CpG methylation and transcriptional activity: a) Porcine Endogenous Retrovirus (PERV) variants exhibit significant variation in CpG content. **b)** PERV-C shows a significant negative correlation between CpG with TpG/CpA content, suggesting prior methylation. **c)** Bisulfite mapping with Bismark shows both PERV-A and PERV-B are heavily methylated across the entire provirus. **d)** PERV-A and PERV-B show no significant difference in median CpG methylation content. **e)** endogenous Koala Retrovirus (KoRV-A) has significantly lower CpG content than the other KoRV subtypes. **f)** KoRV-A shows significant negative correlation between CpG and TpG/CpA frequency, suggesting impacts from DNA methylation. **g)** Bismark bisulfite mapping shows KoRV-A is heavily methylated across the entire provirus. **h)** Endogenous koala betaretrovirus phaCin-β shows a significant negative correlation between CpG and deamination products TpG/CpA. **i)** Median CpG % methylation plotted against median CpG observed/expected (O/E) ratios demonstrates a significant negative correlation between them for all endogenous retroviruses except KoRV-A (in red). Error bars in **(d)** represent interquartile ranges. P-values are reported for Kruskal-Wallis tests for multiple

comparisons **(a)**, Mann-Whitney tests **(d,e)**, Spearman correlation **(b,f,h)**, and Pearson correlation **(i)**. *$p < 0.05$, **$p < 0.01$, *** $p < 0.001$, ****$p < 0.0001$

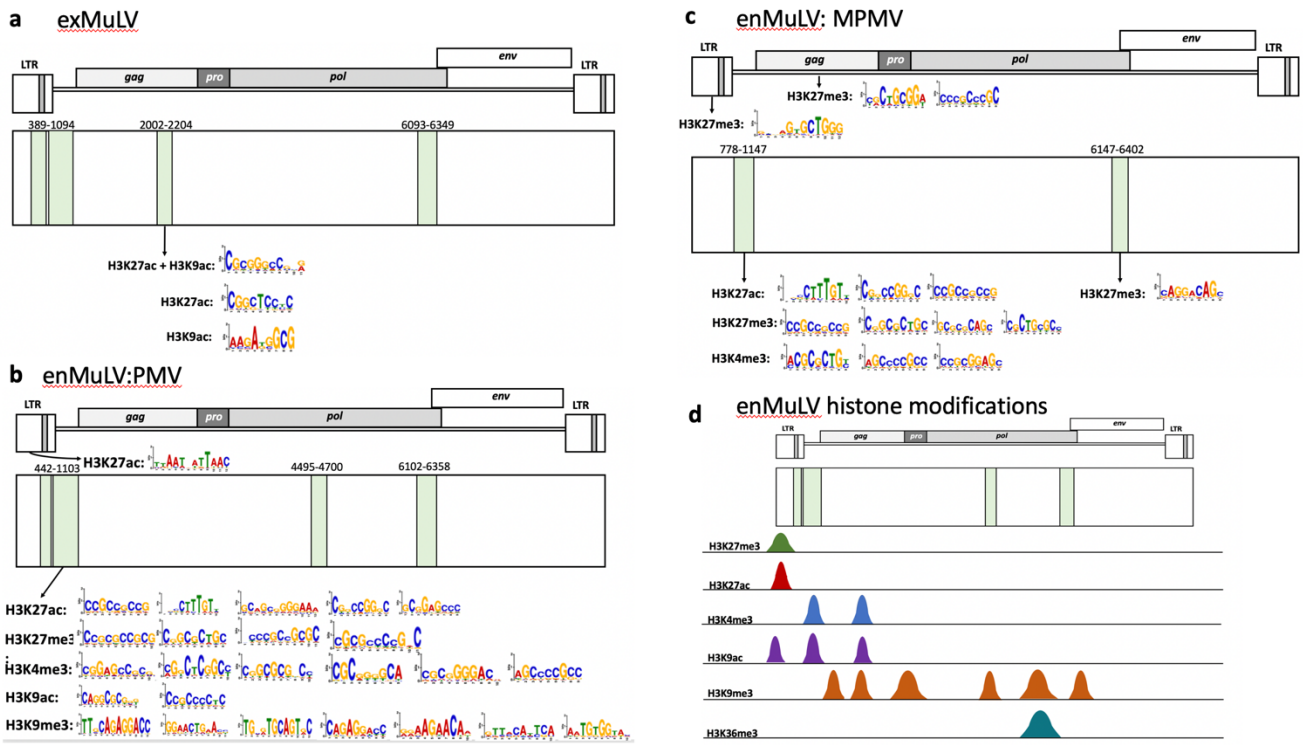


