Manuscript title:

Specialized DNA structures act as genomic beacons for integration by evolutionarily diverse retroviruses

Hinissan P. Kohio[1‡], Hannah O. Ajoge[1‡], Macon D. Coleman[1], Emmanuel Ndashimye[1], Richard M. Gibson[1], Eric J. Arts[1] and Stephen D. Barr[1*]

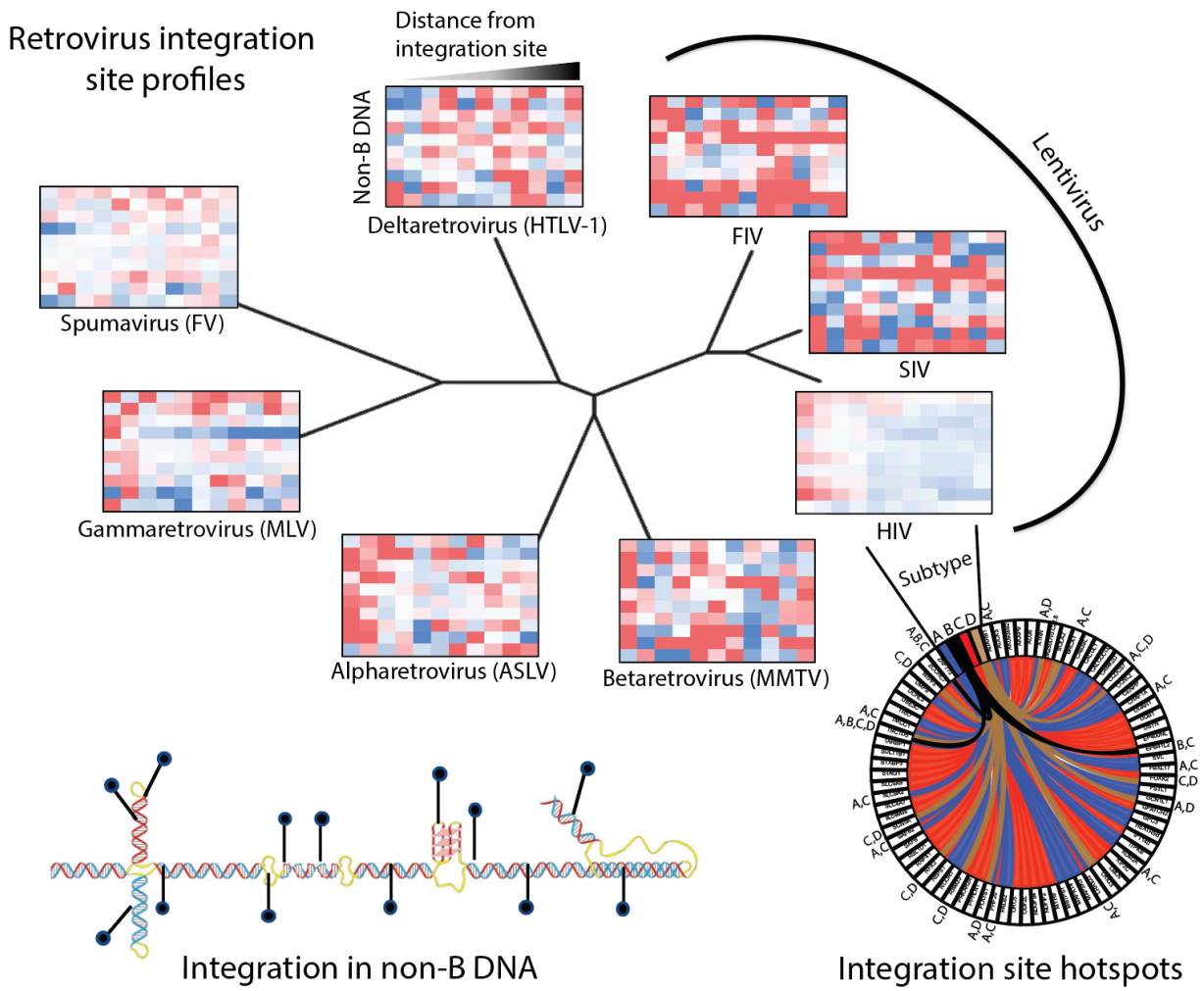
[1] Western University, Schulich School of Medicine and Dentistry, Department of Microbiology and Immunology, Dental Sciences Building Room 3007, London, Ontario, Canada.

* To whom correspondence should be addressed. Tel: 01-519-661-3438; Fax: 01-519-661-3499; Email: stephen.barr@uwo.ca

‡ The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint first authors

ABSTRACT

Retroviral integration site targeting is not random and plays a critical role in expression and long-term survival of the integrated provirus. To better understand the genomic environment surrounding retroviral integration sites, we performed an extensive comparative analysis of new and previously published integration site data from evolutionarily diverse retroviruses from seven genera, including different HIV-1 subtypes. We showed that evolutionarily divergent retroviruses exhibited distinct integration site profiles with strong preferences for non-canonical B-form DNA (non-B DNA). Whereas all lentiviruses and most retroviruses integrate within or near genes and non-B DNA, MMTV and ERV integration sites were highly enriched in heterochromatin and transcription-silencing non-B DNA features (e.g. G4, triplex and Z-DNA). Compared to *in vitro*-derived HIV-1 integration sites, *in vivo*-derived sites are significantly more enriched in transcriptionally silent regions of the genome and transcription-silencing non-B DNA features. Integration sites from individuals infected with HIV-1 subtype A, C or D viruses exhibited different preferences for non-B DNA and were more enriched in transcriptionally active regions of the genome compared to subtype B virus. In addition, we identified several integration site hotspots shared between different HIV-1 subtypes with specific non-B DNA sequence motifs present at these hotspots. Together, these data highlight important similarities and differences in retroviral integration site targeting and provides new insight into how retroviruses integrate into genomes for long-term survival.



Graphical Abstract:
Schematic comparing integration site profiles from evolutionarily diverse retroviruses. Upper left, heatmaps showing the fold-enrichment (blue) and fold-depletion (red) of integration sites near non-B DNA features (lower left). Lower right, circa plot showing integration site hotspots shared between HIV-1 subtype A, B, C and D virus.

INTRODUCTION

Retroviruses are divided into seven genera *alpha*-, *beta*-, *gamma*-, *delta*-, and *epsilon*-*Retroviridae*, *Spumaviridae*, and *Lentiviridae*). Following entry of a retrovirus into cells, the viral RNA genome is converted into double-stranded complimentary DNA (cDNA). In the cytoplasm, the cDNA associates with several viral and host proteins to form a pre-integration complex (PIC). Soon after completion of cDNA synthesis, the viral DNA ends are primed by the enzyme integrase in a process called 3' processing. After docking with a host chromosome, the viral cDNA undergoes a strand transfer reaction resulting in the insertion of the retroviral genome into the chromosome. Host repair enzymes are then thought to complete integration (*1*). Insertion of the retroviral genome results in a persistent life-long infection. Selection of integration sites in the genome by the PIC is not random. The gamma- and deltaretroviruses (e.g. murine leukemia virus (MLV), human T cell leukemia virus Type 1 (HTLV-1)) and foamy virus (FV) favor integration around *transcription start sites* (*TSS*) (*2–4*). Alpharetroviruses (e.g. avian sarcoma leukosis virus (ASLV)) and human endogenous retroviruses (ERVs) show no strong preferences, with integration only slightly favored in transcription units or the 5' end of genes (*5–8*).

Lentiviruses such as human immunodeficiency virus type 1 (HIV-1), simian immunodeficiency virus (isolated from a pig-tailed macaque) (SIV_{mne}) and feline immunodeficiency virus (FIV) strongly favor integration in active transcription units (*9–11*). HIV integrations sites are also associated with regions of high G/C content, high gene density, short introns, high frequencies of short interspersed nuclear elements (*SINEs*) (e.g. Alu repeats), low frequencies of long interspersed nuclear elements (*LINEs*), and characteristic epigenetic modifications (*9*, *12*). Thus far, integration site analyses have only been conducted on HIV-1 subtype B infections. Based on phylogenetic analyses of full-length genomic sequences, HIV-1 isolates are classified into four distinct groups: group M, N, O and P (*13*). HIV-1 M group accounts for the majority of the global pandemic and is subdivided into nine subtypes or clades (A, B, C, D, F, G, H, J and K). Additionally, several circulating recombinant forms (CRF) and unique recombinant forms (URF) have been identified, which are the result of a recombination event between two or more different subtypes. HIV-1 geographical prevalence is extremely diverse. Subtype C represents more than 50% of the infection worldwide and is prevalent in Africa and Asia. Subtype B infection is common in the Americas, Europe, Australia, and part of South Asia, Northern Africa and the Middle East. Subtypes A, D, F, G, H, J and K occur mostly in Sub-Saharan Africa, whereas infections with groups N, O and P have been found in confined regions of West-Central Africa.

Several models, not mutually exclusive, have been proposed to explain integration site selection. In the chromatin accessibility model, the structure of chromatin influences accessibility of target DNA sequences to PICs. *In vivo* target DNA is not expected to be naked but rather wrapped in nucleosomes. Wrapping target DNA in nucleosomes does not reduce integration, but instead creates hotspots for integration at sites of probable DNA distortion (*14*, *15*). Distortion of DNA in several other protein-DNA complexes has also been shown to favour integration in the major grooves facing outwards from the nucleosome core (*16*, *17*). Although chromatin structure can facilitate integration, chromatin accessibility cannot solely explain the differences observed in integration site preferences.

