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 (210 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 genera. Beta- and gammaretroviruses are two types of simple retroviruses that 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 motifs associated with 'active' histone modifications. 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.

Importance

Expression of vertebrate genes is regulated by chemical modifications made directly to the DNA or to the proteins associated with it, termed epigenetics. Because retroviruses integrate into DNA, they are subject to the same epigenetic modifications as regular genes. Retroviruses will tend to endogenize, meaning they will become a permanent part of a species' genome when their hidden DNA is passed down to progeny during reproduction. However, sometimes it is difficult to discern whether a retroviral sequence is endogenous (permanently fixed) or exogenous (an infectious entity). We hypothesized that changes to the retroviral sequences over time after endogenization would result from epigenetic modifications, and that these changes could help distinguish an endogenous retrovirus from an exogenous one. In this paper, we show that changes to the viral sequences associated with epigenetics indeed take place after endogenization.

Introduction

Retroviruses are a unique family 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

75 progeny. When retroviruses infect germ cells, they can become incorporated into every cell of
76 the resultant offspring, a process known as endogenization. Following endogenization,
77 retroviruses will tend to acquire mutations rendering their promoter sequences and/or gene
78 products deficient over the course of generations. Therefore, the degree of mutation of an
79 endogenous retroviral sequence can give a window into how long it has been endogenized in
80 a species.[1] Over 40% of mammalian genomes are reported to be products of reverse
81 transcription, including LINEs and SIINEs, with 8-10% derived from endogenous
82 retroviruses.[2]
83
84 Retroviruses are divided into seven main genera based upon their genome organization and
85 structures.[3] These genera exhibit different propensities for the genomic regions into which
86 they tend to integrate. For example, lentiviruses, which include human immunodeficiency
87 virus (HIV) and feline immunodeficiency virus (FIV), have a tendency to integrate into
88 transcriptionally active gene bodies. On the other hand, gammaretroviruses, which include
89 murine leukemia virus (MuLV) and feline leukemia virus (FeLV), tend to target CpG-rich gene
90 promoters.[4] Betaretroviruses, such as mouse mammary tumor virus (MMTV) and Jaagsiekte
91 sheep retrovirus (JSRV), have shown a weak propensity for transcriptional start sites (TSS),[5]
92 but generally seem to have a more random distribution of integration.[6] Gammaretroviruses
93 and betaretroviruses are much more prone than lentiviruses to infecting germ cells and
94 endogenizing, as evidenced by their retroviral remnants being present in every major
95 mammalian genome examined, including HERV-K (betaretrovirus) and HERV-W
96 (gammaretrovirus) in humans. Interestingly, endogenous beta- and gammaretroviruses can
97 frequently have exogenous counterparts that are fully infectious and replication-competent.

98 Their sequences are often only slightly divergent from each other, usually in the long terminal
99 repeat (LTR) and/or envelope (env), impacting their expression and tropism, respectively.[7]
100
101 Endogenous and exogenous retroviral variants are well-characterized for some species, such
102 as MuLV and FeLV. However, the distinction between recent integrants with transcriptionally
103 intact genomes in more poorly characterized species can be difficult. A case in point is the
104 koala retrovirus, KoRV, which is a gammaretrovirus in the process of endogenizing. The virus
105 has been associated with leukemias and lymphomas, as well as with secondary infections
106 such as chlamydia.[8] Thus far, ten different subtypes have been described, varying with
107 geography and exhibiting different associations with disease. KoRV-A appears to be the
108 endogenous variant that has arisen from this virus, while the other subtypes exhibit more
109 overt associations with disease and are believed to be exogenous.[9]
110
111 To make matters more confusing, endogenous and exogenous retroviruses will recombine.
112 The endogenous variant of FeLV can recombine with the exogenous variant FeLV-A to yield
113 a new virus with different tropism, and recombination events can even occur with other types
114 of less related endogenous retroviruses, giving rise to diversity and evolution within the
115 genera.[10] Most of these recombination events result from viruses within a single or closely
116 related species, but there are examples of recombination between gammaretroviruses from
117 more divergent species. For example, RD114, another endogenous gammaretrovirus found
118 in domestic cats, appears to be the product of a recombination event between an
119 endogenous feline retrovirus (*Felis catus* endogenous retrovirus, or FcEV) and baboon

120 endogenous retrovirus (BaEV) and is believed to have resulted from a cross-species viral
121 transfer from an old-world monkey to a domestic cat.[11]
122
123 Multiple host epigenetic mechanisms used to suppress endogenous retroviral infections have
124 been described, including DNA methylation, histone modifications, and non-coding RNAs.[12]
125 Exogenous retroviruses can also be subject to these same mechanisms as a molecular arm
126 of the host antiviral defense, making them an active area of study in HIV.[13] However,
127 endogenous retroviruses that have co-evolved with their host species are frequently more
128 highly transcriptionally suppressed than their exogenous counterparts,[7] prompting the
129 hypothesis that sequence-specific evidence of epigenetic suppression will generally be more
130 apparent in endogenous variants.
131
132 Previously, HIV was reported to mirror the human genome with its low CpG content,
133 exhibiting evidence that this was primarily a product of deamination following CpG
134 methylation rather than other possible mechanisms, such as CpG depletion by APOBEC3 or
135 due to toll-like receptors.[14] Because of their propensity to integrate into promoter regions, we
136 hypothesized that gammaretroviruses would mirror the CpG content of the promoters in their
137 respective species, and originally, we also hypothesized that CpG content of betaretroviral
138 genomes might mirror that of promoters due to a TSS integration bias. We also hypothesized
139 that endogenous variants would exhibit decreased CpG content relative to exogenous
140 variants, suggesting more CpG methylation and deamination following endogenization, as
141 well as altered CpG trinucleotide motifs consistent with the methylation patterns in their host
142 species. Finally, we hypothesized that motifs consistent with active histone modifications

143 would be more apparent in exogenous proximal proviral sequences, while those more
144 consistent with repressive modifications (e.g. H3K9me3) would be more evident in known
145 endogenous sequences. To test these hypotheses, we performed statistical analyses on CpG
146 frequencies and CpG trinucleotide motif frequencies from various retroviruses and compared
147 endogenous to exogenous variants, as well as all those from promoter sequences of the
148 species in which they are present. We also performed bisulfite mapping to examine
149 methylation patterns in endogenous retroviruses versus promoters. Finally, we examined
150 differences in motifs associated with histone modifications using the Simple Enrichment
151 Analysis tool from the MEME suite.[15] Because betaretroviruses have different integration
152 propensities than gammaretroviruses, we focused on characteristics for the few viruses we
153 were able to examine in each genus, as their integration behavior potentially impacts their
154 epigenetic regulation.
155
156 **Results**
157 *Recently endogenized beta- and gammaretroviruses reflect CpG content of eutherian host*
158 *promoters, with variability attributable to DNA methylation*
159
160 To evaluate how betaretroviruses and gammaretroviruses adapt to the genomes of their
161 respective species, we examined CpG content of full-length endogenous and exogenous
162 retroviruses from cats, mice, pigs, sheep, and koalas from **Tables 1 and 2** (see **Data**
163 **Availability Statement** for accession numbers) using a Python-based Kmer tool. The Kmer
164 tool calculates D-ratios, which are the observed/expected (or O/E ratios) based on a first
165 order Markov model of conditional probabilities (see **Methods**). We assessed the CpG

content of endogenous retroviruses relative to multiple genomic regions (500 bp upstream of the TSS, 1kb upstream of the TSS, 2kb upstream of the TSS, whole gene bodies, 1kb downstream of all genes, and entire chromosomes) in all five species **(Supplementary Figures S1 and S2)**.

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

Gammaretroviruses						
	Exogenous vs endogenous % identity	**Original Integration time**	**Exogenous CpG O/E**	**Endogenous CpG O/E**	**Exog. N**	**Endog. N**
FeLV	86%	<2 MYA[16]	0.53	0.49	8	8
MuLV	90-100%	1 MYA[17]	0.53	0.52	16	49
KoRV	90%	22,000-49,000 YA[18]	0.51	0.5	8	16
Betaretroviruses						
MMTV	95-99%	20 MYA[19]	0.45	0.46	6	5
JSRV	90-98%	5-7 MYA[20]	0.58	0.57	20	19

In most cases, ERVs reflected the CpG content of regions 500 bp upstream of the TSS, especially for the gammaretroviruses we evaluated **(Supplementary Figure S1a-d).** The betaretroviruses we evaluated also trended this way, although enJSRV in particular had a higher CpG O/E ratio relative to the other endogenous retroviruses we examined **(Supplementary Figure S2a-c)**. However, the five different mammalian species exhibited significantly different CpG O/E ratios from each other in promoter regions (defined in **Supplementary S2d** as 1 kb upstream of the TSS), suggesting that while endogenous beta- and gammaretroviruses averaged CpG O/E ratios around 0.5, these could deviate significantly from promoter CpG content, depending on the species they infect. Cats in

179 particular had overall higher CpG content across all regions examined **(Supplementary**

180 **Figure S1a**, **Supplementary Figure S2d)**, while sheep, koalas, and mice exhibited lower

181 relative CpG content closer to 0.4 within 1kb of the TSS **(Supplementary Figure S2d)**.

Table 2: Further endogenous retroviruses used for analysis.

Gammaretroviruses			
	Original Integration time	**CpG O/E**	**Endog. N**
PERV-A	6.6 MYA[21]	0.52	8
PERV-B	6.4 MYA[21]	0.53	14
PERV-C	3.2 MYA[21]	0.48	14
Betaretroviruses			
phaCin-β	0.7-1.5 MYA[22]	0.52	18

182

183 Based on our analysis, we compared CpG content of endogenous and exogenous beta- and

184 gammaretroviruses in each species to that of the region within 500 bp upstream of the TSS

185 **(Figure 1a-e)**. Mice and sheep both displayed significant differences in CpG content of their

186 promoters from their respective retroviruses, reflecting their overall lower promoter CpG

187 content. However, three out of the five species examined had no significant difference

188 between their retroviral CpG content and their promoters. To rule out the potential impact of

189 ancient retroviral sequences inhabiting promoters, we used also LTRHarvest[23] to remove all

190 LTR and retrotransposon-like sequences and re-analyzed CpG content using the Kmer tool,

191 which resulted in a minimal change to the CpG content of the analyzed promoter sequences

192 (data not shown).