Figure 5: Murine leukemia virus (MuLV) shows histone modification motifs consistent with exogenous and variably active endogenous variants: a) Emboss Cpgplot and SEA analysis identify CpG islands (shaded in green) across the exogenous MuLV provirus, with one in particular containing H3K27ac and H3K9ac-associated motifs. **b)** Emboss Cpgplot and SEA analysis of polytropic endogenous MuLV (PMV) show heavy density of histone modification motifs within a CGI in the leader region, including those for active (H3K27ac, H3K4me3, H3K9ac) and repressive (H3K27me3 and H3K9me3) domains. **c)** Emboss Cpgplot and SEA analysis of endogenous modified polytropic endogenous MuLV (MPMV) shows widespread H3K27me3 motifs in the LTR and CGIs throughout the provirus. **d)** Visualization of ENCODE ChIP-Seq data from day 16 mouse embryos in the UCSC Genome Browser shows several histone modifications present in endogenous MuLV sequences (Figure details in **Table 3**). Further details on MuLV SEA analysis can be found in **Supplementary Table S3**.

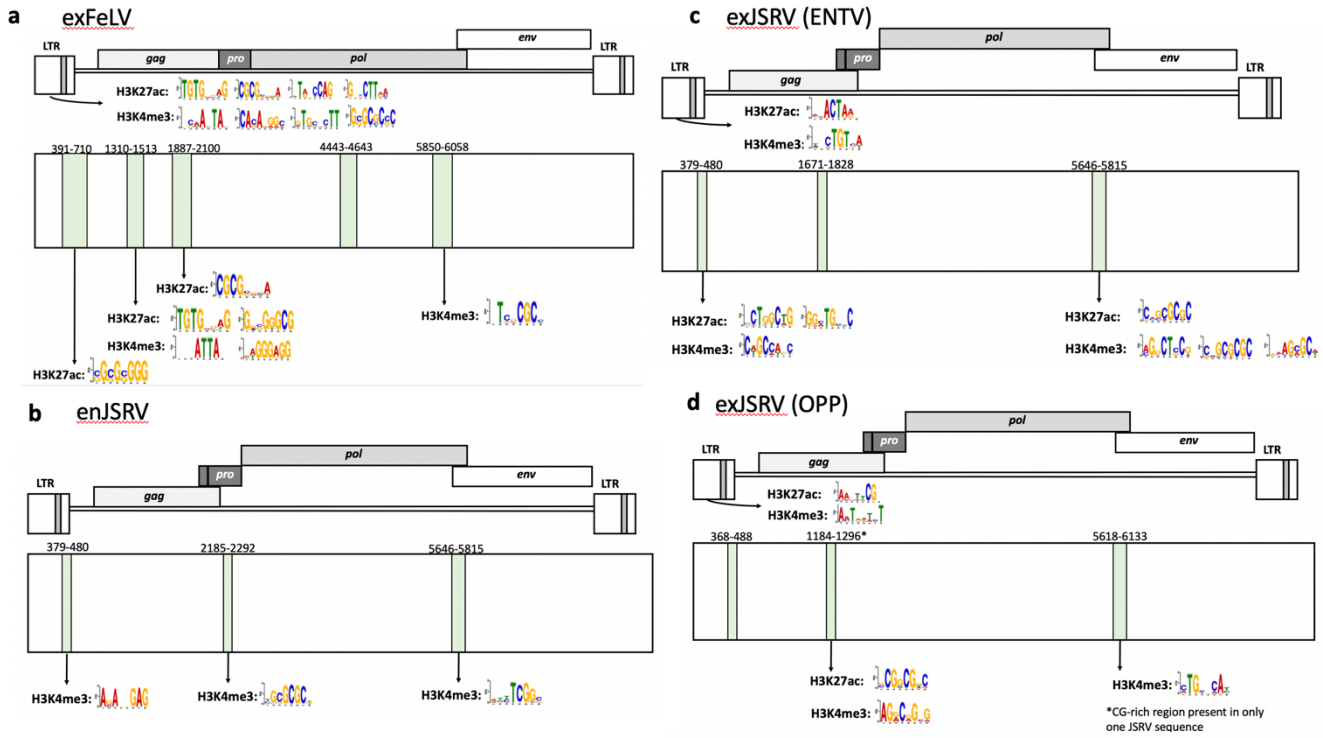


Figure 6: Exogenous Jaagsiekte Sheep Retrovirus (JSRV) and Feline Leukemia Virus (FeLV) variants contain more significant activation-associated histone modification motifs than endogenous variants. a) EMBOSS Cpgplot and SEA analysis show endogenous JSRV contains several CG-rich regions with only H3K4me3 motifs present and no H3K27ac. **b)** EMBOSS Cpgplot and SEA analysis of the exogenous Enzootic Nasal Tumor Virus (ENTV) JSRV variant show significant motifs present in the LTR, leader region, and *pol-env* junction for both H3K27ac and H3K4me3. **c)** EMBOSS Cpgplot and SEA analysis of the exogenous