The protein tethering model suggests that a cellular protein, potentially specific for each retroviral genera, act as tethering factors between chromatin and the PIC. The most characterized tethering factor identified to date is lens epithelium-derived growth factor and co-factor p75 (LEDGF/p75) (also known as PSIP1/p75) (*18–20*). LEDGF/p75 interacts with HIV integrase and tethers the PIC to genomic DNA in transcriptionally active genes marked by specific histone modifications such as H3K20me1, H3K27me1 and H3K36me3. In similar fashion, host protein bromodomain/extraterminal domain proteins (BETs) bound to acetylated histones (e.g. H327ac and H3K9ac) interact with MLV integrase and tethers the PIC to genomic DNA in transcriptionally active promoters, enhancers and super enhancers (*21–24*). LEDGF/p75 and BET depletion studies demonstrated that integration still occurs but with reduced efficiency and an altered integration site selection profile (*20*, *22–28*). Several cellular proteins have been proposed to facilitate integration including barrier to autointegration factor (BAF), high mobility group A1 (HMGA1), integrase interactor 1 (Ini-1), and heat shock protein 60 (Hsp60) (*29–31*). Some of these proteins might contribute to PIC function by coating and condensing the viral DNA, thereby assisting the assembly of the viral nucleoprotein complexes. Several cellular chromatin proteins have also been suggested to influence integration such as Ini-1, EED, SUV39H1, and HP1γ (reviewed in references (*12*, *32*, *33*)). Cleavage and polyadenylation specificity factor 6 (CPSF6) facilitates HIV-1 PIC import and helps direct the PIC to the nuclear interior to locate gene-dense euchromatin for integration (*34–38*). Recently, Apolipoprotein B Editing Complex (APOBEC3) was identified as a potential new host factor that also influences integration site selection by promoting a more transcriptionally silent integration site profile (*39*). Moreover, just as DNA-binding proteins can promote integration, they can also block access of integration complexes, creating regions refractory for integration (*16*, *40*, *41*). Binding to different factors or differential binding affinity to a factor by the PIC could modulate integration site selection between lentivirus types or even between different HIV-1 subtypes.

Previous analyses of primary DNA sequences (~20 base pairs (bp)) flanking integration sites only revealed a weak consensus motif (*42*). We previously assessed a broader window of primary sequence (80 bp) around HIV-1 integration sites discovered that HIV-1 integration sites were highly enriched near specialized genomic features called non-B DNA (*43*). Non-B DNA form functional secondary structures in our genome formed by specific nucleotide sequences that exhibit non-canonical DNA base pairing. At least 10 non-B DNA conformations are identified including inverted repeats, direct repeats, mirror repeats, short-tandem repeats, guanine-quadruplex (G4), A-phased, cruciform, slipped, triplex and Z-DNA (*44*, *45*). Here we present a comparative analysis of integration site profiles of evolutionarily diverse retroviruses and identify striking similarities and differences among all retroviruses, especially towards non-B DNA.

MATERIAL AND METHODS

Ugandan and Zimbabwean cohort description. Details pertaining to the Uganda study population have been reported previously (*46–49*). Briefly, women who became HIV infected while participating in the Hormonal Contraception and Risk of HIV Acquisition Study in Uganda were enrolled upon

114 primary infection with HIV-1 into a subsequent study, the Hormonal Contraception and HIV-1 Genital
115 Shedding and Disease Progression among Women with Primary HIV Infection (GS) Study. Ethical
116 approval was obtained from the Institutional Review boards (IRBs) from the Joint Clinical Research
117 Centre and UNST in Uganda, from University of Zimbabwe, from the University Hospitals of
118 Cleveland, and recently, from Western University. All adult subjects provided written informed consent
119 and no child participants were included in the study. Protocol numbers and documentation of these
120 approvals/renewals are available upon request. Blood and cervical samples were collected every
121 month for the first six months, then every three months for the first two years, and then every six
122 months up to 9.5 years. Women who had CD4 lymphocyte counts of 200 cells/ml and/or who
123 developed severe symptoms of HIV infection (WHO clinical stage IV or advanced stage III disease)
124 were offered combination antiretroviral therapy (cART) and trimethoprim-sulfamethoxazole (for
125 prophylaxis against bacterial infections and *Pneumocystis jeroveci* pneumonia).
126
127 ***HIV-1 integration site library.*** Total genomic DNA was extracted from PBMCs using the QIAmp DNA
128 mini kit (Qiagen) and processed for integration site analysis and sequenced using the Illumina MiSeq
129 platform as previously described (*43*, *50*). Samples were sequenced at the London Regional
130 Genomics Centre/Robarts Research Institute (Western University, Canada) and Case Western
131 Reserve University (USA).
132
133 ***Computational analysis***. Fastq sequencing reads were quality trimmed and unique integration sites
134 identified using our in-house bioinformatics pipeline Barr Lab Integration Site Identification Pipeline
135 (BLISIP version 2.9) and included the following updates: bedtools (v2.25.0), bioawk (awk version
136 20110810), bowtie2 (version 2.3.4.1), and restrSiteUtils (v1.2.9) (*39*, *43*). All non-B DNA motifs were
137 defined according to previously established criteria (*51*). The G-quadruplex algorithm identifies four or
138 more individual G-runs of at least three nucleotides in length. The algorithm requires at least one
139 nucleotide between each run and considers up to seven nucleotides as spacer, including guanines
140 (*51*). Lamina-associated domains (*LADs*) were retrieved from http://dx.doi.org/10.1038/nature06947
141 (*52*). HIV-1 LTR-containing fastq sequences were identified and filtered by allowing up to a maximum
142 of five mismatches with the reference NL4-3 LTR sequence and accepted if the LTR sequence had
143 no match with any region of the human genome (GRCh37/hg19). Matched random control integration
144 sites (28,800 in total) were generated by matching each experimentally determined site with 50
145 random sites *in silico* that were constructed to be the same number of bases from the restriction site
146 as was the experimental site, as previously described (*53*). All genomic sites in each dataset that
147 hosted two or more sites (i.e. identical sites) were collapsed into one unique site for the analysis.
148 Sites that could not be unambiguously mapped to a single region in the genome were excluded from
149 the study. Integration site profile heatmaps were generated using our in-house python program
150 BLISIP heatmap (BHmap v1.0).
151

152 ***Data and software availability.*** The sequences reported in this paper will be deposited in the
153 National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) upon acceptance
154 for publication.

155

156 **RESULTS**

157 **Integration site dataset acquisition and analyses**

158 Previous integration site analyses of evolutionarily diverse retroviruses identified three general
159 integration site profiles in which retroviruses can be grouped. For example, HIV-1, SIV_{mne} and FIV
160 integrate into transcriptionally active genes; MLV, HTLV-1 and FV integrate near *TSS*; ASLV, HTLV
161 and MMTV show little preference for any genomic feature. The integration site profiles of these
162 evolutionarily diverse retroviruses with respect to non-B DNA was previously unknown. To determine
163 integration site profiles of these viruses, including *endogenous retroviruses* (*ERVs*) present within the
164 human genome, we analyzed ~169,910 integration sites from previously published datasets (**Table S1**).
165 To allow for a more comparable analysis of integration site profiles, we updated these integration site
166 profiles using the human genome assembly GRCh37/19. Our update also included an assessment of
167 genomic features not included in some of the previous retroviral integration site studies. We used our
168 previously developed in-house bioinformatics pipeline called the Barr Lab Integration Site Pipeline
169 (BLISIP) to generate integration site profiles from the different retroviral integration site datasets (*39*,
170 *43*, *50*). BLISIP measures integration site enrichment near the genomic features CpG islands, *DNAseI*
171 *hypersensitivity sites* (*DHS*), *ERVs*, heterochromatic DNA regions (e.g. *lamin-associated domains*
172 (*LADs*) and satellite DNA), *SINEs* and *LINEs* respectively, *low complexity repeats* (*LCRs*), oncogenes,
173 genes, simple repeats and *TSS*. In addition, BLISIP measures enrichment near the non-B DNA features
174 A-phased motifs, cruciform motifs, direct repeats, G4 motifs, inverted repeats, mirror repeats, short
175 tandem repeats, slipped motifs, triplex motifs and Z-DNA motifs. Our analyses focused on unique
176 integration sites and excluded sites arising from clonal expansion, sites falling in repeat regions that
177 could not be uniquely identified, and regions that could not be confidently placed on a specific
178 chromosome (e.g. ChrUn). Enrichment of integration sites within genomic features was determined by
179 comparing the proportion of sites with either a matched random control (MRC) to account for restriction
180 site bias in the cloning procedure during library construction, or a random control (RC) for comparison
181 of datasets that used DNA shearing/fragmentation during library construction (**Table S1**).