193

194 *Endogenization impacts CpG content of beta- and gammaretroviruses due to CpG*

195 *methylation*

196

197 When comparing CpG content of endogenous vs. exogenous retroviruses, the endogenous
198 gammaretroviruses FeLV and MuLV exhibited a statistically significant decrease in CpG
199 content compared to their exogenous counterparts (**Figure 1f-g**). However, the
200 betaretroviruses analyzed (MMTV and JSRV) did not have as robust a difference with the
201 sequences available **(Figure 1h-i)**. In fact, no significant difference was noted for
202 endogenous vs. exogenous MMTV, but the distribution of CpG content for MMTV is highly
203 skewed due to very few full-length replicates being publicly available (i.e. 5 for enMMTV and
204 6 for exMMTV), making concrete conclusions difficult to draw. However, JSRV had many
205 more replicates available, with 19 replicates for endogenous JSRV and 20 replicates for
206 exogenous JSRV. JSRV has two exogenous variants, one which causes enzootic nasal
207 tumor virus (ENTV), and another which causes ovine pulmonary adenocarcinoma (OPA).
208 Only the ENTV version of JSRV differed significantly in CpG content compared to the
209 endogenous variant, despite our own phylogenetic analysis of full-length sequences
210 suggesting it to be more closely related to the endogenous JSRV variant than the exogenous
211 OPA variant is (**Supplementary Figure S3**).
212 To examine the role of DNA methylation in the CpG content of all retroviruses examined, we
213 calculated the correlation between CpG, TpG, and CpA O/E ratio to assess for deamination
214 of methylated cytosines. Deamination of methylated cytosines occurs over time, changing the
215 CpG landscape to TpG and CpA on the complementary strand, while non-methylated
216 cytosines are not subject to these same forces.[24] Strikingly, both the endogenous and
217 exogenous JSRV showed a strong significant symmetrical negative correlation **(Figure 2a-c)**.
218 The correlation for exogenous JSRV was surprisingly more profound than that found in the

219 endogenous variant. Of the endogenous retroviruses examined that had a closely related
220 exogenous variant, JSRV has been one of the longest circulating, prompting speculation that
221 perhaps long-standing integrants have relaxed epigenetic control. However, endogenous
222 JSRV showed heavy methylation across the provirus, with a preference for CGC and CGG
223 motifs **(Figure 2d-e)**. Consistent with the predilection for CGC methylation, CGC motifs had
224 the strongest correlation with TGC/CAC **(Figure 2f)**, while endogenous JSRV showed a
225 significant decrease in CGC O/E ratios compared to both exogenous JSRV subtypes **(Figure**
226 **2g)**. MMTV had few full-length replicates, three of which had identical sequences, calling into
227 question the statistical power of the analysis for this virus. However, Bismark bisulfite
228 mapping also demonstrated heavy CpG methylation across the body of the endogenous
229 MMTV provirus, with CGC trinucleotides exhibiting a small but significantly higher percent
230 methylation **(Figure 2h-i)**. CGC O/E ratios did not exhibit significant differences between
231 endogenous and exogenous MMTV (data not shown), but few replicates were available for
232 this analysis, making it difficult to achieve significance.
233
234 Correlations between CpG and TpG/CpA in the two gammaretroviruses we examined were
235 less pronounced; FeLV, which only included 8 replicates of endogenous and exogenous full-
236 length sequences each that were publicly available, showed no significant negative
237 correlation, and differences in TpG and CpA O/E ratios were not significant between
238 endogenous and exogenous variants. However, O/E ratios of the trinucleotide motif CGC
239 showed a significant decrease in enFeLV compared to exFeLV, while TGC and CAC were
240 significantly increased **(Figure 3a)**, suggesting this trinucleotide motif in particular is prone to
241 DNA methylation in the context of enFeLV. Bisulfite mapping to cat promoters (defined as

242 1kb upstream of all transcriptional start sites) showed significantly lower CGC and CGG
243 methylation **(Figure 3b)**, suggesting this pattern is inconsistent with that seen in cat
244 promoters. Unfortunately, mapping the bisulfite converted cat genome to enFeLV failed to
245 produce robust results, preventing definitive conclusions about the methylation patterns in
246 enFeLV and their relationship to that of promoters in the domestic cat.
247
248 More replicate sequences of endogenous and exogenous MuLV were available for analysis
249 from [25], allowing a comprehensive examination of this gammaretrovirus. We analyzed
250 promoters from three endogenous variants: polytropic (PMV), modified polytropic (MPMV),
251 and xenotropic (XMV) MuLV. In addition to a significant decrease in CpG O/E ratios between
252 enMuLV and exMuLV **(Figure 1g)**, enMuLV also exhibited a significant increase in TpG and
253 CpA O/E compared to exMuLV, suggesting a likely role by DNA methylation **(Figure 3c)**. A
254 significant correlation was also observed between CpG O/E and TpG/CpA O/E ratios in
255 enMuLV, providing further support for this conclusion **(Figure 3d)**. Bisulfite analysis of mouse
256 promoters showed no significant difference in frequency of trinucleotide motifs methylated in
257 mouse promoters, and enMuLV also showed no difference **(Supplementary Figure S4)**.
258 CGG trinucleotides showed no significant difference in frequency between enMuLV and
259 exMuLV, and while CCG trinucleotides did show a difference, CCG frequency negatively
260 correlated only with CCA and not CTG **(Supplementary Figure S5a-e)**, suggesting
261 APOBEC3G was a more likely cause of the asymmetric deamination.[26] Analysis of all
262 enMuLV sequences using Hypermut 3[27] demonstrated significant hypermutation due to
263 APOBEC3G for all but three of the 49 sequences examined **(Supplementary Table S1)**,
264 providing further support for this. Of note, the paper from which these enMuLV sequences

265 were retrieved discusses the role of APOBEC3 in genetic diversity among enMuLVs.[25]
266 Interestingly, as was true for FeLV, a significant reduction was found in CGC trinucleotide
267 motifs in the endogenous MuLVs compared to exMuLVs, while a significant increase in
268 corresponding TGC and CAC frequency was also observed **(Figure 3e)**. A modest but
269 significant negative correlation between CGC with TGC and CAC was also found in enMuLV
270 **(Figure 3f)**, supporting the CGC motif as a more frequent past target for DNA
271 methyltransferases in enMuLV. Unlike the methylation patterns in the other endogenous
272 retroviruses, MuLV did not exhibit a significant preference for CGC methylation, nor did
273 mouse promoters **(Supplementary Figure S4)**, calling into question whether CpG
274 methylation is responsible in this case. However, these CpG frequency patterns might reflect
275 initial methylation following integration rather than following endogenization. In all cases
276 comparing endogenous and exogenous variants, there was evidence of a predilection for
277 CGC methylation following endogenization, either via available methylation data, CGC O/E
278 data, or both.
279
280 *CpG content of endogenous retroviruses is variably impacted by DNA methylation and*
281 *corresponds with transcriptional activity*
282
283 Pigs have three highly related endogenous gammaretroviruses known as Porcine
284 Endogenous Retrovirus A, B, and C. The viral sequences are primarily different in their
285 envelope regions, as well as some difference in the LTRs that appear to confer varying levels
286 of transcriptional activity. PERV-A is the oldest, having endogenized approximately 6.6 million
287 years ago. PERV-B is close in age, having endogenized approximately 6.4 million years ago.

288 Finally, PERV-C is the most divergent of the three and is thought to have endogenized 3.2
289 million years ago.[28] Of the three, PERV-B is the most transcriptionally active, while PERV-C
290 is the least, and the CpG contents of the three proviruses reflect this, with PERV-B exhibiting
291 the highest CpG O/E ratio, and PERV-C exhibiting the lowest **(Figure 4a)**. Of the three
292 proviruses, only PERV-C demonstrates a correlation between CpG and TpG/CpA O/E ratios
293 **(Figure 4b)**, suggesting this provirus is the most heavily impacted by DNA methylation. To
294 attempt to confirm this finding, we mapped a bisulfite converted pig genome to full-length
295 PERV-A, PERV-B, and PERV-C. The PERV-C envelope had no reads mapping, suggesting
296 this variant was not present in the publicly available bisulfite converted pig genome we
297 accessed. However, methylation was heavy across both PERV-A and PERV-B **(Figure 4c)**.
298 Because sequences are identical for the two variants across the LTR, gag, and pol, we
299 compared CpG methylation across the env regions of the two subtypes, demonstrating that
300 CpG methylation was comparable between the two at approximately 80% **(Figure 4d)**.
301
302 The koala gammaretrovirus, KoRV, is the youngest known endogenous retrovirus, as
303 evidence points to it being in the midst of endogenizing. KoRV-A was the first subtype
304 discovered and appears to be the most commonly endogenized variant, showing less
305 association with disease than other subtypes. It is not entirely clear whether the other
306 subtypes are endogenous or exogenous, but subtypes B-F are believed to be primarily
307 exogenous.[9] The CpG patterns of the other subtypes suggest they are less impacted by
308 DNA methylation and potentially more transcriptionally active. For example, KoRV-A exhibits
309 a significantly lower frequency of CpG dinucleotides than the other KoRV subtypes **(Figure**
310 **4e)**. Additionally, KoRV-A shows a significant correlation between CpG and TpG/CpA, once

311 again suggesting DNA methylation plays a more significant role for this subtype **(Figure 4f)**,
312 while no significant correlation was found between CpG O/E ratios and TpG/CpA ratios of the
313 other subtypes (data not shown). Mapping of bisulfite converted koala DNA against the
314 KoRV-A sequence showed heavy DNA methylation across the provirus in somatic liver
315 tissue, which was also recently reported by [29] for testicular tissue **(Figure 4g)**, reinforcing the
316 likely impact of DNA methylation upon the CpG composition of this provirus. Interestingly, the
317 endogenous koala betaretrovirus, phaCin-β,[30] also exhibited evidence for past DNA
318 methylation, with a significant correlation between CpG O/E ratios with TpG/CpA ratios
319 **(Figure 4h)**. Calculation of median %CpG methylation from mapping of bisulfite converted
320 genomes for all endogenous retroviruses examined showed a significant correlation to
321 median CpG O/E ratios, with KoRV-A proving to be a major outlier **(Figure 4i)**, perhaps
322 reflecting its recent integration into the koala genome, as CpG frequencies have not yet
323 decreased to the same extent as older endogenous retroviruses.
324
325 *Exogenous retroviruses exhibit more motifs associated with 'active' histone modifications*
326 *than endogenous variants*
327
328 To evaluate differences in histone modifications across endogenous and exogenous
329 retroviruses, we used Simple Enrichment Analysis (SEA) from the MEME suite[15] to compare
330 endogenous to exogenous proviral sequences and discover the most commonly enriched
331 motifs in each. For this analysis we compared each of two types of endogenous MuLV
332 sequences (modified polytropic (MPMV) and polytropic (PMV) to exogenous MuLV
333 sequences. We also compared endogenous FeLV to exogenous FeLV-A, and endogenous

334 JSRV to the two types of exogenous JSRV (Enzootic Nasal Tumor Virus, or ENTV, and
335 Ovine Pulmonary Adenocarcinoma, or OPA).
336
337 Because mice have considerably more genome-wide histone modifications publicly available,
338 a more thorough characterization of motifs associated with histone modifications across
339 MuLVs was possible. We performed CpG island (CGI) analysis with Emboss Cpgplot[31] to
340 refine results for histone modifications typically confined to promoters and CGIs. CGI analysis
341 showed that exogenous MuLV exhibited a CGI not present in the endogenous variants
342 **(Figure 5a)**. Three statistically significant motifs for H3K27ac and H3K9ac, which are histone
343 modifications associated with active transcription, were found in this CGI that were not
344 located in the endogenous MuLVs **(Figure 5a, Supplementary Table S2)**. However, in PMV,
345 which is more transcriptionally active than MPMV,[17] a considerable number of motifs
346 associated with both active (H3K9ac, H3K27ac, H3K4me3) and repressive (H3K27me3,
347 H3K9me3) modifications were concentrated in the CGI traversing the U5 region of the 5' LTR
348 and the leader region **(Figure 5b)**. Interestingly, in addition to some of the motifs associated
349 with 'active' histone modifications, MPMV had multiple repressive H3K27me3 motifs
350 throughout the provirus, including in the LTR, the CGI in the leader region, located in *gag*,
351 and also in the CGI located at the *pol-env* junction **(Figure 5c)**.
352
353 To attempt to validate these findings, we examined enMuLVs aligned with mouse genome
354 assembly MM10 in the UCSC Genome Browser with ENCODE histone modification tracks.
355 Surprisingly, 73% of enMuLV's (57/78) showed no histone modifications at all, but a few did,
356 summarized in **Table 3**. Repressive H3K9me3 and H3K27me3 were the most commonly