Ovine Progressive Pleuropneumonia (OPP) JSRV variant shows several H3K27ac and H3K4me3-associated motifs within CG-rich regions primarily in the LTR and *gag*. **d)** EMBOSS Cpgplot and SEA analysis of exogenous FeLV showed significant H3K27ac and H3K4me3 motifs in the LTR and CGIs throughout the provirus. Further details on SEA analysis can be found in **Supplementary Tables S4 and S5**.

616 References
617
618 1. Meyer TJ, Rosenkrantz JL, Carbone L, Chavez SL. Endogenous Retroviruses: With Us and
619 against Us. Front Chem. 2017;5:23. Epub 20170407. doi: 10.3389/fchem.2017.00023. PubMed PMID:
620 28439515; PMCID: PMC5384584.
621 2. Mager DL, Stoye JP. Mammalian Endogenous Retroviruses. Microbiol Spectr.
622 2015;3(1):MDNA3-0009-2014. doi: 10.1128/microbiolspec.MDNA3-0009-2014. PubMed PMID:
623 26104559.
624 3. Coffin J, Blomberg J, Fan H, Gifford R, Hatziioannou T, Lindemann D, Mayer J, Stoye J, Tristem
625 M, Johnson W, Ictv Report C. ICTV Virus Taxonomy Profile: Retroviridae 2021. J Gen Virol.
626 2021;102(12). doi: 10.1099/jgv.0.001712. PubMed PMID: 34939563; PMCID: PMC8744268.
627 4. De Rijck J, de Kogel C, Demeulemeester J, Vets S, El Ashkar S, Malani N, Bushman FD, Landuyt
628 B, Husson SJ, Busschots K, Gijsbers R, Debyser Z. The BET family of proteins targets moloney murine
629 leukemia virus integration near transcription start sites. Cell Rep. 2013;5(4):886-94. Epub 20131031.
630 doi: 10.1016/j.celrep.2013.09.040. PubMed PMID: 24183673; PMCID: PMC4197836.
631 5. Felice B, Cattoglio C, Cittaro D, Testa A, Miccio A, Ferrari G, Luzi L, Recchia A, Mavilio F.