182

183 **Evolutionarily divergent retroviruses exhibit distinct integration site profiles**

184 Integration sites were quantified and placed in five bins based on their distance from each
185 genomic feature (within the feature, 1-499 bp, 500-4,999 bp, 5,000-49,999 bp, and >49,000 bp).
186 Heatmaps from each retrovirus showing the fold-enrichment and fold-depletion in each bin were
187 compared to MRC or RC (**Figure 1A**). Consistent with previous studies (*9–11*), HIV-1, SIV_{mne} and FIV
188 integration sites are significantly enriched within genes (64%, 84%, 90% respectively) ($P < 0.0001$)
189 **(Figure 1A and 1B, Table S2)**. Our analysis of HTLV-1, ASLV and MLV also agreed with previous
190 reports, confirming that these viruses exhibit only modest preferences for integration within *genes* with

191 54%, 56% and 56% of integration sites found within *genes* respectively **(Figure 1A and 1B, Table S2)**
192 (*54*, *55*)**.** In contrast, MMTV, FV and ERVs showed no preference for integration into genes (47%, 43%
193 and 32% respectively). No retrovirus showed a preference for integration directly into *TSS*; however,
194 MLV, HTLV-1 and FV showed significant enrichment of integration sites near *TSS* and CpG islands (*P*
195 <0.0001) **(Figure 1A, Table S2)**. Repetitive elements, such as *LINEs*, *SINEs*, *ERVs* (e.g.
196 retrotransposons), satellite DNA, simple repeats (e.g. microsatellites), and *LCRs* account for nearly half
197 of the human genome sequence. However, no strong preference for integration into these regions was
198 observed for any of the retroviruses except for HIV-1 and FV which targeted *SINEs*, and FV, MMTV
199 and HTLV-1 which targeted satellite DNA **(Figure 1A).** Pairwise analyses of the different integration
200 site profile preferences (fold enrichment and depletion values) showed that HIV-1, MMTV and FIV had
201 the most similar integration site preferences for genomic features overall, whereas SIV, MLV and ERVs
202 had the least similarity to the other retroviruses (**Figure 1C**).
203
204 **MMTV and ERVs strongly favour integration into heterochromatin**
205 The nuclear architecture influences HIV-1 integration site selection and proviral expression (*52*,
206 *56*). HIV-1 strongly disfavors integration into heterochromatin positioned in *LADs* at the nuclear
207 periphery, although some integration does occur in these regions, contributing to the latent reservoir
208 (*57*). We asked whether other retroviruses similarly disfavor integration into *LADs*. Like HIV-1, most
209 retroviruses significantly disfavored integration into *LADs* with only 11-28% of sites falling within *LADs*
210 **(Figure 1D, Table S2).** In contrast, 48% of MMTV and 44% of ERV integration sites were significantly
211 enriched in *LADs*.
212
213 **Evolutionarily divergent retroviruses target non-B DNA for integration**
214 Non-B DNA is a genomic correlate of HIV-1 integration site selection, but its influence on
215 integration site targeting of other retroviruses was previously unknown (*43*). Analysis of the different
216 retroviral integration site profiles showed that all retroviruses exhibited enriched integration near non-B
217 DNA (**Figure 2A and Table S3**). FV, HTLV-1 and SIV_{mne} were the only retroviruses with an enrichment
218 of sites within 50 bp of multiple non-B DNA features, whereas other retroviruses tended to integrate
219 more distal (100-500 bp) to the features. HIV-1 integration sites were enriched near all non-B DNA
220 features except A-phased DNA. SIV_{mne} and FIV sites were enriched near inverted repeats, mirror
221 repeats and short tandem repeats. HTLV-1 sites were enriched near A-phased, cruciform and inverted
222 repeats. FV sites were enriched near G4 and Z-DNA. MLV sites were enriched near G4, triplex and Z-
223 DNA. ASLV sites were enriched in slipped and triplex DNA. MMTV sites were enriched in short-tandem
224 repeats, G4, slipped, triplex and Z-DNA. ERV sites were enriched near cruciform and triplex DNA.
225 Pairwise analyses of the different integration site profiles showed that HIV-1 and FV had the most similar
226 preferences for targeting non-B DNA, whereas SIV_{mne}, ERVs and FIV had the least similarity to the
227 other retroviruses (**Figure 2B**). Taken together, these data show that evolutionarily diverse retroviruses
228 target non-B DNA for integration and that each virus exhibits distinct preferences for certain non-B DNA
229 features.
230

231 **HIV-1 Integration site profiles differ between *in vitro*- and *in vivo*-derived datasets**

232 Despite a large number of integration sites from HIV-1 infected individuals, many of the previous
233 integration site studies have been performed on infections done *in vitro* (*12*, *58*). To determine if the
234 HIV-1 integration site profiles from *in vitro*-derived infections differed from those from *in vivo*-derived
235 infections, we analyzed and compared nine previously published HIV-1 integration site datasets from
236 publicly available databases, totalling 13,601 *in vitro*-derived sites and 139,480 *in vivo*-derived sites
237 **(Table S1)**. As expected, integration sites were significantly enriched in genes compared to the random
238 controls in both the *in vitro*- and *in vivo*-derived datasets; however, only 62% of total *in vivo*-derived
239 integration sites were in genes compared to 83% in the *in vitro* dataset ($P < 0.0001$) (**Figure 3A, 3B**
240 **and Tables S4**). In contrast with the *in vitro* dataset, integration sites in the *in vivo*-derived dataset were
241 highly enriched near most non-B DNA except A-phased DNA **(Figure 3C and Table S4)**. Notably, *in*
242 *vivo*-derived sites were significantly more enriched near G4, triplex and Z-DNA than *in vitro*-derived
243 sites **(Figure 3C, 3D and Table S4)**. Together, these data show that there are substantial differences
244 in HIV-1 integration site targeting preferences between *in vitro*-derived and *in vivo*-derived datasets.
245

246 **Integration site profiles differ in individuals infected with HIV-1 subtype A, B, C or D**

247 Thus far, integration site profiles have been extensively analyzed for HIV-1 subtype B infections,
248 which represents only ~10% of the infections worldwide. We asked if the integration site profiles from
249 individuals infected with HIV-1 non-subtype B virus were similar to those infected with subtype B virus.
250 In the integrase enzyme, the amino acid difference between subtypes is as high as 16% (subtype B vs.
251 C) but typically less than 8% in IN of HIV-1 isolates of a specific subtype. This level of amino acid
252 diversity is the highest among the enzymes encoded by the HIV-1 *pol* gene. The greatest diversity
253 within HIV-1 IN is found in the C-terminal domain (CTD), which is involved in genomic DNA binding.
254 Thus, it is reasonable to propose that we would observe differences in selectivity for integrations in non-
255 B form DNA. Genomic DNA was isolated from peripheral blood mononuclear cells (PBMCs) from a
256 cohort of women in Uganda and Zimbabwe infected with HIV-1 subtype A, C or D and used to generate
257 integration site libraries. Integration site profiles were generated from a total of 48 infected individuals
258 (16 subtype A, 19 subtype C and 13 subtype D) and compared to the integration site profile from 14
259 individuals infected with subtype B virus generated from previously published datasets **(Table S5)** (*59*,
260 *60*)**.**

261 As previously observed with HIV-1 subtype B infections, integration sites from all HIV-1 subtype
262 viruses were enriched in genes (**Figure 4A and 4B and Table S6**). Notably, subtypes A, C and D had
263 significantly more integration sites in genes compared to subtype B (A:82%, B:63%, C:71%, D:78%)
264 (**Figure 4C**). Integration sites for all subtypes were enriched near *DHS*; however, subtypes A, C and D
265 exhibited a weaker preference compared to subtype B. In addition, subtypes A, C and D had
266 significantly less integration in heterochromatin (e.g. *LADs*) and *SINEs* compared to subtype B, which
267 likely reflects their higher integration into genes. Pairwise analyses of the different integration site
268 profiles showed that subtypes C and D, and B and D, shared the most similarity to each other, and
269 subtype A differed the most from the other subtypes (**Figure 4D**). This relates in part to the sequence
270 distance between IN coding regions of subtypes A, B, C, and D, with B and D being most genetically

271 similar (**Figure 4E and 4F**). However, it is important to stress that any slight differences in integration
272 profiles could also relate to discreet single amino acid differences between subtypes, especially in the
273 CTD domain and/or the LEDGF binding domain.
274 Analysis of the non-B DNA integration site profiles from the different subtype viruses showed
275 enriched integration near non-B DNA (**Figure 5A, 5B and Table S7**). Notably, subtypes A and C had
276 more integration sites near A-phased DNA compared to subtypes B and D. Subtypes A, C and D virus
277 had significantly less sites near G4 DNA compared to subtype B virus. In addition, subtype C had
278 significantly less sites near triplex and Z-DNA. Pairwise analysis of the different non-B DNA integration
279 site profiles showed that the profiles of all subtypes differed substantially from each other, with subtypes
280 B and D showing slightly more similarity to each other compared to the other subtypes (**Figure 5C**).
281 Together, these data show that the integration site profiles from individuals infected with HIV-1 subtype
282 B differs from those infected with non-subtype B virus, especially with respect to their preferences for
283 non-B DNA. As discussed below, these slight but significant differences in selection for integration at
284 different non-B DNA features may relate to discreet amino acid differences in the CTD domains of these
285 IN that evolved in different subtypes. Although there are trends suggesting impact by discreet amino
286 acid differences, for significance, we need to analyze more integration patterns by more subtype A, C,
287 and D infections with known IN sequences.
288
289 **Integration hotspots are shared between HIV-1 subtypes**
290 The concept of an HIV-1 integration "hotspot" was introduced to describe areas of the
291 genome where integrations accumulate more than expected by chance in the absence of any
292 selection process (*9*, *61*). We catalogued and analyzed 1,000 bp windows containing two or more
293 integration sites from each HIV-1 subtype dataset. This yielded a total of 10,892 hotspots between all
294 subtypes (A: 41/429, 9.6%; B: 10,762/118,727, 9.1%; C: 74/484, 15.3%; D: 15/323, 4.6%; and MRC:
295 2/2982, 0.07%) (**Figure 6A**). 16% of the hotspots were shared between subtypes A and C, 14%
296 between subtypes A and D and 10% between subtypes C and D. Despite the large number of
297 hotspots identified in the subtype B dataset, less than 0.03% of the hotspots were shared between
298 subtype B and the other subtypes. Comparison of hotspots revealed one hotspot in the *tbc1d5* gene
299 that was shared by all subtypes (**Figure 6B and 6C**).
300 To identify potential local genomic sequences that may serve as 'beacons' for increased
301 integration, we analyzed a nucleotide window of 100 bp upstream and 100 bp downstream of each
302 integration site and generated a consensus sequence using DiffLogo (**Figure 6D**) (*62*). Comparison of
303 the sequence motifs surrounding integration sites located in hotspots with the sequence motifs from
304 sites not located in hotspots revealed the presence of a slipped DNA motif at the integration site in
305 each of the HIV-1 subtype datasets. In addition, a G4 motif was located ~40-100 bp downstream from
306 the integration site in each of the different HIV-1 subtype datasets. Subtypes B and C also had a
307 triplex motif and subtype B had a short tandem repeat located downstream of the integration site.
308 Together, these data show that sites located in integration hotspot regions are located in slipped DNA
309 motifs flanked downstream by other non-B DNA motifs.
310