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

357 found modifications, with some modifications associated with active transcription, including
358 H3K27ac, H3K4me3, and H3K9ac **(Figure 5d)**. H3K27me3 and H3K27ac were most
359 commonly located in the proximal LTR, while H3K4me3 and H3K9ac were distributed among
360 CGIs and CpG-rich regions. H3K9me3 was broadly distributed across gene bodies, while
361 H3K36me3 was located in the gene body of one provirus, with H3K9me3 marking the same
362 spot. CpG methylation in embryonic stem cells and neural progenitors was usually found in
363 sparse or moderate amounts in all proviruses examined and frequently overlapped histone
364 modifications when they were present **(Table 3, Supplementary Figure S6a-b)**. However,
365 mapping of bisulfite converted DNA from liver tissue showed heavy average CpG methylation
366 across the provirus, suggesting CpG methylation might be more critical for proviral epigenetic
367 regulation after differentiation **(Supplementary Figure S6c)**, although these data represent
368 an average of proviral CpG methylation from different integrants at different sites vs. the
369 individual sites examined for histone modifications.
370
371 To examine another gammaretrovirus, we generated H3K4me3 and H3K27ac motifs for the
372 domestic cat and used SEA to assess enrichment throughout the endogenous and
373 exogenous FeLVs. As both of these marks are associated with active transcription, there was
374 a marked absence of any of these motifs meeting the relatively high significance threshold in
375 enFeLVs (e-value < 0.05). However, exFeLV-A had multiple CGIs with which these
376 modifications were associated **(Figure 6a, Supplementary Table S3)**.
377
378 We also examined betaretroviruses for significantly enriched histone motifs. SEA analysis
379 was overall unsuccessful for MMTV because too few full-length sequences were publicly

380 available. However, JSRV analysis yielded some interesting findings. Both JSRV and MMTV
381 exhibit considerably fewer conventional CGIs than the gammaretroviruses, so we narrowed
382 the minimum length to 100 bp to pick up CpG-rich regions where histone modifying enzymes
383 would be more likely to bind. All JSRV variants had short CpG stretches in the U5 region of
384 the LTR and the CGI located in the pol-env junction. Endogenous JSRV had a CpG-rich
385 region located in the *gag-pro-pol* junction, while the exogenous JSRVs had one more
386 proximally located in *gag* **(Figure 6b-d)**. A few statistically significant H3K4me3 sites were
387 found in the endogenous JSRV, but overall they were sparse **(Figure 6b, Supplementary**
388 **Table S4)**. Both exogenous JSRVs had H3K27ac and H3K4me3 motifs in the proximal LTR
389 **(Figure 6c-d)**. However, the ENTV JSRV variant had more of these motifs in the CpG-rich
390 leader region and the *pol-env* junction, while the OPA variant had significant motifs in *gag,*
391 suggesting these two exogenous variants are potentially regulated differently.
392

393 **Discussion**

394 We have shown that beta- and gammaretroviruses frequently reflect the CpG content of their
395 host promoters, but this is not always true in species where promoter CpG content is
396 naturally lower. As shown in **Figure 1**, both beta- and gammaretroviruses can reflect host
397 promoter CpG content, where in most species, ~67% of promoters will contain CpG islands.[32]
398 Interestingly, promoter CpG content varies significantly between the different species
399 examined **(Figure 1 and Supplementary Figure S2d)**, and those with lower CpG content
400 appeared more divergent from the CpG content of their respective retroviruses. In some part,
401 this could reflect retroviral evolution within other species first. KoRV in particular is
402 hypothesized to have arisen in either rodents or bats based on phylogenomic studies.[33, 34]

403 JSRV is an older betaretrovirus thought to have integrated around 5-7 million years ago, but
404 there is evidence of more recent endogenous variants having integrated only 200 years ago,
405 and it is likely that the currently circulating exogenous variants continue to endogenize.[20] The
406 strong negative correlation between CpG and TpA/CpG motifs in the exogenous variants
407 appears to reflect this evolutionary history, suggesting a long history of host CpG methylation
408 and therefore repeated integration within the species. Overall, sheep promoters exhibit low
409 CpG content compared to other species, resulting in a disparity between CpG content of
410 promoters and JSRV. Originally, we hypothesized that the CpG content of the betaretroviral
411 genomes might mirror that of host promoters due to a TSS integration bias. However, while
412 this integration preference is not supported in the literature,[6] the CpG content similarity to
413 gammaretroviruses is still striking. The viral p12 protein of gammaretroviruses is known to
414 tether the preintegration complex to nucleosomes via bromodomain and extraterminal
415 domain (BET) proteins, playing a role in both integration and in nucleosome loading onto the
416 provirus while targeting the PIC to CpG promoters and enhancers.[35, 36] However, a tethering
417 mechanism still has not been elucidated for betaretroviruses,[37] so it is more difficult to
418 hypothesize how their evolution might be impacted by host CpG content. The similarities
419 between the genera in this respect could reflect convergent pressures for these endogenizing
420 viruses. Instead of reflecting integration propensities, perhaps CpG content reflects the
421 proviral tendency to harness promoter-like epigenetic machinery for successful transcription.
422
423 We have also shown that endogenization clearly impacts CpG content and CpG trinucleotide
424 motif patterns, and that these results are likely due to CpG methylation. Every endogenous
425 retrovirus to which we mapped bisulfite sequencing exhibited >80% methylation across the

426 entire provirus **(Figures 2 and 4, Supplementary Figure S6c)**. CpG methylation has long
427 been known to silence endogenous retroviruses during embryonic development.[38] However,
428 CpG methylation within gene bodies is generally not silencing and instead associated with
429 active transcription.[39] From an epigenetic standpoint, this could suggest that the entire
430 provirus is treated similarly to a silenced host promoter. That said, even highly transcribed
431 endogenous retroviruses (e.g. PERV[28]) showed heavy CpG methylation in somatic cells
432 **(Figure 4c)**, calling into question what this mechanism achieves beyond transcriptional
433 silencing. However, bisulfite mapping shows an average of all integrants in different regions,
434 and it is possible that only a few escape repression to produce viral particles, while the
435 majority are silenced.
436
437 Almost across the board, endogenous retroviruses exhibited significantly lower CpG content
438 than their exogenous counterparts **(Figures 2-4)**, with some showing evidence more refined
439 to methylation of particular CpG trinucleotide motifs. In general, these did not appear to
440 reflect the methylation patterns of specific trinucleotide motifs in the host species. Different
441 DNA methyltransferases will target different CpG trinucleotide motifs more heavily,[40, 41] and it
442 is unknown whether these might change between species. The varying promoter methylation
443 patterns of CpG trinucleotides between the different species seem to suggest this is the case.
444 Retroviral methylation is often initiated during embryonic development by *de novo*
445 methyltransferases DNMT3a and DNMT3b.[40-42] This may not be reflective of the overall
446 methylation pattern in somatic cells, which is normally maintained by DNMT1.[43]
447

448 Time from initial ERV endogenization did not appear to correlate well with CpG methylation.
449 For example, MMTV was believed to have endogenized initially around 20 million years ago
450 **(Table 1)**, but it exhibited the most methylation and least amount of CpG loss compared to
451 the other endogenous viruses **(Figure 4i)**, which integrated into their respective genomes
452 initially ranging from 1-7 million years ago **(Table 1)**. However, it is critical to keep in mind
453 that the endogenous retroviruses with exogenous counterparts, such as MMTV, are
454 continuing to produce new endogenous variants that are younger and perhaps more prone to
455 CpG methylation, and their integration sites likely also play a role in methylation frequencies.
456 Therefore, while the approach of using CpG frequencies to estimate time from initial
457 endogenization might work for stable integrants, it cannot be used to generalize time from
458 endogenization for viruses that continue to circulate and create new integrants.
459
460 Using enMuLV data, we found that histone modifications were generally sparse among
461 endogenous retroviruses, suggesting that the seemingly ubiquitous DNA methylation we
462 observed **(Figures 2 and 4, Supplementary Figure S6c)** was more likely important for
463 endogenous retroviral epigenetic regulation. Of proviruses that did have modifications, those
464 associated with both active (H3K27ac, H3K9ac, H3K4me3, H3K36me3) and repressive
465 (H3K27me3, H3K9me3) chromatin could be found **(Figure 5d, Table 3)**. It is probable the
466 proviral histone modification pattern is dictated by the integration site environment and less
467 by the endogenous provirus itself, which requires further exploration. However, there did
468 appear to be significant differences in motifs associated with active and repressive histone
469 modifications when comparing endogenous and exogenous retroviruses **(Figures 5 and 6)**,
470 suggesting histone modifications could be playing an important role early after

471 endogenization, but future work will need to evaluate this hypothesis. Alterations to histone
472 modifications frequently precede DNA methylation for vertebrate gene regulation,[44] so newly
473 endogenized proviruses might be treated similarly. However, previous work on the more
474 ancient human endogenous retroviruses suggested that DNA methylation was an earlier
475 modification, while histone modifications were found in older and more degenerate LTRs.[45]
476 The profound (~93%) DNA methylation across the provirus of the recently integrated KoRV-A
477 supports this hypothesis **(Figure 4g)**. Additionally, KoRV-A defied the clear relationship
478 between proviral DNA methylation and CpG frequency observed for older endogenous
479 proviruses **(Figure 4i)**, suggesting changes to the viral sequence associated with methylated
480 cytosine deamination have thus far been minimal, and supporting the idea that DNA
481 methylation is a common initial epigenetic method for regulation of recent integrants.
482 Interestingly, with the exception of MMTV, the amount of CpG methylation across
483 endogenous proviruses appeared to correlate roughly to their ages, with lower amounts of
484 methylation corresponding to older integrants **(Table 1, Figure 4i)**, further supporting this
485 hypothesis.
486
487 We confined our analysis of enMuLV histone modifications to sequences over 4000 bp,
488 selecting for more intact sequences, and the majority of these sequences (73%) contained no
489 histone modifications. Of those that did, CpG methylation was frequently present overlapping
490 both activation-associated and repressive histone marks in neural progenitors **(Table 3,**
491 **Supplementary Figure S6a-b)**. CpG methylation in embryonic stem cells and neural
492 progenitors was almost ubiquitously present, but generally either in sparse or moderate
493 amounts. However, heavy average CpG methylation is present across the enMuLV provirus

494 in mouse liver **(Supplementary Figure S6c)**, suggesting DNA methylation is dynamic across
495 enMuLV during development and also suggesting marked regulation via CpG methylation in
496 somatic tissue.
497
498 In conclusion, retroviral endogenization shapes the sequence landscape of the provirus due
499 to host epigenetic factors. CpG methylation clearly impacts proviral CpG content over time,
500 while motifs associated with histone modifications appear to reflect the transcriptional state of
501 the provirus, whether endogenous or exogenous. Generally speaking, of the three
502 retroviruses we evaluated for motifs associated with histone modifications, exogenous
503 variants appeared to have more prominent activation-associated motifs (i.e. H3K27ac and
504 H3K4me3) than their endogenous counterparts relative to repression-associated motifs
505 **(Figures 5 and 6)**. Endogenous MuLVs, for which we had more well-defined motifs to
506 examine, exhibited more significant motifs associated with repression than exogenous
507 variants, although those classified as PMV also harbored a number of activation-associated
508 motifs, consistent with their known elevated transcriptional activity compared to other
509 enMuLVs.[46] While no one clear pattern emerged to distinguish endogenous from exogenous
510 variants, we are still able to show that endogenization impacts the proviral sequence
511 landscape in generally predictable ways, despite our analyses traversing multiple
512 endogenous integrants in what are likely a wide variety of chromatin environments from
513 different species.
514
515 **Methods**
516

517 *CG content and trinucleotide motif O/E calculation*:

518

519 To calculate trinucleotide D-ratios (Observed/Expected or O/E ratios), a first order Markov
520 model of conditional probabilities was used.[14, 47, 48] The expected frequency of each motif is
521 estimated using the individual base components with the following example equation:

522

523 $$P_{exp}(CpGpT) = \frac{P_{obs}(CpG) \times P_{obs}(GpT)}{P_{obs}(G)}$$

524

525

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

527

528 CG trinucleotide D-ratios were calculated for all proviral sequences using a Kmer tool
529 previously published and provided by Dr. Diako Ebrahimi,[14] as well as genomic sequences
530 retrieved from each respective species using the *Ensembl* Biomart Tool.[49] Genomic
531 sequences extracted included 500 bp, 1 kb upstream, and 2 kb upstream of each TSS, as
532 well as complete gene bodies and 1kb downstream of all Refseq annotated genes.