632 Transcription factor binding sites are genetic determinants of retroviral integration in the human
633 genome. PLoS One. 2009;4(2):e4571. Epub 20090224. doi: 10.1371/journal.pone.0004571. PubMed
634 PMID: 19238208; PMCID: PMC2642719.
635 6. Jarosz AS, Halo JV. Transcription of Endogenous Retroviruses: Broad and Precise Mechanisms
636 of Control. Viruses. 2024;16(8). Epub 20240817. doi: 10.3390/v16081312. PubMed PMID: 39205286;
637 PMCID: PMC11359688.
638 7. Tarlinton R, Meers J, Hanger J, Young P. Real-time reverse transcriptase PCR for the
639 endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease
640 in koalas. J Gen Virol. 2005;86(Pt 3):783-7. doi: 10.1099/vir.0.80547-0. PubMed PMID: 15722540.
641 8. Hashem MA, Kayesh MEH, Maetani F, Eiei T, Mochizuki K, Ochiai S, Ito A, Ito N, Sakurai H, Asai
642 T, Tsukiyama-Kohara K. Koala retrovirus (KoRV) subtypes and their impact on captive koala
643 (Phascolarctos cinereus) health. Arch Virol. 2021;166(7):1893-901. Epub 20210426. doi:
644 10.1007/s00705-021-05078-y. PubMed PMID: 33900468.
645 9. Ngo MH, AbuEed L, Kawasaki J, Oishi N, Pramono D, Kimura T, Sakurai M, Watanabe K,
646 Mizukami Y, Ochi H, Anai Y, Odahara Y, Umehara D, Kawamura M, Watanabe S, Miyake A, Nishigaki K.
647 Multiple recombination events between endogenous retroviral elements and feline leukemia virus. J
648 Virol. 2024;98(2):e0140023. Epub 20240119. doi: 10.1128/jvi.01400-23. PubMed PMID: 38240589;
649 PMCID: PMC10878261.
650 10. van der Kuyl AC, Dekker JT, Goudsmit J. Discovery of a new endogenous type C retrovirus
651 (FcEV) in cats: evidence for RD-114 being an FcEV(Gag-Pol)/baboon endogenous virus BaEV(Env)
652 recombinant. J Virol. 1999;73(10):7994-8002. doi: 10.1128/JVI.73.10.7994-8002.1999. PubMed PMID:
653 10482547; PMCID: PMC112814.
654 11. Du Toit A. Leaving a mark on retroviruses. Nat Rev Microbiol. 2021;19(3):138-9. doi:
655 10.1038/s41579-021-00514-x. PubMed PMID: 33469167.
656 12. Letchumanan P, Theva Das K. The role of genetic diversity, epigenetic regulation, and sex-
657 based differences in HIV cure research: a comprehensive review. Epigenetics Chromatin.
658 2025;18(1):1. Epub 20250103. doi: 10.1186/s13072-024-00564-4. PubMed PMID: 39754177; PMCID:
659 PMC11697457.

660 13. Alinejad-Rokny H, Anwar F, Waters SA, Davenport MP, Ebrahimi D. Source of CpG Depletion in
661 the HIV-1 Genome. Mol Biol Evol. 2016;33(12):3205-12. Epub 20160928. doi:
662 10.1093/molbev/msw205. PubMed PMID: 27682824.
663 14. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res.
664 2015;43(W1):W39-49. Epub 20150507. doi: 10.1093/nar/gkv416. PubMed PMID: 25953851; PMCID:
665 PMC4489269.
666 15. Johnson BJ, Robbins A, Gyawali N, Ong O, Loader J, Murphy AK, Hanger J, Devine GJ. The
667 environmental and ecological determinants of elevated Ross River Virus exposure in koalas residing in
668 urban coastal landscapes. Sci Rep. 2021;11(1):4419. Epub 20210224. doi: 10.1038/s41598-021-
669 83919-1. PubMed PMID: 33627779; PMCID: PMC7904799.
670 16. Upadhyay M, Samal J, Kandpal M, Vasaikar S, Biswas B, Gomes J, Vivekanandan P. CpG
671 dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. J
672 Virol. 2013;87(24):13816-24. Epub 20131009. doi: 10.1128/JVI.02515-13. PubMed PMID: 24109231;
673 PMCID: PMC3838256.
674 17. Arnaud F, Varela M, Spencer TE, Palmarini M. Coevolution of endogenous betaretroviruses of
675 sheep and their host. Cell Mol Life Sci. 2008;65(21):3422-32. doi: 10.1007/s00018-008-8500-9.