311 **DISCUSSION**

312 Here we showed that evolutionarily diverse retroviruses from all seven genera exhibit strong
313 preferences for integration within or near non-B DNA features. Whereas all lentiviruses and most
314 retroviruses integrate within or near genes and a preference for non-B DNA, we show that MMTV and
315 ERV integration sites are highly enriched in heterochromatin but with still a preference for non-B DNA.
316 Nonetheless, each retrovirus has distinct preferences for integration at specific types of non-B DNA.
317 Highly distinct preferences for these non-B motifs among different HIV-1 subtypes is less apparent
318 and yet, divergent IN evolution may still result in some differential selection/inhibition of HIV DNA
319 integration in host cellular DNA. Finally, we showed that certain regions of the genome attract more
320 integration sites than others and that the integration sites in these regions are located in slipped DNA
321 motifs and flanked downstream by other non-B DNA motifs.

322 All retroviruses exhibited enriched integration near non-B DNA with specific non-B DNA
323 features (e.g. slipped, G4, triplex and short-tandem repeat DNA) serving as integration site hotspots.
324 How retroviral PICs recognize non-B DNA for integration is not fully understood. Non-B DNA motifs are
325 enriched near endogenous retroviruses, are near sites of genomic variability, and are targets for
326 preferential homologous recombination (>20-fold) in human cells (*63–65*). It is possible that integrase
327 itself, components of the PIC and/or host proteins can bind non-B DNA directly and promote integration
328 in these regions. We recently showed that PIC-binding host factors APOBEC3 (*39*), LEDGF/p75 and
329 CPSF6 (submitted) influence the distribution of integration sites near non-B DNA features, potentially
330 indicating that a host-directed mechanism is involved. Intriguingly, APOBEC3G binding and
331 deamination hotspots also comprise slipped DNA motifs (*66*, *67*). It is possible that viral integrase is
332 also involved in recognizing non-B DNA given that HIV-1 integrase has been shown to bind directly to
333 G4 DNA (*68*).

334 Analysis of integration site profiles after acute infection showed that HIV-1, FIV and SIV_{mne} had
335 the strongest preference for integrating into transcriptionally active regions of the genome, whereas
336 MMTV, FV and ERVs had the strongest preference for transcriptionally inactive regions. The targeting
337 of transcriptionally active regions of the genome is believed to help maximize provirus expression for
338 establishing infection and promoting its spread (*9*). Over time, especially with the help of antiretroviral
339 drugs, the immune system likely clears these high virus-producing cells leaving behind infected cells
340 that produce little to no virus that escape detection by the immune system. The integration profile of
341 these latently infected cells showed that sites are located in more transcriptionally silent regions of the
342 genome, similar to the profile shown here for ERVs and after acute infection with MMTV and FV.
343 Consistent with the latter observation, latently infected cells have reduced integration in genes and
344 increased integration in *SINEs* and heterochromatin, typically found in gene deserts or transcriptionally
345 silent regions (*57*, *60*). Transcriptionally silent regions of the genome are characterized by several
346 features. For example, *LADs* represent a repressive chromatin environment tightly associated with the
347 nuclear periphery (*56*, *57*). *SINEs* (e.g. Alu repeats) and other transposed sequences are known to
348 serve as direct silencers of gene expression due to their repressed chromatin marks (histone H3
349 methylated at Lys 9) (*69*, *70*). Moreover, non-B DNA structures such as G4, cruciform, triplex and Z-
350 DNA have been shown to silence expression of adjacent genes (*65*, *71*, *80*, *72–79*). The ability to target

351 silent regions of the genome for integration could be a long-term survival mechanism for both
352 retroviruses and their hosts. By integrating into transcriptionally silent regions of the genome, these
353 viruses can minimize their expression and avoid detection by the immune system. As an extreme
354 example, endogenous retroviruses, some of which are ancient viruses that have been co-evolving with
355 their hosts for hundreds of millions of years (e.g. ERVs and FV), are commonly integrated in
356 transcriptionally silent regions of the genome away from genes (Figures 1 and 2) (*81–83*). The host also
357 benefits with a lower risk of disease. This is seen with FV which persists in their primate hosts in the
358 absence of disease, and MMTV which typically does not cause cancer unless it integrates near an
359 oncogene (*84*). It is currently unknown if integration in transcriptionally silent regions of the genome is
360 driven by retroviruses, the host or both. Interestingly, host APOBEC3G and APOBEC3F appear to
361 promote a transcriptionally silent integration site profile of HIV-1, suggesting that the host may indeed
362 contribute to this silent phenotype (*39*).

363 Genomic position effects have been shown to influence HIV-1 expression, which can ultimately
364 impact disease progression (*57*, *85–87*). As described above, HIV-1 infection of a host cell results in
365 preferential integration in actively transcribing genes, due in part to reduced chromatin structure. Thus,
366 it is not surprising to observe this finding following *in vitro* infections. However, even *in vitro*, there is
367 evidence of preferential integration near non-B DNA motifs. Within a human host, enhanced HIV-1
368 integration within genes remains evident but now, there is greater enrichment of those exact integration
369 sites being near transcription-silencing non-B DNA motifs, suggesting an enrichment in the pool of HIV-
370 1 proviral DNA within transcriptionally silent sites in the genome. As described above, this may be the
371 consequence of the immune system consistently clearing away cells propagating HIV-1 while the cells
372 with silent integrants are slowly enriched. HIV-1 integration sites from these *in vitro* experiments were
373 typically analyzed from cells infected for 2-3 days, whereas the *in vivo*-derived integration sites were
374 obtained from chronically infected individuals whose cells contain integrated proviruses that have
375 undergone more than 2 months of selection in the host. Consistent with previous findings (*57*, *60*), HIV
376 integrants were found enriched within features associated with reduced expression such as *LADs*,
377 satellite DNA and *SINEs* when compared to *in vitro*-derived sites. Moreover, we identified a sequence-
378 based integration site bias towards non-B DNA, particularly transcription silencing G4, triplex and Z-
379 DNA.

380 Much of our knowledge of HIV-1 integration site selection has come from studies using subtype
381 B virus. Our analyses showed that HIV-1 non-subtype B viruses have a much stronger preference for
382 integrating into transcriptionally active regions of the genome compared to subtype B virus. This was
383 characterized by increased integration in genes and decreased integration in genomic features
384 associated with transcriptional silencing such as *LADs*, *SINEs*, G4, triplex and Z-DNA. Despite the
385 enriched proviral integration into non-B DNA among all HIV-1 subtypes, there was some integration site
386 preferences that differed between each of the HIV-1 subtypes. In the integrase coding region, amino
387 acid differences between subtypes are as a high as 16% (subtype B vs. C) but typically less than 8%
388 IN diversity among HIV-1 isolates of one subtype (*13*, *88*, *89*). This level of amino acid diversity is the
389 highest among the enzymes encoded by the HIV-1 *pol* gene. The greatest diversity within HIV-1 IN is
390 found in the CTD, which is involved in genomic DNA binding. Thus, it is reasonable to suspect that HIV-

391 1 subtypes may target different non-B DNA motifs with differential selectivity. Despite the multiple
392 substitutions that segregate HIV-1 variants of each subtype, there may be specific amino acid
393 differences in one subtype versus another that impact preference for integration sites. For example, the
394 isoleucine found at position 20 in subtype C and D versus the threonine in subtype A and B in the IN
395 CTD may relate the greater similarity of integration site selection of subtypes C and D (Figure 4E and
396 4F). Despite the closer genetic relationship between HIV-1 subtype B and D across the genome, similar
397 genetic distances separating subtypes A, B, C, and D are within the IN coding region. Polymorphisms
398 in HIV-1 integrase have been reported that retarget integration away from gene dense regions, which
399 also correlated with increased disease progression and virulence (*87*). This suggests that
400 characteristics of integrase itself is a driver of integration site selection in retroviruses. We are currently
401 exploring if subtype-specific polymorphisms in integrase may account for differences in the targeting of
402 transcriptionally silent regions of the genome and of non-B DNA.