533

534 To identify and remove long terminal repeats from analyzed promoter sequences, we used
535 LTRharvest[23] from GenomeTools to identify and remove LTR sequences from all promoter
536 regions in each species. Identification parameters included a minimum length of 10 bp,
537 maximum length of 1000 bp, a minimum interval between LTRs of 15 bp with a maximum
538 interval of 1000 bp.

539

540 *Bisulfite sequencing alignment and cytosine methylation calculation*

541 Bisulfite converted genomes used in these analyses are reported in **Table 4**. SRA files
542 obtained from NCBI were converted to fastq using the NCBI prefetch utility from the SRA
543 toolkit. Fastq files were subjected to Trim Galore!.[50] Alignment of retrieved promoter and
544 retroviral sequences (accessions in **Table 5**) to bisulfite converted genomes was performed
545 using Bismark software.[51] Deduplication and methylation extraction were also performed
546 using Bismark. CpG methylation reports generated from Bismark were used to calculate CpG
547 methylation for CGA, CGC, CGG, and CGT trinucleotides.

548

549 *Histone motif generation*

550 Motifs for all mouse histone modifications were retrieved from.[52] Epigram was used as
551 described in [52] to generate motifs for H3K4me3 and H3K27ac in sheep and domestic cats
552 using the following data sets: H3K4me3 and H3K27ac from sheep liver GSE206736[21] aligned
553 with the Ovis aries genome assembly Oar_rambouillet_v1.0, and H3K4me3 and H3K27ac
554 from cat heart GSE182952 aligned with the *Felis catus* genome assembly
555 F.catus_Fca126_mat1.0.

556

557 *CpG Island Identification*

558 For CpG Island identification, consensus sequences for each provirus were generated using
559 MacVector® software. Sequences were entered into Emboss Cpgplot[31] using a window size
560 of 100, a minimum length of 200, a minimum observed CpG frequency of 0.6, and a minimum
561 CpG percentage of 50. For JSRV sequences, CGI minimum lengths were dropped to 100.

562

563 *SEA analysis*

564 For Simple Enrichment Analysis (SEA), we used the SEA tool from the MEME suite.[15]
565 Histone modification motif analyses for exogenous retrovirus variants were submitted for
566 enrichment compared to their corresponding endogenous variants as negative controls, and
567 vice versa with histone motifs generated from Epigram analyzed for enrichment as described
568 above with an E-value cutoff of 0.05. Motifs for H3K4me3, H3K27ac, and H3K27me3 were
569 retrieved only if they aligned with CpG islands from the corresponding proviruses being
570 analyzed. In all cases, only sequence motifs enriched in all inputted sequences with no more
571 than one of the negative control sequences, or one less than all inputted sequences with no
572 negative control sequences, were considered significant, even when E-value and CGI criteria
573 were met.

574

575 *enMuLV histone modification identification*

576 Consensus sequences for endogenous MuLV variants MPMV, PMV, and XMV were all
577 entered into the UCSC Genome Browser BLAT tool.[53, 54] ENCODE histone tracks for
578 H3K27ac, H3K27me3, H3K9ac, H3K9me3, H3K4me3, H3K36me3, and H3K4me1 from
579 forebrain, liver, kidney, and intestine from day 16 embryonic mice aligning ≥ 4000 bp of
580 enMuLV sequence were visualized to identify sites of modification throughout the provirus.
581 ENCODE tracks for CpG methylation in mouse embryonic stem cells and neural progenitor
582 cells were also included.

583

584 *Statistical analysis*

585 Statistical analyses and data visualization were performed using GraphPad Prism® software
586 version 10.4.2 for macOS, as well as the stats and rstatix packages in R for statistics and
587 ggplot2 for visualization. Multiple comparisons for CpG motifs between proviruses and
588 promoters were performed using the Kruskal-Wallis test for multiple comparisons followed by
589 Dunn's tests. Two-sided independent t-tests or Mann-Whitney tests were used for individual
590 comparisons where appropriate. All correlation analyses were performed using Spearman's
591 rank correlation test with a 95% confidence interval. Pearson correlation was used in one
592 case **(Figure 4i)** after the normality assumption was confirmed by Shapiro-Wilk and
593 Komogorov-Smirnov tests. In all analyses, statistical significance was defined as p-value <
594 0.05.
595
596 *Phylogenetic analysis*
597 Phylogenetic analysis for JSRV sequences reported in **Supplementary Figure S2** were
598 conducted using MacVector® software. Neighbor-joining trees were built using 1000
599 bootstraps from from enJSRV and exJSRV sequences aligned with Clustal W.
600
601 **Data Availability Statement**
602 *Accession numbers:*
603 Chip-seq data from GSE206736 was retrieved to analyze for H3K4me3 and H3K27ac motifs
604 in sheep. Chip-seq data from GSE182952 was retrieved to analyze for H3K4me3 and
605 H3K27ac motifs in domestic cats.
606

607 **Table 4** lists bisulfite converted genome sequences accessed on the NCBI BioProjects

608 database with the following Project, SRA Experiment, and Accession numbers:

609

Table 4: Bisulfite converted genomes from the NCBI BioProjects database used for this study.

Species	Tissue	Bioproject	Experiment	Accession numbers					
Cat	Blood	PRJNA16726	SRX548709	SRR1293903	SRR1293905				
Sheep	Skeletal Muscle	PRJNA622418	SRX8073310	SRR11497387					
Pig	Blood	PRJEB55066	ERX11037418	ERR11635337					
Koala	Kidney	PRJNA629340	SRX8207656	SRR11646494	SRR11646495				
Mouse	Liver	PRJNA416505	SRX3347189	SRR6239777	SRR6239784	SRR6239792	SRR6239799	SRR6239807	SRR6239814
				SRR6239778	SRR6239785	SRR6239793	SRR6239801	SRR6239808	SRR6239815
				SRR6239779	SRR6239786	SRR6239794	SRR6239802	SRR6239809	SRR6239816
				SRR6239780	SRR6239787	SRR6239795	SRR6239803	SRR6239810	
				SRR6239781	SRR6239788	SRR6239796	SRR6239804	SRR6239811	
				SRR6239782	SRR6239789	SRR6239797	SRR6239805	SRR6239812	
				SRR6239783	SRR6239790	SRR6239798	SRR6239806	SRR6239813	

610

611 **Table 5** lists accession numbers for all viral sequences accessed on NCBI Genbank:

Table 5: Sources and NCBI GenBank Accession numbers for proviral sequences used in this study.

Gammaretroviruses					
enMuLV [25]					
Mpmv1	Mpmv6	Pmv14	Pmv22	Xmv10	Xmv42
Mpmv10	Mpmv7	Pmv15	Pmv23	Xmv12	Xmv43
Mpmv11	Mpmv8	Pmv16	Pmv24	Xmv13	Xmv8

Mpmv12	Mpmv9	Pmv17	Pmv4	Xmv15	Xmv9
Mpmv13	Pmv1	Pmv18	Pmv5	Xmv16	
Mpmv2	Pmv10	Pmv19	Pmv6	Xmv17	
Mpmv3	Pmv11	Pmv2	Pmv7	Xmv18	
Mpmv4	Pmv12	Pmv20	Pmv8	Xmv19	
Mpmv5	Pmv13	Pmv21	Pmv9	Xmv41	
exMuLV					
KY574504.1	KY574507.1	KY574510.1	KY574513.1	KY574516.1	MG525170.1
KY574505.1	KY574508.1	KY574511.1	KY574514.1	KY574517.1	
KY574506.1	KY574509.1	KY574512.1	KY574515.1	KY574518.1	
enFeLV					
AY364318.1	LC196053.1	LC196055.1	OP595707.1		
AY364319.1	LC196054.1	OP595706.1	OP595708.1		
exFeLV-A					
AB672612.1	AB672612.1	M18247.1	MZ964580.1		
AF052723.1	AF052723.1	MT129531.1	NC_001940.1		
KoRV-A full-length					
AB721500.1	KF786280.1	KF786282.1	KF786284.1	KF786286.1	
KC779547.1	KF786281.1	KF786283.1	KF786285.1	NC_039228.1	
KoRV-A env					
AB721500.1	JQ244835.1	JQ244838.1	KF786281.1	KF786284.1	NC_039228.1
AB823238.1	JQ244836.1	JQ244839.1	KF786282.1	KF786285.1	
AF151794.2	JQ244837.1	KF786280.1	KF786283.1	KF786286.1	
KoRV-B-F env					
AB822553.1	AB828005.1	KP792564.1	KU533852.1		
AB828004.1	KC779547.1	KP792565.1	KU533853.1		
PERV-A					
AF435966.1	AY099323.1	HM159246.1	KY484771.1		
AJ133817.1	EF133960.1	HQ540591.1	OR574834.1		
PERV-B					
AJ133816.1	AJ293657.1	EU523109.1	HQ536010.1	HQ540594.1	
AJ133818.1	AY056035.1	HQ536008.1	HQ536011.1	HQ540595.1	
AJ279057.1	AY099324.1	HQ536009.1	HQ540593.1		
PERV-C					
AF038599.1	AY570980.1	HQ536014.1	KC116219.1	KY352351.2	
AF038600.1	HQ536012.1	HQ536015.1	KC116220.1	MT984604.1	
AM229312.2	HQ536013.1	HQ536016.1	KC116221.1		
Betaretroviruses					
enMMTV					

AC122322.4	AF228550.1	NG_005612.1			
AC140344.3	AL683884.26				
exMMTV					
AF033807.1	AF228552.1	M15122.1			
AF228551.1	D16249.1	NC_001503.1			
enJSRV					
AF136224.1	EF680297.1	EF680302.1	EF680307.1	EF680314.1	MF175071.1
AF136225.1	EF680298.1	EF680303.1	EF680308.1	MF175067.1	
AF153615.1	EF680299.1	EF680304.1	EF680309.1	MF175068.1	
DQ838493.1	EF680300.1	EF680305.1	EF680310.1	MF175069.1	
EF680296.1	EF680301.1	EF680306.1	EF680312.1	MF175070.1	
exJSRV OPA					
AF105220.1	CQ964469.1	KP691837.1	NC_001494.1		
AF357971.1	DQ838494.1	M80216.1			
exJSRV ENTV					
FJ744146.1	FJ744149.1	GU292315.1	GU292318.1	PP707036.1	
FJ744147.1	FJ744150.1	GU292316.1	KC189895.1		
FJ744148.1	GU292314.1	GU292317.1	NC_007015.1		
phaCin-β [30]					
pCi1002	pCi237	pCi366	pCi47	pCi921	pCi989
pCi194	pCi273	pCi414	pCi643	pCi933	pCi998
pCi233	pCi303	pCi457	pCi786	pCi983	

Tools and software:

Genomic sequence retrieval was performed with Ensembl Biomart.

LTRHarvest[23] from GenomeTools was used to extract LTR sequences from promoters.

SRA files obtained from NCBI were converted to fastq using the NCBI prefetch utility from the SRA toolkit. Fastq files were trimmed using Trim Galore!.[50] Bisulfite converted sequences were aligned with Bismark[51] software.

Epigram[52, 55] was used to predict motifs associated with histone modifications.

Proviral alignments, consensus generation, and phylogenetic analysis was performed with MacVector® software.