676 PubMed PMID: 18818869; PMCID: PMC4207369.
677 18. Jern P, Stoye JP, Coffin JM. Role of APOBEC3 in genetic diversity among endogenous murine
678 leukemia viruses. PLoS Genet. 2007;3(10):2014-22. Epub 20070910. doi:
679 10.1371/journal.pgen.0030183. PubMed PMID: 17967065; PMCID: PMC2041998.
680 19. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, Neuberger MS,
681 Malim MH. DNA deamination mediates innate immunity to retroviral infection. Cell. 2003;113(6):803-
682 9. doi: 10.1016/s0092-8674(03)00423-9. PubMed PMID: 12809610.
683 20. Lapp Z, Yoon H, Foley B, Leitner T. Hypermut 3: identifying specific mutational patterns in a
684 defined nucleotide context that allows multistate characters. Bioinform Adv. 2025;5(1):vbaf025. Epub
685 20250210. doi: 10.1093/bioadv/vbaf025. PubMed PMID: 39981109; PMCID: PMC11842049.
686 21. Chen Y. CM, Duan X., Cui J. Ancient origin and complex evolution of porcine endogenous
687 retroviruses. Biosafety and Health. 2020;2:142-51. doi:
688 http://dx.doi.org/10.1016/j.bsheal.2020.03.003.
689 22. Yu T, Blyton MBJ, Abajorga M, Koppetsch BS, Ho S, Xu B, Hu Z, Luban J, Chappell K, Weng Z,
690 Theurkauf WE. Evolution of KoRV-A transcriptional silencing in wild koalas. Cell. 2025. Epub
691 20250226. doi: 10.1016/j.cell.2025.02.006. PubMed PMID: 40056902.
692 23. Lillie M, Hill J, Pettersson ME, Jern P. Expansion of a retrovirus lineage in the koala genome.
693 Proc Natl Acad Sci U S A. 2022;119(25):e2201844119. Epub 20220613. doi:
694 10.1073/pnas.2201844119. PubMed PMID: 35696585; PMCID: PMC9231498.
695 24. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite.
696 Trends Genet. 2000;16(6):276-7. doi: 10.1016/s0168-9525(00)02024-2. PubMed PMID: 10827456.
697 25. Stocking C, Kozak CA. Murine endogenous retroviruses. Cell Mol Life Sci. 2008;65(21):3383-98.
698 doi: 10.1007/s00018-008-8497-0. PubMed PMID: 18818872; PMCID: PMC4802364.
699 26. Antequera F. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci.
700 2003;60(8):1647-58. doi: 10.1007/s00018-003-3088-6. PubMed PMID: 14504655; PMCID:
701 PMC11138798.

702 27. Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-
703 host evolution. Proc Natl Acad Sci U S A. 2013;110(50):20146-51. Epub 20131125. doi:
704 10.1073/pnas.1315419110. PubMed PMID: 24277832; PMCID: PMC3864273.
705 28. Hayward JA, Tachedjian M, Kohl C, Johnson A, Dearnley M, Jesaveluk B, Langer C, Solymosi PD,
706 Hille G, Nitsche A, Sanchez CA, Werner A, Kontos D, Crameri G, Marsh GA, Baker ML, Poumbourios P,
707 Drummer HE, Holmes EC, Wang LF, Smith I, Tachedjian G. Infectious KoRV-related retroviruses
708 circulating in Australian bats. Proc Natl Acad Sci U S A. 2020;117(17):9529-36. Epub 20200413. doi:
709 10.1073/pnas.1915400117. PubMed PMID: 32284399; PMCID: PMC7196810.