403 In conclusion, we identified non-B DNA as a feature surrounding integration sites that are
404 targeted differentially by evolutionarily diverse retroviruses. We have also presented the first
405 comparative look at the integration site profiles of several HIV-1 non-subtype B viruses and showed
406 that they differed from HIV-1 subtype B profiles. Together, this data highlights important similarities and
407 differences in retroviral integration site targeting that can be used in future studies to better understand
408 the evolution of retroviral integration site targeting and how these viruses integrate into our genomes
409 for long-term survival.

410

411 **FUNDING**

412 This work was supported by the Canadian Institutes of Health Research (CIHR) Operating Grants
413 [IBC-150406 and HBF-143164] to S.D.B.; and in part by CIHR [385787, 377790 to E.J.A]; Canada
414 Research Chair Tier 1 [230811 to E.J.A.]; National Institute of Allergy and Infectious Diseases-
415 National Institutes of Health [AI49170 to E.J.A].

416

417 **CONFLICT OF INTEREST**

418 The authors declare no competing interests.

419

420 **REFERENCES**

421 1. R. Daniel *et al.*, Evidence that stable retroviral transduction and cell survival following DNA
422 integration depend on components of the nonhomologous end joining repair pathway. *J Virol*.
423 **78**, 8573–8581 (2004).
424 2. X. Wu, Y. Li, B. Crise, S. M. Burgess, Transcription start regions in the human genome are
425 favored targets for MLV integration. *Science*. **300**, 1749–51 (2003).

426 3. B. Felice *et al.*, Transcription Factor Binding Sites Are Genetic Determinants of Retroviral
427 Integration in the Human Genome. *PLoS One*. **4**, e4571 (2009).
428 4. G. D. Trobridge *et al.*, Foamy virus vector integration sites in normal human cells. *Proc. Natl.*
429 *Acad. Sci. U. S. A.* **103**, 1498–503 (2006).
430 5. S. D. Barr, J. Leipzig, P. Shinn, J. R. Ecker, F. D. Bushman, Integration targeting by avian
431 sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J. Virol.*
432 **79**, 12035–12044 (2005).
433 6. R. Mitchell *et al.*, Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site
434 Preferences. *PLoS Biol.* **2**, E234 (2004).
435 7. A. Narezkina *et al.*, Genome-wide analyses of avain sarcoma virus integration sites. *J.Virol.*
436 **78**, 11656–11663 (2004).
437 8. T. Brady *et al.*, Integration target site selection by a resurrected human endogenous retrovirus.
438 *Genes Dev.* **23**, 633–642 (2009).
439 9. A. Schroder *et al.*, HIV-1 Integration in the Human Genome Favors Active Genes and Local
440 Hotspots. *Cell*. **110**, 521–529 (2002).
441 10. B. Crise *et al.*, Simian Immunodeficiency Virus Integration Preference Is Similar to That of
442 Human Immunodeficiency Virus Type 1. *J. Virol.* **79**, 12199–12204 (2005).
443 11. Y. Kang *et al.*, Integration Site Choice of a Feline Immunodeficiency Virus Vector. *J. Virol.* **80**,
444 8820–8823 (2006).
445 12. F. Bushman *et al.*, Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3**,
446 848–58 (2005).
447 13. B. Taylor *et al.*, The Challenge of HIV-1 Subtipe Diversity. *N. Engl. J. Med.* **358**, 1590–1602
448 (2008).
449 14. D. Pruss, F. D. Bushman, A. P. Wolffe, Human immunodeficiency virus integrase directs
450 integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci.*
451 *U. S. A.* **91**, 5913–7 (1994).
452 15. D. Pruss, R. Reeves, F. D. Bushman, A. P. Wolffe, The influence of DNA and nucleosome
453 structure on integration events directed by HIV integrase. *J. Biol. Chem.* **269**, 25031–41
454 (1994).
455 16. Y. C. Bor, F. D. Bushman, L. E. Orgel, In vitro integration of human immunodeficiency virus
456 type 1 cDNA into targets containing protein-induced bends. *Proc. Natl. Acad. Sci. U. S. A.* **92**,
457 10334–10338 (1995).
458 17. H.-P. Muller, H. E. Varmus, DNA bending creates favored sites for retroviral integration: an
459 explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**, 4704–4714 (1994).
460 18. A. Engelman, P. Cherepanov, The lentiviral integrase binding protein LEDGF/p75 and HIV-1
461 replication. *PLoS Pathog.* **4**, e1000046 (2008).
462 19. E. M. Poeschla, Integrase, LEDGF/p75 and HIV replication. *Cell. Mol. Life Sci.* **65**, 1403–24
463 (2008).
464 20. M. Llano *et al.*, An essential role for LEDGF/p75 in HIV integration. *Science*. **314**, 461–4
465 (2006).

466 21. R. C. Larue *et al.*, Bimodal high-affinity association of Brd4 with murine leukemia virus
467 integrase and mononucleosomes. *Nucleic Acids Res.* **42**, 4868–4881 (2014).
468 22. A. Sharma *et al.*, BET proteins promote efficient murine leukemia virus integration at
469 transcription start sites. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12036–41 (2013).
470 23. S. S. Gupta *et al.*, Bromo- and extraterminal domain chromatin regulators serve as cofactors
471 for murine leukemia virus integration. *J. Virol.* **87**, 12721–36 (2013).
472 24. S. Aiyer *et al.*, Altering murine leukemia virus integration through disruption of the integrase
473 and BET protein family interaction. *Nucleic Acids Res.* **42**, 5917–5928 (2014).
474 25. M.-C. Shun *et al.*, LEDGF/p75 functions downstream from preintegration complex formation to
475 effect gene-specific HIV-1 integration. *Genes Dev.* **21**, 1767–78 (2007).
476 26. H. M. Marshall *et al.*, Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration
477 targeting. *PLoS One*. **2**, e1340 (2007).
478 27. A. Ciuffi *et al.*, A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11**, 1287–9
479 (2005).
480 28. L. Vandekerckhove *et al.*, Transient and stable knockdown of the integrase cofactor
481 LEDGF/p75 reveals its role in the replication cycle of human immunodeficiency virus. *J. Virol.*
482 **80**, 1886–96 (2006).
483 29. S. P. Goff, Host factors exploited by retroviruses. *Nat. Rev.* **5**, 253–263 (2007).
484 30. A. Engelman, The roles of cellular factors in retroviral integration. *Curr. Top. Microbiol.*
485 *Immunol.* **281**, 209–38 (2003).
486 31. F. D. Bushman *et al.*, Host cell factors in HIV replication: meta-analysis of genome-wide
487 studies. *PLoS Pathog.* **5**, e1000437 (2009).
488 32. W. C. Greene, B. M. Peterlin, Charting HIV's remarkable voyage through the cell: Basic
489 science as a passport to future therapy. *Nat. Med.* **8**, 673–80 (2002).
490 33. Y. Suzuki, R. Craigie, The road to chromatin - nuclear entry of retroviruses. *Nat. Rev.*
491 *Microbiol.* **5**, 187–96 (2007).
492 34. V. Achuthan *et al.*, Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of
493 Viral DNA Integration. *Cell Host Microbe* (2018), doi:10.1016/j.chom.2018.08.002.
494 35. C. R. Chin *et al.*, Direct Visualization of HIV-1 Replication Intermediates Shows that Capsid
495 and CPSF6 Modulate HIV-1 Intra-nuclear Invasion and Integration. *Cell Rep.* **13**, 1717–1731
496 (2015).
497 36. K. Peng *et al.*, Quantitative microscopy of functional HIV post-entry complexes reveals
498 association of replication with the viral capsid. *Elife*. **3** (2014), doi:10.7554/eLife.04114.
499 37. A. Dharan *et al.*, KIF5B and Nup358 Cooperatively Mediate the Nuclear Import of HIV-1 during
500 Infection. *PLOS Pathog.* **12**, e1005700 (2016).
501 38. C. Buffone *et al.*, Nup153 Unlocks the Nuclear Pore Complex for HIV-1 Nuclear Translocation
502 in Nondividing Cells. *J. Virol.* **92** (2018), doi:10.1128/JVI.00648-18.
503 39. H. O. Ajoge *et al.*, Antiretroviral APOBEC3 Cytidine Deaminases Alter HIV-1 Provirus
504 Integration Site Profiles. *bioRxiv*, 833475 (2019).
505 40. F. D. Bushman, Tethering human immunodeficiency virus 1 integrase to a DNA site directs