622 Emboss Cpgplot[31] was used for CpG Island identification.

623 The SEA tool from the MEME suite[15] was used for Simple Enrichment Analysis.

624 The UCSC Genome Browser BLAT tool[53, 54] was used to locate endogenous proviral

625 sequences in the mouse genome for histone modification identification with ENCODE histone

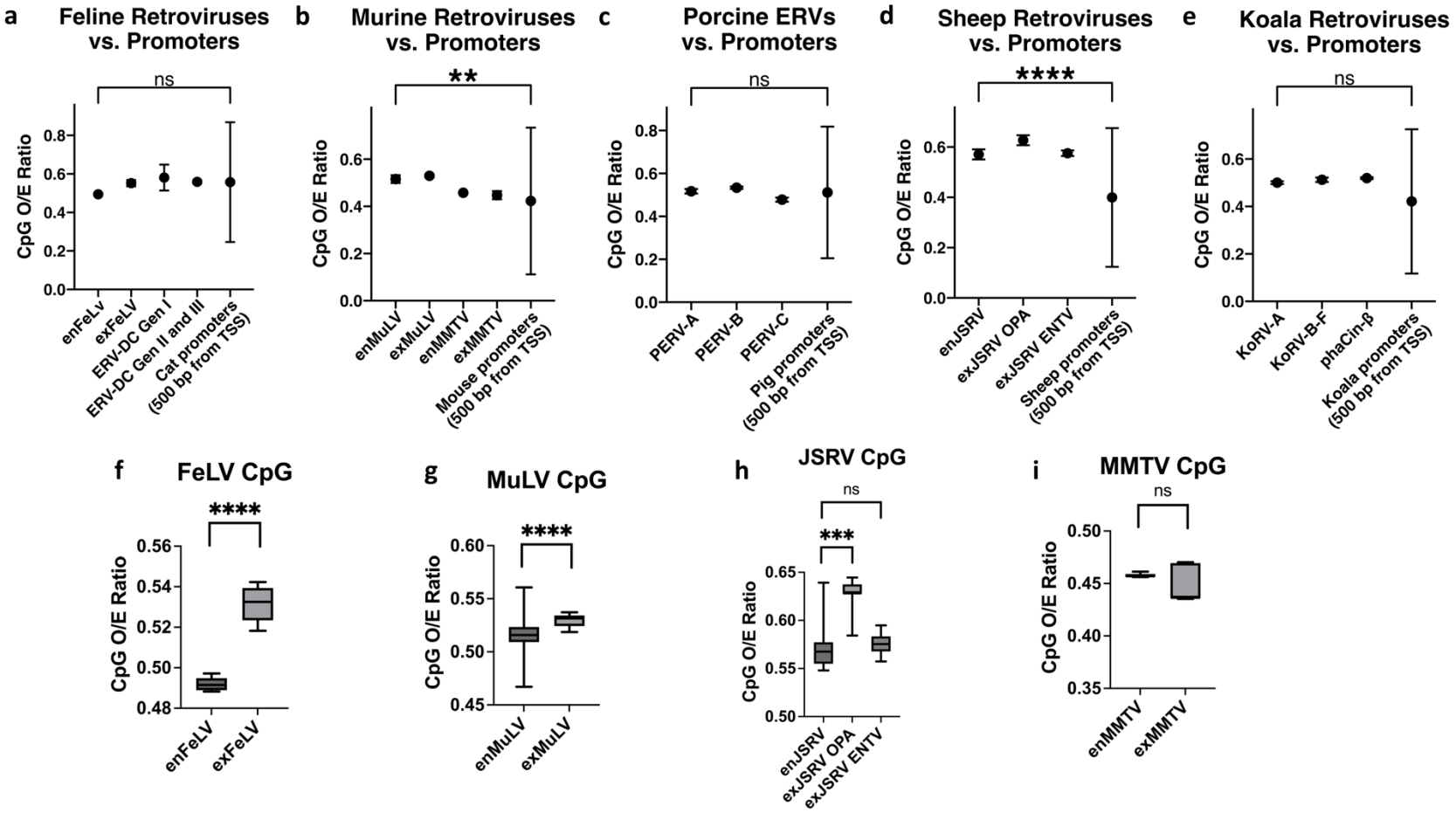
626 modification tracks.

627 Statistical analyses were performed using GraphPad Prism® software, as well as the rtstatix

628 package in R. Visualization was also performed with GraphPad Prism® software and ggplot2

629 in R.

630 **Figures:**



631

632 **Figure 1: CpG content of beta- and gammaretroviruses commonly reflects CpG content of promoter sequences,**

633 **and CpG content is usually lower with endogenization. a-e)** Endogenous and exogenous retroviral median CpG D-

634 ratios are compared to those of promoters (500 bp upstream of the TSS) for **a)** domestic cats **b)** mice **c)** pigs **d)** sheep

635 and **e)** koalas. **f-h)** Endogenized retroviral variants exhibit lower CpG content than exogenous counterparts for **f)** Feline

636 Leukemia Virus (FeLV), **g)** Murine Leukemia Virus (MuLV), and **h)** Jaagsiekte sheep retrovirus (JSRV). **i)** Mouse

637 Mammary Tumor Virus (MMTV) shows no significant change to median CpG content. P-values are reported from Kruskal-

638 Wallis tests for multiple comparisons **(a-e)** and Mann-Whitney comparisons **(f-i)**. Error bars represent interquartile ranges.

639 n.s. $p > 0.05$, *** $p < 0.001$, **** $p < 0.0001$

640

641

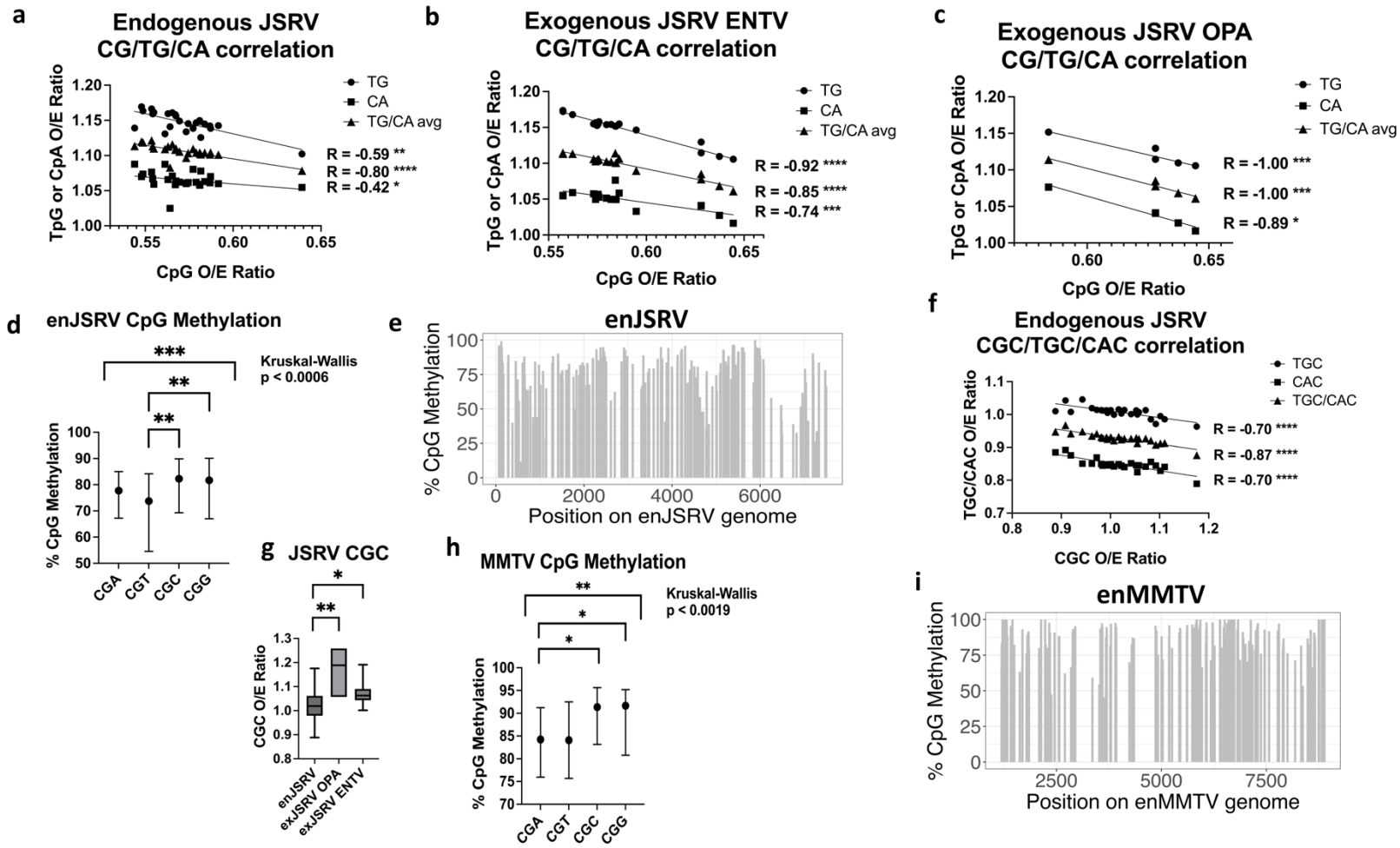


Figure 2: Endogenous betaretroviruses JSRV and MMTV are highly methylated and exhibit CpG decline. a-c) JSRV shows a strong negative correlation between CpG content and TpG/CpA content in **a)** endogenous **b)** exogenous Enzootic Nasal Tumor Virus (ENTV) and **c)** exogenous Ovine Pulmonary Adenocarcinoma (OPA) variants. **d)** CpG methylation across endogenous JSRV shows uneven CpG trinucleotide distribution, with a preference for CGC and CGG

trinucleotides. **e)** Bismark mapping shows heavy CpG methylation across the entire JSRV provirus. **f)** Endogenous JSRV shows a strong negative correlation between CGC content and TGC/CAC. **g)** CGC content is significantly decreased in endogenous JSRV compared to both exogenous JSRV strains. **h)** Trinucleotide analysis of MMTV methylation shows a significant preference for CGC motifs. **i)** Bismark mapping shows heavy CpG methylation across the body of the MMTV provirus.

R and p-values for correlation analyses are from Spearman correlations **(a-c,f)**. P-values are reported for the Kruskal-Wallis test for multiple comparisons with Dunn's test **(d,g,h)**, while error bars in **d)**, **g)**, and **h)** represent interquartile ranges. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$