710 29. Wang GZ, Goff SP. Moloney Murine Leukemia Virus p12 Is Required for Histone Loading onto
711 Retroviral DNAs. J Virol. 2021;95(15):e0049521. Epub 20210712. doi: 10.1128/JVI.00495-21. PubMed
712 PMID: 34011543; PMCID: PMC8274600.
713 30. Miklik D, Slavkova M, Kucerova D, Mekadim C, Mrazek J, Hejnar J. Long Terminal Repeats of
714 Gammaretroviruses Retain Stable Expression after Integration Retargeting. Viruses. 2024;16(10).
715 Epub 20240925. doi: 10.3390/v16101518. PubMed PMID: 39459853; PMCID: PMC11512309.
716 31. Desfarges S, Ciuffi A. Retroviral integration site selection. Viruses. 2010;2(1):111-30. Epub
717 20100112. doi: 10.3390/v2010111. PubMed PMID: 21994603; PMCID: PMC3185549.
718 32. Jahner D, Stuhlmann H, Stewart CL, Harbers K, Lohler J, Simon I, Jaenisch R. De novo
719 methylation and expression of retroviral genomes during mouse embryogenesis. Nature.
720 1982;298(5875):623-8. doi: 10.1038/298623a0. PubMed PMID: 6285203.
721 33. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-
722 body DNA methylation. Oncotarget. 2012;3(4):462-74. doi: 10.18632/oncotarget.497. PubMed PMID:
723 22577155; PMCID: PMC3380580.
724 34. Mao SQ, Cuesta SM, Tannahill D, Balasubramanian S. Genome-wide DNA Methylation
725 Signatures Are Determined by DNMT3A/B Sequence Preferences. Biochemistry. 2020;59(27):2541-50.
726 Epub 20200628. doi: 10.1021/acs.biochem.0c00339. PubMed PMID: 32543182; PMCID:
727 PMC7364778.
728 35. Mallona I, Ilie IM, Karemaker ID, Butz S, Manzo M, Caflisch A, Baubec T. Flanking sequence
729 preference modulates de novo DNA methylation in the mouse genome. Nucleic Acids Res.
730 2021;49(1):145-57. doi: 10.1093/nar/gkaa1168. PubMed PMID: 33290556; PMCID: PMC7797059.
731 36. Gao L, Emperle M, Guo Y, Grimm SA, Ren W, Adam S, Uryu H, Zhang ZM, Chen D, Yin J, Dukatz
732 M, Anteneh H, Jurkowska RZ, Lu J, Wang Y, Bashtrykov P, Wade PA, Wang GG, Jeltsch A, Song J.
733 Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de
734 novo DNA methylation mechanisms. Nat Commun. 2020;11(1):3355. Epub 20200703. doi:
735 10.1038/s41467-020-17109-4. PubMed PMID: 32620778; PMCID: PMC7335073.
736 37. Alvarez-Ponce D, Torres-Sanchez M, Feyertag F, Kulkarni A, Nappi T. Molecular evolution of
737 DNMT1 in vertebrates: Duplications in marsupials followed by positive selection. PLoS One.
738 2018;13(4):e0195162. Epub 20180405. doi: 10.1371/journal.pone.0195162. PubMed PMID:
739 29621315; PMCID: PMC5886458.
740 38. Stewart KR, Veselovska L, Kim J, Huang J, Saadeh H, Tomizawa S, Smallwood SA, Chen T,
741 Kelsey G. Dynamic changes in histone modifications precede de novo DNA methylation in oocytes.
742 Genes Dev. 2015;29(23):2449-62. Epub 20151119. doi: 10.1101/gad.271353.115. PubMed PMID:
743 26584620; PMCID: PMC4691949.

744 39. Ohtani H, Liu M, Zhou W, Liang G, Jones PA. Switching roles for DNA and histone methylation
745 depend on evolutionary ages of human endogenous retroviruses. Genome Res. 2018;28(8):1147-57.
746 Epub 20180703. doi: 10.1101/gr.234229.118. PubMed PMID: 29970451; PMCID: PMC6071641.
747 40. Boi S, Rosenke K, Hansen E, Hendrick D, Malik F, Evans LH. Endogenous retroviruses mobilized
748 during friend murine leukemia virus infection. Virology. 2016;499:136-43. Epub 20160919. doi:
749 10.1016/j.virol.2016.07.009. PubMed PMID: 27657834; PMCID: PMC5102782.