506 integration to nearby sequences. *Proc.Natl.Acad.Sci.USA*. **91**, 9233–9237 (1994).
507 41. P. M. Pryciak, H. E. Varmus, Nucleosomes, DNA-binding proteins, and DNA sequence
508 modulate retroviral integration target site selection. *Cell*. **69**, 769–80 (1992).
509 42. X. Wu, Y. Li, B. Crise, S. M. Burgess, D. J. Munroe, Weak palindromic consensus sequences
510 are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79**,
511 5211–4 (2005).
512 43. R. G. McAllister *et al.*, Lentivector integration sites in ependymal cells from a model of
513 metachromatic leukodystrophy: non-B DNA as a new factor influencing integration. *Mol. Ther.*
514 *Nucleic Acids*. **3**, e187 (2014).
515 44. Jungkweon Choi and Tetsuro Majima, Conformational changes of non-BDNA. *Chem. Soc.*
516 *Rev.* **40**, 5893–5909 (2011).
517 45. A. Bacolla, R. D. Wells, Non-B DNA conformations, genomic rearrangements, and human
518 disease. *J. Biol. Chem.* **279**, 47411–4 (2004).
519 46. C. M. Venner *et al.*, Infecting HIV-1 Subtype Predicts Disease Progression in Women of Sub-
520 Saharan Africa. *EBioMedicine*. **13**, 305–314 (2016).
521 47. C. S. Morrison *et al.*, Hormonal Contraceptive Use and HIV Disease Progression Among
522 Women in Uganda and Zimbabwe. *JAIDS J. Acquir. Immune Defic. Syndr.* **57**, 157–164
523 (2011).
524 48. C. S. Morrison *et al.*, Hormonal contraception and the risk of HIV acquisition. *AIDS*. **21**, 85–95
525 (2007).
526 49. T. L. Lemonovich *et al.*, Differences in Clinical Manifestations of Acute and Early HIV-1
527 Infection between HIV-1 Subtypes in African Women. *J. Int. Assoc. Provid. AIDS Care*. **14**,
528 415–422 (2015).
529 50. A. Ciuffi, S. D. Barr, Identification of HIV integration sites in infected host genomic DNA.
530 *Methods*. **53**, 39–46 (2011).
531 51. R. Z. Cer *et al.*, Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its
532 associated tools. *Nucleic Acids Res.* **41**, D94–D100 (2013).
533 52. L. Guelen *et al.*, Domain organization of human chromosomes revealed by mapping of nuclear
534 lamina interactions. *Nature*. **453**, 948–51 (2008).
535 53. S. D. Barr *et al.*, HIV integration site selection: targeting in macrophages and the effects of
536 different routes of viral entry. *Mol. Ther.* **14**, 218–225 (2006).
537 54. D. Derse *et al.*, Human T-cell leukemia virus type 1 integration target sites in the human
538 genome: comparison with those of other retroviruses. *J. Virol.* **81**, 6731–41 (2007).
539 55. A. Faschinger *et al.*, Mouse Mammary Tumor Virus Integration Site Selection in Human and
540 Mouse Genomes. *J. Virol.* **82**, 1360–1367 (2008).
541 56. B. Marini *et al.*, Nuclear architecture dictates HIV-1 integration site selection. *Nature*. **521**,
542 227–31 (2015).
543 57. E. Battivelli *et al.*, Distinct chromatin functional states correlate with HIV latency reactivation in
544 infected primary CD4+ T cells. *Elife*. **7** (2018), doi:10.7554/eLife.34655.
545 58. V. Poletti, F. Mavilio, *Interactions between Retroviruses and the Host Cell Genome* (2018;

546 http://www.ncbi.nlm.nih.gov/pubmed/29159201), vol. 8.
547 59. F. Maldarelli *et al.*, Specific HIV integration sites are linked to clonal expansion and
548 persistence of infected cells. *Science*. **345**, 179–83 (2014).
549 60. L. B. Cohn *et al.*, HIV-1 Integration Landscape during Latent and Active Infection. *Cell*. **160**,
550 420–432 (2015).
551 61. C. Cattoglio *et al.*, Hot spots of retroviral integration in human CD34+ hematopoietic cells.
552 *Blood*. **110**, 1770–8 (2007).
553 62. M. Nettling *et al.*, DiffLogo: a comparative visualization of sequence motifs. *BMC*
554 *Bioinformatics*. **16**, 387 (2015).
555 63. W. P. Wahls, L. J. Wallace, P. D. Moore, The Z-DNA motif d(TG)30 promotes reception of
556 information during gene conversion events while stimulating homologous recombination in
557 human cells in culture. *Mol. Cell. Biol.* **10**, 785–93 (1990).
558 64. J. Y. Chin, E. B. Schleifman, P. M. Glazer, Repair and recombination induced by triple helix
559 DNA. *Front. Biosci.* **12**, 4288 (2007).
560 65. V. Brázda, R. C. Laister, E. B. Jagelská, C. Arrowsmith, Cruciform structures are a common
561 DNA feature important for regulating biological processes. *BMC Mol. Biol.* **12**, 33 (2011).
562 66. C. M. Holtz, H. A. Sadler, L. M. Mansky, APOBEC3G cytosine deamination hotspots are
563 defined by both sequence context and single-stranded DNA secondary structure. *Nucleic*
564 *Acids Res.* **41**, 6139–48 (2013).
565 67. S. J. Ziegler *et al.*, Insights into DNA substrate selection by APOBEC3G from structural,
566 biochemical, and functional studies. *PLoS One*. **13**, e0195048 (2018).
567 68. A. N. Mazumder *et al.*, Inhibition of human immunodeficiency virus type 1 integrase by
568 guanosine quartet structures. *Biochemistry*. **35**, 13762–13771 (1996).
569 69. J.-C. Jiang, K. R. Upton, Human transposons are an abundant supply of transcription factor
570 binding sites and promoter activities in breast cancer cell lines. *Mob. DNA*. **10**, 16 (2019).
571 70. Y. Kondo, J.-P. J. Issa, Enrichment for histone H3 lysine 9 methylation at Alu repeats in human
572 cells. *J. Biol. Chem.* **278**, 27658–62 (2003).
573 71. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss, L. H. Hurley, Direct evidence for a G-quadruplex in
574 a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc.*
575 *Natl. Acad. Sci. U. S. A.* **99**, 11593–8 (2002).
576 72. A. Verma, V. K. Yadav, R. Basundra, A. Kumar, S. Chowdhury, Evidence of genome-wide G4
577 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.* **37**, 4194–204
578 (2009).
579 73. S. Waga, S. Mizuno, M. Yoshida, Chromosomal protein HMG1 removes the transcriptional
580 block caused by the cruciform in supercoiled DNA. *J. Biol. Chem.* **265**, 19424–8 (1990).
581 74. S. Waga, S. Mizuno, M. Yoshida, Nonhistone protein HMG1 removes the transcriptional block
582 caused by left-handed Z-form segment in a supercoiled DNA. *Biochem. Biophys. Res.*
583 *Commun.* **153**, 334–9 (1988).
584 75. A. Jain, M. Magistri, S. Napoli, G. M. Carbone, C. V. Catapano, Mechanisms of triplex DNA-
585 mediated inhibition of transcription initiation in cells. *Biochimie*. **92**, 317–320 (2010).

586 76. L. J. Maher, P. B. Dervan, B. Wold, Analysis of promoter-specific repression by triple-helical
587 DNA complexes in a eukaryotic cell-free transcription system. *Biochemistry*. **31**, 70–81 (1992).
588 77. M. L. Bochman, K. Paeschke, V. A. Zakian, DNA secondary structures: stability and function of
589 G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–80 (2012).
590 78. J. Delic, R. Onclercq, M. Moisan-Coppey, Inhibition and enhancement of eukaryotic gene
591 expression by potential non-B DNA sequences. *Biochem. Biophys. Res. Commun.* **180**, 1273–
592 83 (1991).
593 79. S. Tornaletti, S. Park-Snyder, P. C. Hanawalt, G4-forming sequences in the non-transcribed
594 DNA strand pose blocks to T7 RNA polymerase and mammalian RNA polymerase II. *J. Biol.*
595 *Chem.* **283**, 12756–62 (2008).
596 80. B. P. Belotserkovskii *et al.*, A triplex-forming sequence from the human c-MYC promoter
597 interferes with DNA transcription. *J. Biol. Chem.* **282**, 32433–41 (2007).
598 81. R. Löwer, J. Löwer, R. Kurth, The viruses in all of us: characteristics and biological significance
599 of human endogenous retrovirus sequences. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 5177–84
600 (1996).
601 82. G.-Z. Han, M. Worobey, An endogenous foamy-like viral element in the coelacanth genome.
602 *PLoS Pathog.* **8**, e1002790 (2012).
603 83. W. M. Switzer *et al.*, Ancient co-speciation of simian foamy viruses and primates. *Nature*. **434**,
604 376–380 (2005).
605 84. S. R. Ross, Mouse mammary tumor virus molecular biology and oncogenesis. *Viruses*. **2**,
606 2000–12 (2010).
607 85. H.-C. Chen, J. P. Martinez, E. Zorita, A. Meyerhans, G. J. Filion, Position effects influence HIV
608 latency reversal. *Nat. Struct. Mol. Biol.* **24**, 47–54 (2017).
609 86. L. S. Vranckx *et al.*, LEDGIN-mediated Inhibition of Integrase–LEDGF/p75 Interaction
610 Reduces Reactivation of Residual Latent HIV. *EBioMedicine* (2016),
611 doi:10.1016/j.ebiom.2016.04.039.
612 87. J. Demeulemeester *et al.*, HIV-1 integrase variants retarget viral integration and are associated
613 with disease progression in a chronic infection cohort. *Cell Host Microbe*. **16**, 651–662 (2014).
614 88. R. E. Myers, D. Pillay, Analysis of Natural Sequence Variation and Covariation in Human
615 Immunodeficiency Virus Type 1 Integrase. *J. Virol.* **82**, 9228–9235 (2008).
616 89. S. Y. Rhee *et al.*, Natural variation of HIV-1 group M integrase: Implications for a new class of
617 antiretroviral inhibitors. *Retrovirology*. **5**, 1–11 (2008).
618