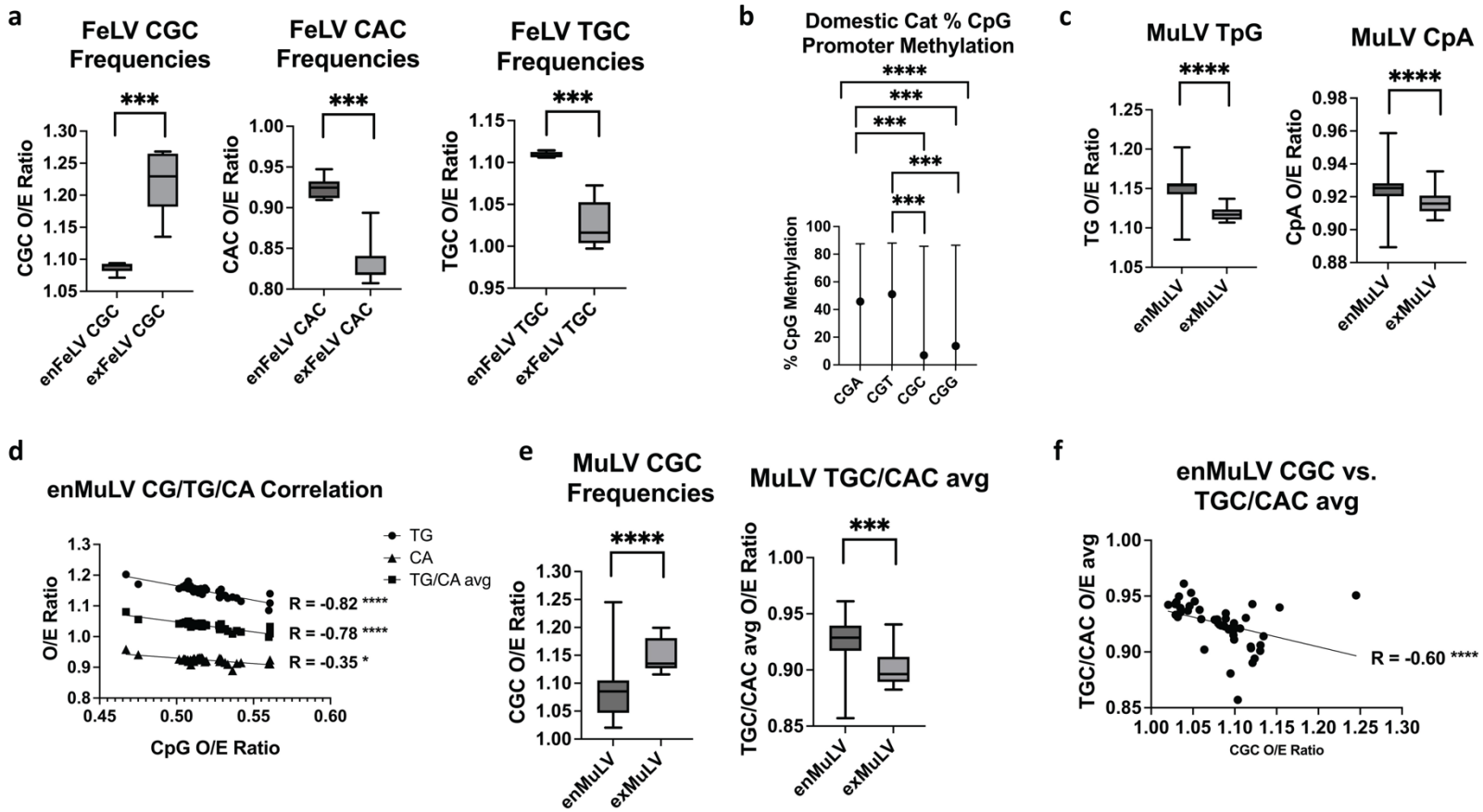


Figure 3: Endogenization of gammaretroviruses FeLV and MuLV impacts retroviral CpG content: a) FeLV endogenization leads to a significant decrease in CGC content, with corresponding increases in CAC and TGC content. **b)** Domestic cat promoters exhibit uneven median CpG methylation across trinucleotides, with CGC being the least methylated. **c)** Endogenous MuLV exhibits significantly increased TpG and CpA content compared to exogenous variants.

662 **d)** enMuLV shows a strong negative correlation between CpG, TpG, and CpA, suggesting methylated cytosine
663 deamination of enMuLV. **e)** EnMuLV CGC frequencies are significantly decreased compared to exMuLV, with
664 corresponding increases in TGC/CAC frequencies. **f)** enMuLV shows a strong negative correlation between CGC and
665 TGC/CAC, suggesting methylated cytosine deamination of enMuLV at CGC trinucleotides. P-values are reported for
666 Mann-Whitney tests **(a, c, e)**, Kruskal-Wallis test for multiple comparisons **(b)**, and Spearman correlation **(d, f)**. Error bars
667 for **b)** represent interquartile ranges. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$
668
669
670

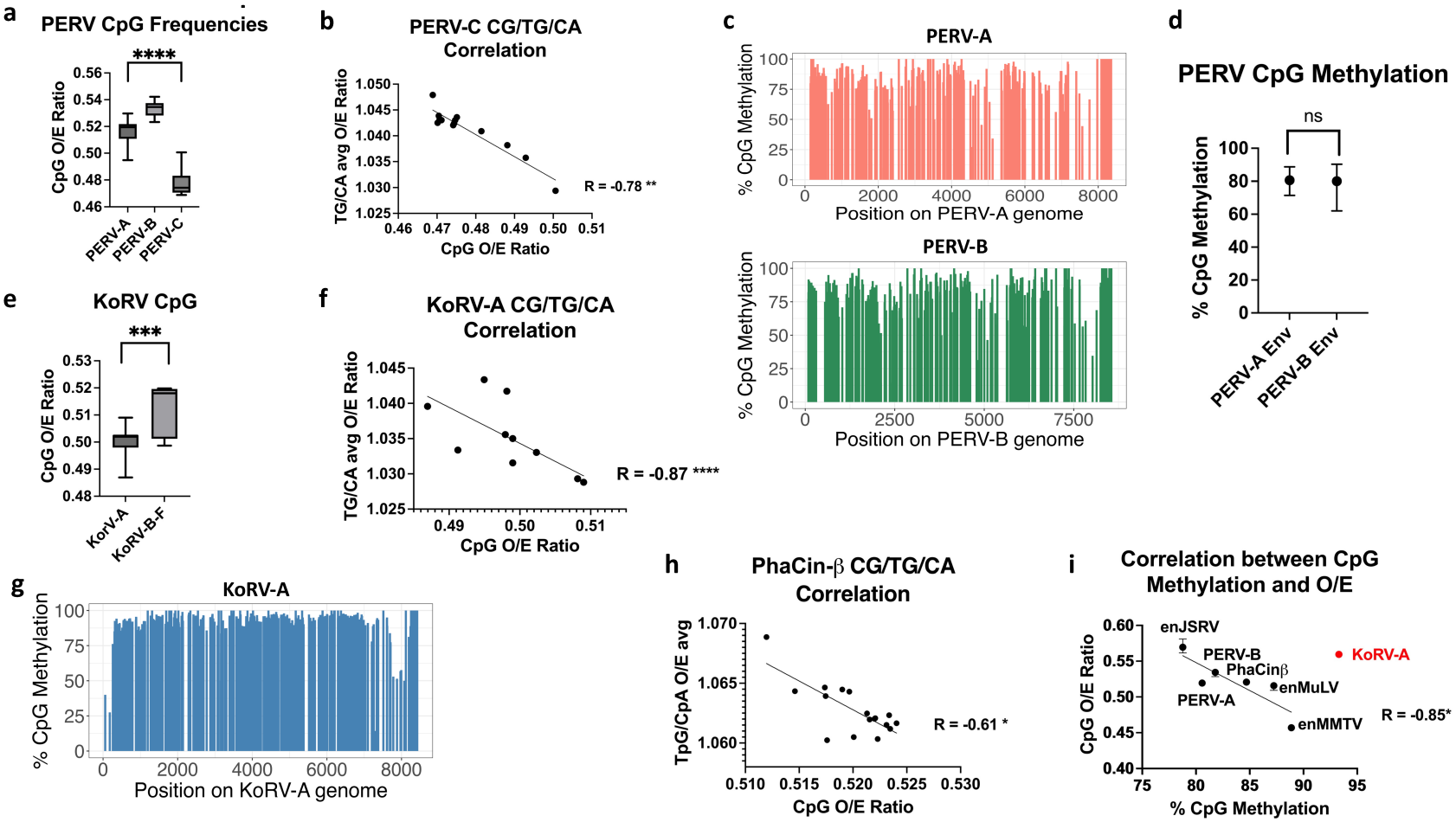


Figure 4: CpG content of seven vertebrate endogenous retroviruses corresponds with prior CpG methylation and transcriptional activity: a) Porcine Endogenous Retrovirus (PERV) variants exhibit significant variation in CpG content. **b)** PERV-C shows a significant negative correlation between CpG with 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)**

676 PERV-A and PERV-B show no significant difference in median CpG methylation content. **e)** Endogenous Koala Retrovirus
677 (KoRV-A) has significantly lower CpG content than the other KoRV subtypes. **f)** KoRV-A shows significant negative
678 correlation between CpG and TpG/CpA frequency, suggesting impacts from DNA methylation. **g)** Bismark bisulfite
679 mapping shows KoRV-A is heavily methylated across the entire provirus. **h)** Endogenous koala betaretrovirus phaCin-β
680 shows a significant negative correlation between CpG and deamination products TpG/CpA. **i)** Median CpG % methylation
681 plotted against median CpG observed/expected (O/E) ratios demonstrates a significant negative correlation between them
682 for all endogenous retroviruses except KoRV-A (in red). Error bars in **(d)** represent interquartile ranges. P-values are
683 reported for Kruskal-Wallis tests for multiple comparisons **(a)**, Mann-Whitney tests **(d,e)**, Spearman correlation **(b,f,h)**, and
684 Pearson correlation **(i)**. *$p < 0.05$, **$p < 0.01$, *** $p < 0.001$, ****$p < 0.0001$

685
686

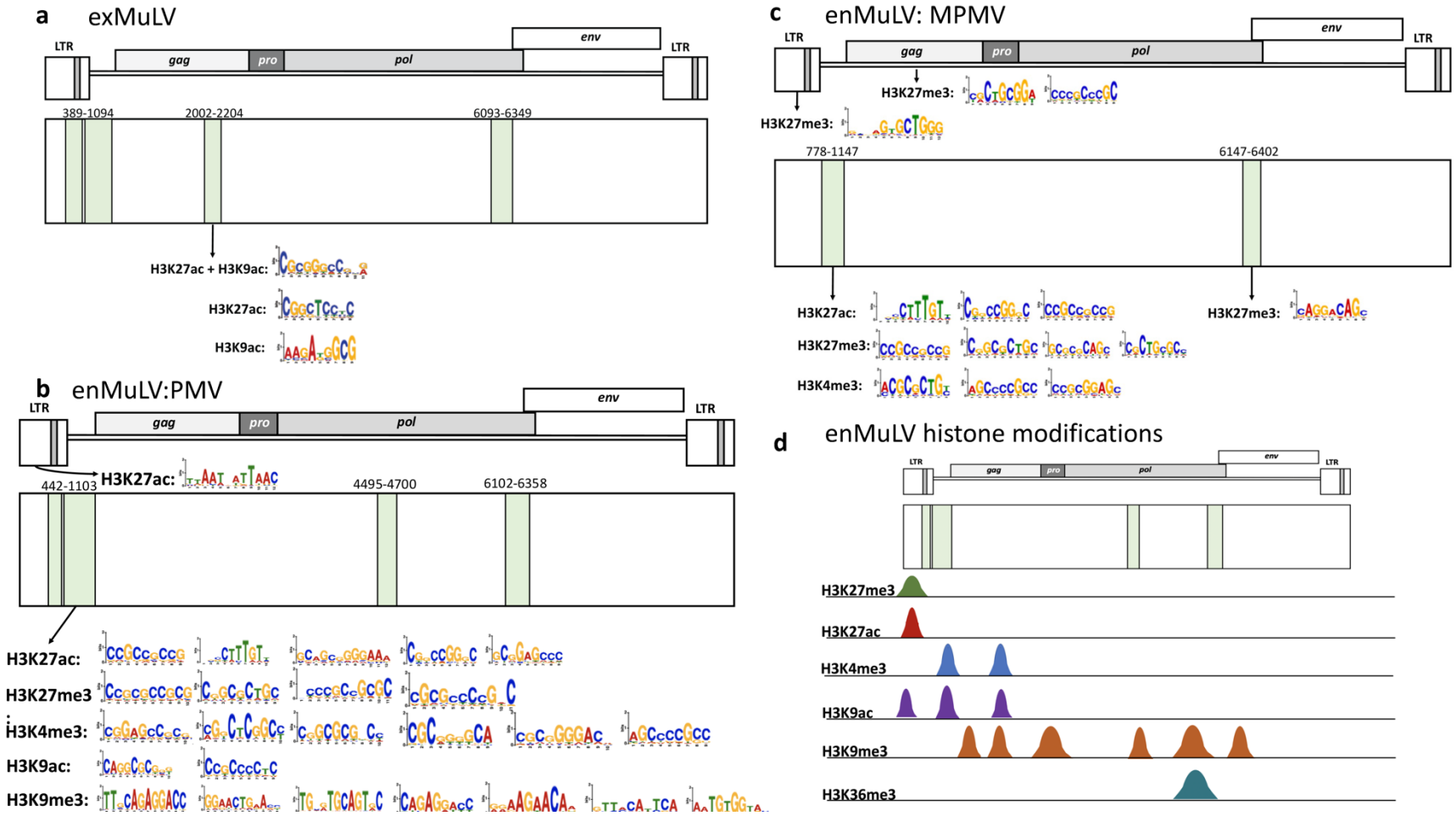


Figure 5: Murine leukemia virus (MuLV) shows motifs associated with histone modifications consistent with exogenous and variably active endogenous variants: a) Emboss Cpgplot and SEA analysis identify CpG islands (shaded in green) across the exogenous MuLV provirus, with one in particular containing H3K27ac and H3K9ac-associated motifs. **b)** Emboss Cpgplot and SEA analysis of polytropic endogenous MuLV (PMV) show heavy density of