750 41. Ebrahimi D, Anwar F, Davenport MP. APOBEC3G and APOBEC3F rarely co-mutate the same
751 HIV genome. Retrovirology. 2012;9:113. Epub 20121220. doi: 10.1186/1742-4690-9-113. PubMed
752 PMID: 23256516; PMCID: PMC3532371.
753 42. Anwar F, Davenport MP, Ebrahimi D. Footprint of APOBEC3 on the genome of human
754 retroelements. J Virol. 2013;87(14):8195-204. Epub 20130522. doi: 10.1128/JVI.00298-13. PubMed
755 PMID: 23698293; PMCID: PMC3700199.
756 43. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D,
757 Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across
758 taxonomic space. Database (Oxford). 2011;2011:bar030. Epub 20110723. doi:
759 10.1093/database/bar030. PubMed PMID: 21785142; PMCID: PMC3170168.
760 44. Krueger F. Trim Galore! Available from:
761 https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
762 45. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq
763 applications. Bioinformatics. 2011;27(11):1571-2. Epub 20110414. doi:
764 10.1093/bioinformatics/btr167. PubMed PMID: 21493656; PMCID: PMC3102221.
765 46. Ngo V, Chen Z, Zhang K, Whitaker JW, Wang M, Wang W. Epigenomic analysis reveals DNA
766 motifs regulating histone modifications in human and mouse. Proc Natl Acad Sci U S A.
767 2019;116(9):3668-77. Epub 20190212. doi: 10.1073/pnas.1813565116. PubMed PMID: 30755522;
768 PMCID: PMC6397554.
769 47. Chen S, Liu S, Shi S, Jiang Y, Cao M, Tang Y, Li W, Liu J, Fang L, Yu Y, Zhang S. Comparative
770 epigenomics reveals the impact of ruminant-specific regulatory elements on complex traits. BMC Biol.
771 2022;20(1):273. Epub 20221208. doi: 10.1186/s12915-022-01459-0. PubMed PMID: 36482458;
772 PMCID: PMC9730597.
773 48. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN,
774 Hinrichs AS, Lee CM, Nassar LR, Raney BJ, Speir ML, van Baren MJ, Vaske CJ, Haussler D, Kent WJ,
775 Haeussler M. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res.
776 2025;53(D1):D1243-D9. doi: 10.1093/nar/gkae974. PubMed PMID: 39460617; PMCID: PMC11701590.
777 49. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64. doi:
778 10.1101/gr.229202. PubMed PMID: 11932250; PMCID: PMC187518.
779 50. Roca AL, Pecon-Slattery J, O'Brien SJ. Genomically intact endogenous feline leukemia viruses
780 of recent origin. J Virol. 2004;78(8):4370-5. doi: 10.1128/jvi.78.8.4370-4375.2004. PubMed PMID:
781 15047851; PMCID: PMC374283.
782 51. Ishida Y, Zhao K, Greenwood AD, Roca AL. Proliferation of endogenous retroviruses in the
783 early stages of a host germ line invasion. Mol Biol Evol. 2015;32(1):109-20. Epub 20140925. doi:
784 10.1093/molbev/msu275. PubMed PMID: 25261407; PMCID: PMC4271524.
785 52. Parisi F, Freer G, Mazzanti CM, Pistello M, Poli A. Mouse Mammary Tumor Virus (MMTV) and
786 MMTV-like Viruses: An In-depth Look at a Controversial Issue. Viruses. 2022;14(5). Epub 20220506.
787 doi: 10.3390/v14050977. PubMed PMID: 35632719; PMCID: PMC9147501.

788 53. Lillie M, Pettersson M, Jern P. Contrasting segregation patterns among endogenous
789 retroviruses across the koala population. Commun Biol. 2024;7(1):350. Epub 20240321. doi:
790 10.1038/s42003-024-06049-0. PubMed PMID: 38514810; PMCID: PMC10957985.
791