619 **FIGURE LEGENDS**

620 **Figure 1: Evolutionarily diverse retroviruses exhibit distinct integration site preferences. A,**
621 Heatmaps depicting the fold enrichment or depletion of integration sites near common genomic
622 features compared to matched random controls. Darker shades represent higher fold-changes in the
623 ratio of integration sites to matched random control sites. Bins represent the distance of the
624 integration sites from each genomic feature. Bin 0 = within the feature; Bin 1= 1-499 bp; Bin 2 = 500-
625 4,999 bp; Bin 3 = 5,000-49,999 bp; Bin 4 = >49,999 bp. Heatmaps of the diverse retrovirus genera

were superimposed on a BioNJ tree constructed using their reverse transcriptase amino acid sequences using the Dayhoff substitution model with 1000 bootstraps. All branches are scaled according to number of amino acids changes per site. The phylogenetic tree shows the evolutionary relatedness of the different retrovirus genera only. Significant differences are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (Fisher's Exact test). **B**, Proportion of retroviral integration sites located within genes compared to the random control (blue line). **C**, Pairwise analysis was performed on the retroviral integration site profile preferences (fold enrichment and depletion values) using Euclidean distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site profiles are indicated by darker blue color in the pairwise distance matrix. **D**, Nuclear localization of integration sites were determined by quantifying the proportion of total integrations that fell within a lamin-associated domain (LAD) (=1) as opposed to outside a LAD (=0). *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001; Fisher's Exact test. Infinite number (inf), 1 or more integrations were observed when 0 integrations were expected by chance. Not a number (nan), 0 integrations were observed and 0 were expected by chance. HIV-1 = human immunodeficiency virus, SIV_{mne} = simian immunodeficiency virus (isolated from a pig-tailed macaque), FIV= feline immunodeficiency virus, HTLV-1= human T-lymphotrophic virus Type 1, MLV = murine leukemia virus, FV = foamy virus, ASLV = avian sarcoma leucosis virus, MMTV = mouse mammary tumor virus and ERVs = endogenous retroviruses.

Figure 2: Evolutionarily diverse retroviruses target non-B DNA for integration. A, Heatmaps illustrating the fold-enrichment or depletion of unique retroviral integration sites near non-B DNA features compared to matched random controls. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (Fisher's Exact test). **B**, Pairwise analysis was performed on the retroviral integration site profile preferences using Euclidean distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site profiles are indicated by darker blue color in the pairwise distance matrix.

Figure 3: Integration site profiles differ between *in vitro*- and *in vivo*-derived datasets. A, Heatmaps illustrating the fold-enrichment or depletion of unique integration sites (compared to the matched random control) near common genomic features from *in vivo*-derived datasets (n= 167,098 sites) and *in* vitro-derived datasets (n= 13,601 sites). Inset numbers represent the fold-change in the percentage of integration sites. **B,** Comparison of the percentage of integration sites within 5,000 bp of common genomic features between *in vitro*- and *in vivo*-derived datasets. **C,** Heatmaps illustrating the fold-enrichment of unique integration sites compared to the matched random control near non-B DNA from *in vivo*- and *in vitro*-derived datasets. **D,** Comparisons of the percentage of integration sites within 500 bp of G4, triplex and Z-DNA between *in vitro*- and *in vivo*-derived datasets. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (Fisher's Exact test).

666 **Figure 4: HIV-1 subtype A, B, C and D have different integration site preferences for genomic**
667 **features. A,** Comparison of the percentage of integration sites near common genomic features
668 between HIV-1 subtypes A, B, C and D. **B,** Heatmaps depicting the fold enrichment or depletion of
669 integration sites near common genomic features compared to the matched random control. Darker
670 shades represent higher fold-changes in the ratio of integration sites to matched random control sites.
671 Bins in A and B represent the distance of the integration sites from the genomic feature. **C**,
672 Comparison of the percentage of integration sites within genes from HIV-1 subtypes A, B, C and D. **D,**
673 Pairwise analysis was performed on the retroviral integration site profile preferences using Euclidean
674 distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships
675 between retroviral integration site profiles are indicated by darker blue color in the pairwise distance
676 matrix. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P <
677 0.0001 (Fisher's Exact test). **E**, Estimates of Evolutionary Divergence over Sequence Pairs between
678 Groups. The number of amino acid substitutions per site from averaging over all sequence pairs
679 between groups are shown. Analyses were conducted using the Poisson correction model. This
680 analysis involved 486 amino acid sequences. The coding data was translated assuming a Standard
681 genetic code table. All ambiguous positions were removed for each sequence pair (pairwise deletion
682 option). There were a total of 265 positions in the final dataset. Evolutionary analyses were conducted
683 in MEGA X. **F**, Amino acid alignment of the C-terminal domain of HIV-1 integrase from subtypes A, B,
684 C and D.
685
686 **Figure 5: HIV-1 subtype A, B, C and D have different integration site preferences for non-B**
687 **DNA. A,** Comparison of the percentage of integration sites near non-B DNA features between HIV-1
688 subtypes A, B, C and D. **B,** Heatmaps depicting the fold enrichment or depletion of integration sites
689 near non-B DNA compared to the matched random control. Darker shades represent higher fold-
690 changes in the ratio of integration sites to matched random control sites. Bins in A and B represent
691 the distance of the integration sites from the non-B DNA feature. Comparison of the percentage of
692 integration sites within genes from HIV-1 subtypes A, B, C and D. **C,** Pairwise analysis was performed
693 on the retroviral integration site profile preferences using Euclidean distance as the measurement
694 method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site
695 profiles are indicated by darker blue color in the pairwise distance matrix. Significant differences are
696 denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001 (Fisher's Exact test).
697
698
699 **Figure 6: Integration hotspots identified from individuals infected with different HIV-1**
700 **subtypes. A,** Venn diagram showing the percentage of unique (colored red or light blue) and shared
701 (colored dark blue) integration site hotspots among the different HIV-1 subtypes. **B** and **C,** circos plots
702 showing integration site hotspots inside (B) or outside (C) genes. The subtypes whose hotspots were
703 shared is denoted in capital letters on the outside of the circos plots. For integration site hotspots
704 outside genes (C), the closest gene and the distance from that gene are shown on the outside of the
705 circos plots separated by a colon. Hotspots were defined as a 1 kb window in the genome hosting 2

706 or more integration sites. Note: (+) Due to the large number of subtype B integration site hotspots
707 (10,762), only those hotspots shared between subtypes A, C and/or D are shown on the circos plots.
708 No integration site hotspots outside of genes were shared between subtype B and any of the other
709 subtypes. **D**, Consensus sequences were generated from a window of 100 nucleotides upstream and
710 100 nucleotides downstream of each integration site. Consensus sequences from integration sites
711 located in hotspots were compared to consensus sequences from sites not located in hotspots using
712 DiffLogo. Consensus sequences were analyzed for the presence of non-B DNA motifs (solid colored
713 lines).