692 motifs associated with histone modifications within a CGI in the leader region, including those for active (H3K27ac,
693 H3K4me3, H3K9ac) and repressive (H3K27me3 and H3K9me3) domains. **c)** Emboss Cpgplot and SEA analysis of
694 endogenous modified polytropic endogenous MuLV (MPMV) shows widespread H3K27me3 motifs in the LTR and CGIs
695 throughout the provirus. **d)** Visualization of ENCODE ChIP-Seq data from day 16 mouse embryos in the UCSC Genome
696 Browser shows several histone modifications present in endogenous MuLV sequences (Figure details in **Table 3**). Further
697 details on MuLV SEA analysis can be found in **Supplementary Table S2**.
698
699

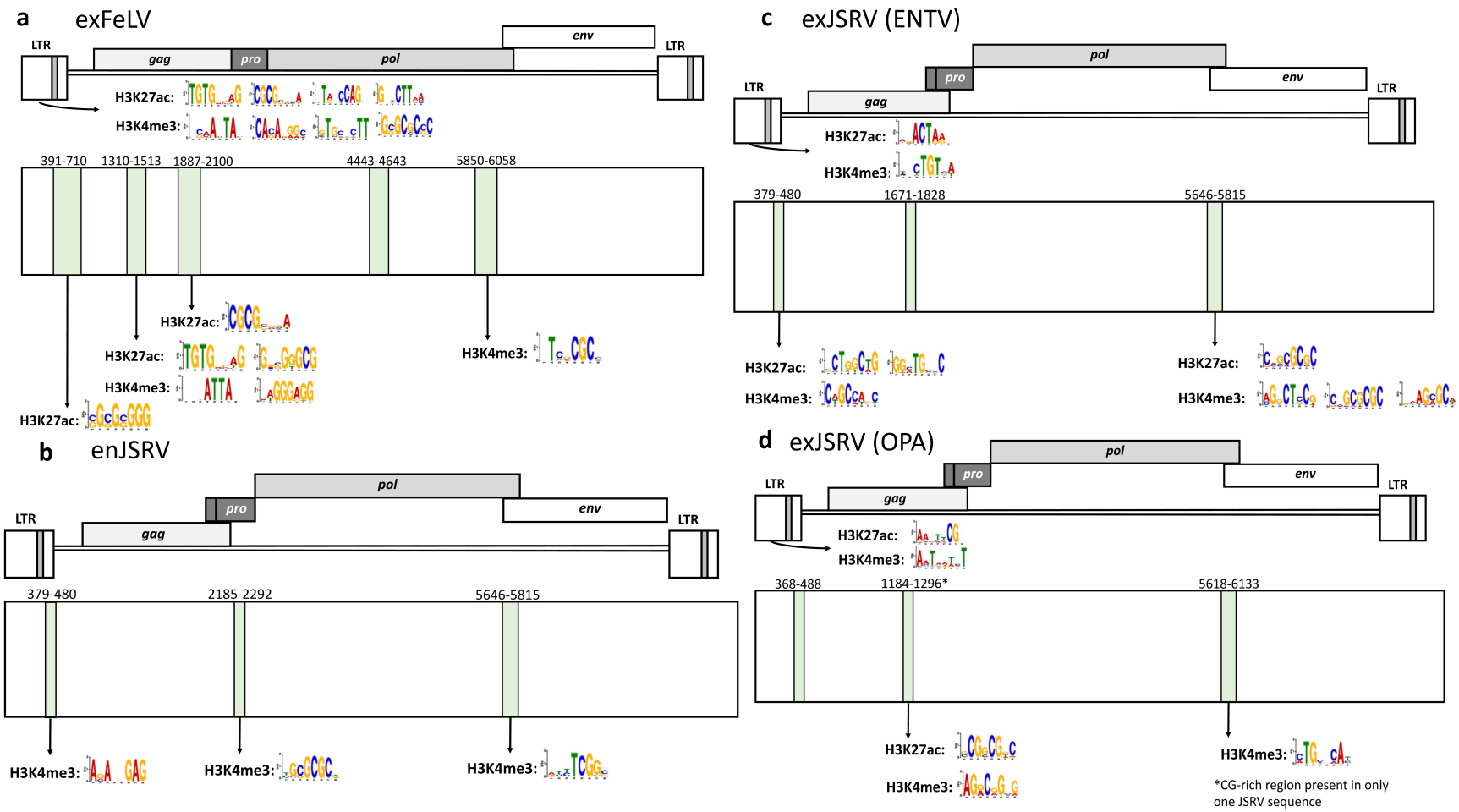


Figure 6: Exogenous Feline Leukemia Virus (FeLV) and Jaagsiekte Sheep Retrovirus (JSRV) variants contain more significant motifs associated with 'active' histone modifications than endogenous variants. a) EMBOSS Cpgplot and SEA analysis of exogenous FeLV showed significant H3K27ac and H3K4me3 motifs in the LTR and CGIs throughout the provirus. Further details on SEA analysis can be found in **Supplementary Tables S3 and S4**.

705 **b)** EMBOSS Cpgplot and SEA analysis show endogenous JSRV contains several CG-rich regions with only H3K4me3
706 motifs present and no H3K27ac. **c)** EMBOSS Cpgplot and SEA analysis of the exogenous Enzootic Nasal Tumor Virus
707 (ENTV) JSRV variant show significant motifs present in the LTR, leader region, and *pol-env* junction for both H3K27ac
708 and H3K4me3. **d)** EMBOSS Cpgplot and SEA analysis of the exogenous OPA JSRV variant shows several H3K27ac and
709 H3K4me3-associated motifs within CG-rich regions primarily in the LTR and *gag*.
710

711 References
712
713 1. Meyer TJ, Rosenkrantz JL, Carbone L, Chavez SL. Endogenous Retroviruses: With Us and
714 against Us. Front Chem. 2017;5:23. Epub 20170407. doi: 10.3389/fchem.2017.00023. PubMed PMID:
715 28439515; PMCID: PMC5384584.
716 2. Mager DL, Stoye JP. Mammalian Endogenous Retroviruses. Microbiol Spectr.
717 2015;3(1):MDNA3-0009-2014. doi: 10.1128/microbiolspec.MDNA3-0009-2014. PubMed PMID:
718 26104559.
719 3. Coffin J, Blomberg J, Fan H, Gifford R, Hatziioannou T, Lindemann D, Mayer J, Stoye J, Tristem
720 M, Johnson W, Ictv Report C. ICTV Virus Taxonomy Profile: Retroviridae 2021. J Gen Virol.
721 2021;102(12). doi: 10.1099/jgv.0.001712. PubMed PMID: 34939563; PMCID: PMC8744268.
722 4. De Rijck J, de Kogel C, Demeulemeester J, Vets S, El Ashkar S, Malani N, Bushman FD, Landuyt
723 B, Husson SJ, Busschots K, Gijsbers R, Debyser Z. The BET family of proteins targets moloney murine
724 leukemia virus integration near transcription start sites. Cell Rep. 2013;5(4):886-94. Epub 20131031.
725 doi: 10.1016/j.celrep.2013.09.040. PubMed PMID: 24183673; PMCID: PMC4197836.
726 5. Goubran M, Wang W, Indik S, Faschinger A, Wasilenko ST, Bintner J, Carpenter EJ, Zhang G,
727 Nuin P, Macintyre G, Wong GK, Mason AL. Isolation of a Human Betaretrovirus from Patients with
728 Primary Biliary Cholangitis. Viruses. 2022;14(5). Epub 20220424. doi: 10.3390/v14050886. PubMed
729 PMID: 35632628; PMCID: PMC9146342.
730 6. Faschinger A, Rouault F, Sollner J, Lukas A, Salmons B, Gunzburg WH, Indik S. Mouse
731 mammary tumor virus integration site selection in human and mouse genomes. J Virol.
732 2008;82(3):1360-7. Epub 20071121. doi: 10.1128/JVI.02098-07. PubMed PMID: 18032509; PMCID:
733 PMC2224419.
734 7. Jarosz AS, Halo JV. Transcription of Endogenous Retroviruses: Broad and Precise Mechanisms
735 of Control. Viruses. 2024;16(8). Epub 20240817. doi: 10.3390/v16081312. PubMed PMID: 39205286;
736 PMCID: PMC11359688.
737 8. Tarlinton R, Meers J, Hanger J, Young P. Real-time reverse transcriptase PCR for the
738 endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease
739 in koalas. J Gen Virol. 2005;86(Pt 3):783-7. doi: 10.1099/vir.0.80547-0. PubMed PMID: 15722540.
740 9. Hashem MA, Kayesh MEH, Maetani F, Eiei T, Mochizuki K, Ochiai S, Ito A, Ito N, Sakurai H, Asai
741 T, Tsukiyama-Kohara K. Koala retrovirus (KoRV) subtypes and their impact on captive koala
742 (Phascolarctos cinereus) health. Arch Virol. 2021;166(7):1893-901. Epub 20210426. doi:
743 10.1007/s00705-021-05078-y. PubMed PMID: 33900468.
744 10. Ngo MH, AbuEed L, Kawasaki J, Oishi N, Pramono D, Kimura T, Sakurai M, Watanabe K,
745 Mizukami Y, Ochi H, Anai Y, Odahara Y, Umehara D, Kawamura M, Watanabe S, Miyake A, Nishigaki K.
746 Multiple recombination events between endogenous retroviral elements and feline leukemia virus. J
747 Virol. 2024;98(2):e0140023. Epub 20240119. doi: 10.1128/jvi.01400-23. PubMed PMID: 38240589;
748 PMCID: PMC10878261.
749 11. van der Kuyl AC, Dekker JT, Goudsmit J. Discovery of a new endogenous type C retrovirus
750 (FcEV) in cats: evidence for RD-114 being an FcEV(Gag-Pol)/baboon endogenous virus BaEV(Env)
751 recombinant. J Virol. 1999;73(10):7994-8002. doi: 10.1128/JVI.73.10.7994-8002.1999. PubMed PMID:
752 10482547; PMCID: PMC112814.

753 12. Zamudio N, Bourc'his D. Transposable elements in the mammalian germline: a comfortable
754 niche or a deadly trap? Heredity (Edinb). 2010;105(1):92-104. Epub 20100505. doi:
755 10.1038/hdy.2010.53. PubMed PMID: 20442734.
756 13. Letchumanan P, Theva Das K. The role of genetic diversity, epigenetic regulation, and sex-
757 based differences in HIV cure research: a comprehensive review. Epigenetics Chromatin.
758 2025;18(1):1. Epub 20250103. doi: 10.1186/s13072-024-00564-4. PubMed PMID: 39754177; PMCID:
759 PMC11697457.
760 14. Alinejad-Rokny H, Anwar F, Waters SA, Davenport MP, Ebrahimi D. Source of CpG Depletion in
761 the HIV-1 Genome. Mol Biol Evol. 2016;33(12):3205-12. Epub 20160928. doi:
762 10.1093/molbev/msw205. PubMed PMID: 27682824.
763 15. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res.
764 2015;43(W1):W39-49. Epub 20150507. doi: 10.1093/nar/gkv416. PubMed PMID: 25953851; PMCID:
765 PMC4489269.
766 16. Roca AL, Pecon-Slattery J, O'Brien SJ. Genomically intact endogenous feline leukemia viruses
767 of recent origin. J Virol. 2004;78(8):4370-5. doi: 10.1128/jvi.78.8.4370-4375.2004. PubMed PMID:
768 15047851; PMCID: PMC374283.
769 17. Stocking C, Kozak CA. Murine endogenous retroviruses. Cell Mol Life Sci. 2008;65(21):3383-98.
770 doi: 10.1007/s00018-008-8497-0. PubMed PMID: 18818872; PMCID: PMC4802364.
771 18. Ishida Y, Zhao K, Greenwood AD, Roca AL. Proliferation of endogenous retroviruses in the
772 early stages of a host germ line invasion. Mol Biol Evol. 2015;32(1):109-20. Epub 20140925. doi:
773 10.1093/molbev/msu275. PubMed PMID: 25261407; PMCID: PMC4271524.
774 19. Parisi F, Freer G, Mazzanti CM, Pistello M, Poli A. Mouse Mammary Tumor Virus (MMTV) and
775 MMTV-like Viruses: An In-depth Look at a Controversial Issue. Viruses. 2022;14(5). Epub 20220506.
776 doi: 10.3390/v14050977. PubMed PMID: 35632719; PMCID: PMC9147501.
777 20. Arnaud F, Varela M, Spencer TE, Palmarini M. Coevolution of endogenous betaretroviruses of
778 sheep and their host. Cell Mol Life Sci. 2008;65(21):3422-32. doi: 10.1007/s00018-008-8500-9.
779 PubMed PMID: 18818869; PMCID: PMC4207369.
780 21. Chen S, Liu S, Shi S, Jiang Y, Cao M, Tang Y, Li W, Liu J, Fang L, Yu Y, Zhang S. Comparative
781 epigenomics reveals the impact of ruminant-specific regulatory elements on complex traits. BMC Biol.
782 2022;20(1):273. Epub 20221208. doi: 10.1186/s12915-022-01459-0. PubMed PMID: 36482458;
783 PMCID: PMC9730597.
784 22. Lillie M, Pettersson M, Jern P. Contrasting segregation patterns among endogenous
785 retroviruses across the koala population. Commun Biol. 2024;7(1):350. Epub 20240321. doi:
786 10.1038/s42003-024-06049-0. PubMed PMID: 38514810; PMCID: PMC10957985.
787 23. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo
788 detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18. Epub 20080114. doi:
789 10.1186/1471-2105-9-18. PubMed PMID: 18194517; PMCID: PMC2253517.
790 24. Upadhyay M, Samal J, Kandpal M, Vasaikar S, Biswas B, Gomes J, Vivekanandan P. CpG
791 dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. J
792 Virol. 2013;87(24):13816-24. Epub 20131009. doi: 10.1128/JVI.02515-13. PubMed PMID: 24109231;
793 PMCID: PMC3838256.
794 25. Jern P, Stoye JP, Coffin JM. Role of APOBEC3 in genetic diversity among endogenous murine
795 leukemia viruses. PLoS Genet. 2007;3(10):2014-22. Epub 20070910. doi:
796 10.1371/journal.pgen.0030183. PubMed PMID: 17967065; PMCID: PMC2041998.

797 26. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, Neuberger MS,
798 Malim MH. DNA deamination mediates innate immunity to retroviral infection. Cell. 2003;113(6):803-
799 9. doi: 10.1016/s0092-8674(03)00423-9. PubMed PMID: 12809610.
800 27. Lapp Z, Yoon H, Foley B, Leitner T. Hypermut 3: identifying specific mutational patterns in a
801 defined nucleotide context that allows multistate characters. Bioinform Adv. 2025;5(1):vbaf025. Epub
802 20250210. doi: 10.1093/bioadv/vbaf025. PubMed PMID: 39981109; PMCID: PMC11842049.
803 28. Chen Y. CM, Duan X., Cui J. Ancient origin and complex evolution of porcine endogenous
804 retroviruses. Biosafety and Health. 2020;2:142-51. doi:
805 http://dx.doi.org/10.1016/j.bsheal.2020.03.003.
806 29. Yu T, Blyton MBJ, Abajorga M, Koppetsch BS, Ho S, Xu B, Hu Z, Luban J, Chappell K, Weng Z,
807 Theurkauf WE. Evolution of KoRV-A transcriptional silencing in wild koalas. Cell. 2025. Epub
808 20250226. doi: 10.1016/j.cell.2025.02.006. PubMed PMID: 40056902.
809 30. Lillie M, Hill J, Pettersson ME, Jern P. Expansion of a retrovirus lineage in the koala genome.
810 Proc Natl Acad Sci U S A. 2022;119(25):e2201844119. Epub 20220613. doi:
811 10.1073/pnas.2201844119. PubMed PMID: 35696585; PMCID: PMC9231498.
812 31. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite.
813 Trends Genet. 2000;16(6):276-7. doi: 10.1016/s0168-9525(00)02024-2. PubMed PMID: 10827456.
814 32. Antequera F. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci.
815 2003;60(8):1647-58. doi: 10.1007/s00018-003-3088-6. PubMed PMID: 14504655; PMCID:
816 PMC11138798.
817 33. Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-
818 host evolution. Proc Natl Acad Sci U S A. 2013;110(50):20146-51. Epub 20131125. doi:
819 10.1073/pnas.1315419110. PubMed PMID: 24277832; PMCID: PMC3864273.
820 34. Hayward JA, Tachedjian M, Kohl C, Johnson A, Dearnley M, Jesaveluk B, Langer C, Solymosi PD,
821 Hille G, Nitsche A, Sanchez CA, Werner A, Kontos D, Crameri G, Marsh GA, Baker ML, Poumbourios P,
822 Drummer HE, Holmes EC, Wang LF, Smith I, Tachedjian G. Infectious KoRV-related retroviruses
823 circulating in Australian bats. Proc Natl Acad Sci U S A. 2020;117(17):9529-36. Epub 20200413. doi:
824 10.1073/pnas.1915400117. PubMed PMID: 32284399; PMCID: PMC7196810.
825 35. Wang GZ, Goff SP. Moloney Murine Leukemia Virus p12 Is Required for Histone Loading onto
826 Retroviral DNAs. J Virol. 2021;95(15):e0049521. Epub 20210712. doi: 10.1128/JVI.00495-21. PubMed
827 PMID: 34011543; PMCID: PMC8274600.
828 36. Miklik D, Slavkova M, Kucerova D, Mekadim C, Mrazek J, Hejnar J. Long Terminal Repeats of
829 Gammaretroviruses Retain Stable Expression after Integration Retargeting. Viruses. 2024;16(10).
830 Epub 20240925. doi: 10.3390/v16101518. PubMed PMID: 39459853; PMCID: PMC11512309.
831 37. Desfarges S, Ciuffi A. Retroviral integration site selection. Viruses. 2010;2(1):111-30. Epub
832 20100112. doi: 10.3390/v2010111. PubMed PMID: 21994603; PMCID: PMC3185549.
833 38. Jahner D, Stuhlmann H, Stewart CL, Harbers K, Lohler J, Simon I, Jaenisch R. De novo
834 methylation and expression of retroviral genomes during mouse embryogenesis. Nature.
835 1982;298(5875):623-8. doi: 10.1038/298623a0. PubMed PMID: 6285203.
836 39. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-
837 body DNA methylation. Oncotarget. 2012;3(4):462-74. doi: 10.18632/oncotarget.497. PubMed PMID:
838 22577155; PMCID: PMC3380580.
839 40. Mao SQ, Cuesta SM, Tannahill D, Balasubramanian S. Genome-wide DNA Methylation
840 Signatures Are Determined by DNMT3A/B Sequence Preferences. Biochemistry. 2020;59(27):2541-50.

841 Epub 20200628. doi: 10.1021/acs.biochem.0c00339. PubMed PMID: 32543182; PMCID:
842 PMC7364778.
843 41. Mallona I, Ilie IM, Karemaker ID, Butz S, Manzo M, Caflisch A, Baubec T. Flanking sequence
844 preference modulates de novo DNA methylation in the mouse genome. Nucleic Acids Res.
845 2021;49(1):145-57. doi: 10.1093/nar/gkaa1168. PubMed PMID: 33290556; PMCID: PMC7797059.
846 42. Gao L, Emperle M, Guo Y, Grimm SA, Ren W, Adam S, Uryu H, Zhang ZM, Chen D, Yin J, Dukatz
847 M, Anteneh H, Jurkowska RZ, Lu J, Wang Y, Bashtrykov P, Wade PA, Wang GG, Jeltsch A, Song J.
848 Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de
849 novo DNA methylation mechanisms. Nat Commun. 2020;11(1):3355. Epub 20200703. doi:
850 10.1038/s41467-020-17109-4. PubMed PMID: 32620778; PMCID: PMC7335073.
851 43. Alvarez-Ponce D, Torres-Sanchez M, Feyertag F, Kulkarni A, Nappi T. Molecular evolution of
852 DNMT1 in vertebrates: Duplications in marsupials followed by positive selection. PLoS One.
853 2018;13(4):e0195162. Epub 20180405. doi: 10.1371/journal.pone.0195162. PubMed PMID:
854 29621315; PMCID: PMC5886458.
855 44. Stewart KR, Veselovska L, Kim J, Huang J, Saadeh H, Tomizawa S, Smallwood SA, Chen T,
856 Kelsey G. Dynamic changes in histone modifications precede de novo DNA methylation in oocytes.
857 Genes Dev. 2015;29(23):2449-62. Epub 20151119. doi: 10.1101/gad.271353.115. PubMed PMID:
858 26584620; PMCID: PMC4691949.
859 45. Ohtani H, Liu M, Zhou W, Liang G, Jones PA. Switching roles for DNA and histone methylation
860 depend on evolutionary ages of human endogenous retroviruses. Genome Res. 2018;28(8):1147-57.
861 Epub 20180703. doi: 10.1101/gr.234229.118. PubMed PMID: 29970451; PMCID: PMC6071641.
862 46. Boi S, Rosenke K, Hansen E, Hendrick D, Malik F, Evans LH. Endogenous retroviruses mobilized
863 during friend murine leukemia virus infection. Virology. 2016;499:136-43. Epub 20160919. doi:
864 10.1016/j.virol.2016.07.009. PubMed PMID: 27657834; PMCID: PMC5102782.
865 47. Ebrahimi D, Anwar F, Davenport MP. APOBEC3G and APOBEC3F rarely co-mutate the same
866 HIV genome. Retrovirology. 2012;9:113. Epub 20121220. doi: 10.1186/1742-4690-9-113. PubMed
867 PMID: 23256516; PMCID: PMC3532371.
868 48. Anwar F, Davenport MP, Ebrahimi D. Footprint of APOBEC3 on the genome of human
869 retroelements. J Virol. 2013;87(14):8195-204. Epub 20130522. doi: 10.1128/JVI.00298-13. PubMed
870 PMID: 23698293; PMCID: PMC3700199.
871 49. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D,
872 Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across
873 taxonomic space. Database (Oxford). 2011;2011:bar030. Epub 20110723. doi:
874 10.1093/database/bar030. PubMed PMID: 21785142; PMCID: PMC3170168.
875 50. Krueger F. Trim Galore! Available from:
876 https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
877 51. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq
878 applications. Bioinformatics. 2011;27(11):1571-2. Epub 20110414. doi:
879 10.1093/bioinformatics/btr167. PubMed PMID: 21493656; PMCID: PMC3102221.
880 52. Ngo V, Chen Z, Zhang K, Whitaker JW, Wang M, Wang W. Epigenomic analysis reveals DNA
881 motifs regulating histone modifications in human and mouse. Proc Natl Acad Sci U S A.
882 2019;116(9):3668-77. Epub 20190212. doi: 10.1073/pnas.1813565116. PubMed PMID: 30755522;
883 PMCID: PMC6397554.

884 53. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN,
885 Hinrichs AS, Lee CM, Nassar LR, Raney BJ, Speir ML, van Baren MJ, Vaske CJ, Haussler D, Kent WJ,
886 Haeussler M. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res.
887 2025;53(D1):D1243-D9. doi: 10.1093/nar/gkae974. PubMed PMID: 39460617; PMCID: PMC11701590.
888 54. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64. doi:
889 10.1101/gr.229202. PubMed PMID: 11932250; PMCID: PMC187518.
890 55. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat
891 Methods. 2015;12(3):265-72, 7 p following 72. Epub 20140921. doi: 10.1038/nmeth.3065. PubMed
892 PMID: 25240437; PMCID: PMC4344378.
893