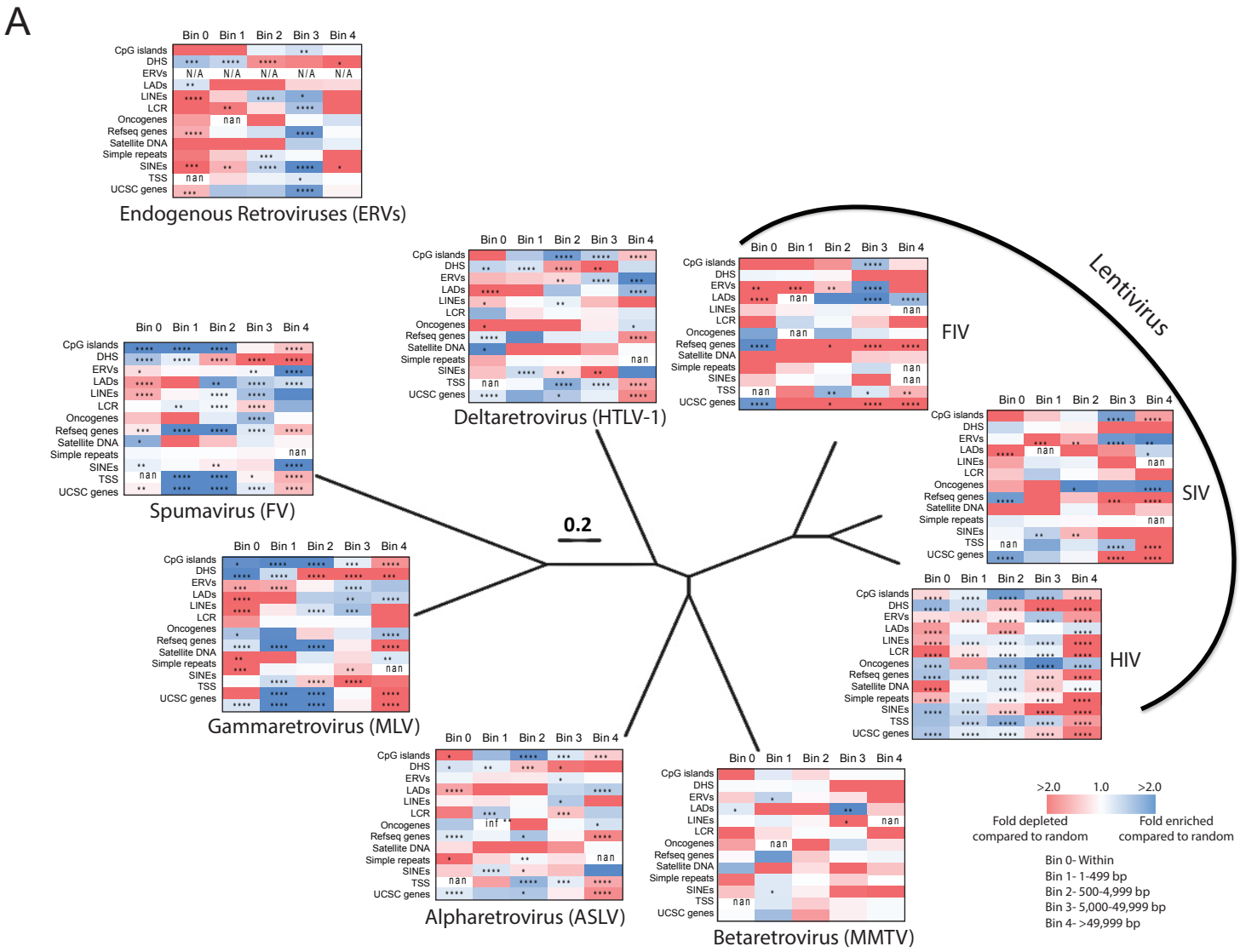
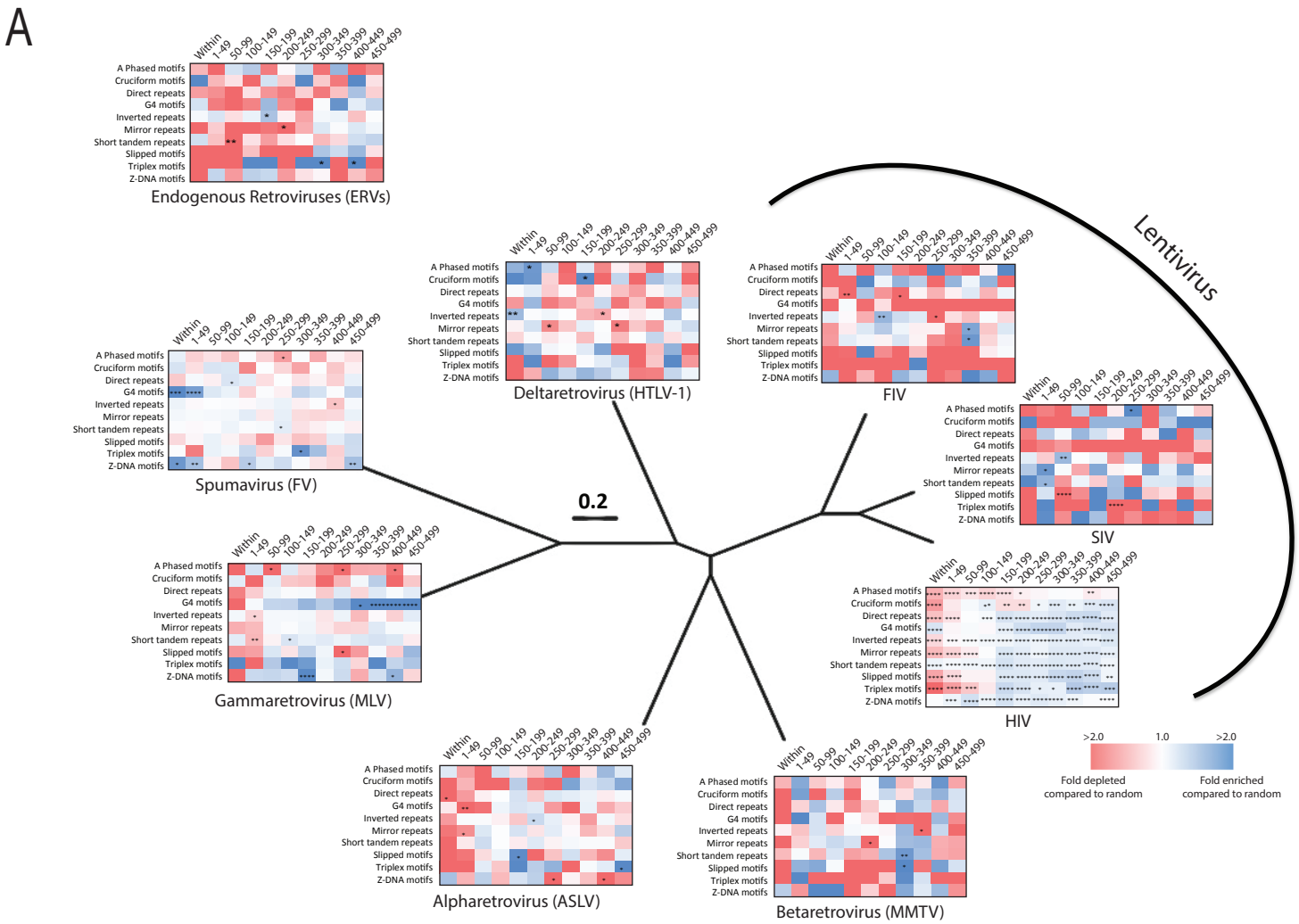

A
Endogenous Retroviruses (ERVs)
Deltaretrovirus (HTLV-1)
FIV
Lentivirus
Spumavirus (FV)
0.2
SIV
Gammaretrovirus (MLV)
HIV
Alpharetrovirus (ASLV)
Betaretrovirus (MMTV)
Bin 0 Bin 1 Bin 2 Bin 3 Bin 4
CpG islands
DHS
ERVs
LADs
LINEs
LCR
Oncogenes
Refseq genes
Satellite DNA
Simple repeats
SINEs
TSS
UCSC genes
>2.0 1.0 >2.0
Fold depleted compared to random
Fold enriched compared to random
Bin 0- Within
Bin 1- 1-499 bp
Bin 2- 500-4,999 bp
Bin 3- 5,000-49,999 bp
Bin 4- >49,999 bp
B
random control
Proportion of integration sites within genes
Retrovirus
ASLV ERVs FIV FV HIV HTLV MLV MMTV SIV
C
Values
D
Control
Test
Proportion of integration sites within a LAD
Retroviral integration site dataset

Figure 1

Figure 1: Evolutionarily diverse retroviruses exhibit distinct integration site preferences. A, Heatmaps depicting the fold enrichment or depletion of integration sites near common genomic features compared to matched random controls. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Bins represent the distance of the integration sites from each genomic feature. Bin 0 = within the feature; Bin 1= 1-499 bp; Bin 2 = 500-4,999 bp; Bin 3 = 5,000-49,999 bp; Bin 4 = >49,999 bp. Heatmaps of the diverse retrovirus genera were superimposed on a BioNJ tree constructed using their reverse transcriptase amino acid sequences using the Dayhoff substitution model with 1000 bootstraps. All branches are scaled according to number of amino acids changes per site. The phylogenetic tree shows the evolutionary relatedness of the different retrovirus genera only. Significant differences are denoted by asterisks (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (Fisher's Exact test). **B**, Proportion of retroviral integration sites located within genes compared to the random control (blue line). **C**, Pairwise analysis was performed on the retroviral integration site profile preferences (fold enrichment and depletion values) using Euclidean distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site profiles are indicated by darker blue color in the pairwise distance matrix. **D**, Nuclear localization of integration sites were determined by quantifying the proportion of total integrations that fell within a lamin-associated domain (LAD) (=1) as opposed to outside a LAD (=0). Infinite number (inf), 1 or more integrations were observed when 0 integrations were expected by chance. Not a number (nan), 0 integrations were observed and 0 were expected by chance. HIV-1 = human immunodeficiency virus, SIV = simian immunodeficiency virus, FIV= feline immunodeficiency virus, HTLV-1= human T-lymphotrophic virus Type 1, MLV = murine leukemia virus, FV = foamy virus, ASLV = avian sarcoma leucosis virus, MMTV = mouse mammary tumor virus and ERVs = endogenous retroviruses.



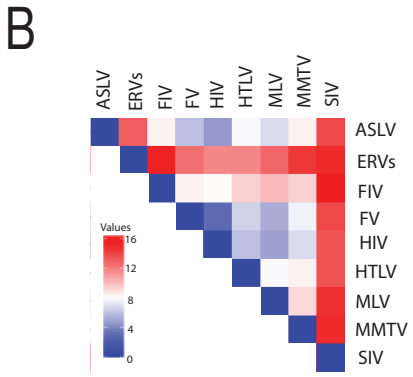


Figure 2: Evolutionarily diverse retroviruses target non-B DNA for integration. A, Heatmaps illustrating the fold-enrichment or depletion of unique retroviral integration sites near non-B DNA features compared to matched random controls. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (Fisher's Exact test). **B**, Pairwise analysis was performed on the retroviral integration site profile preferences using Euclidean distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site profiles are indicated by darker blue color in the pairwise distance matrix.

A

In vivo dataset compared to random

	Within	1-499	500-4,999	5,000-49,999	>50,000
CpG islands	0.90**	1.19****	1.82****	1.50****	0.82****
DNaseI HS	1.65****	1.27****	0.75****	0.23****	0.01****
ERVs	0.94****	0.90****	0.96****	1.44****	0.61****
LADs	0.71****	0.89	0.68****	0.98	1.20****
LINEs	0.71****	1.18****	1.17****	1.26****	0.01****
LCR	0.40****	0.91****	1.08****	1.14****	0.11****
Oncogenes	1.45****	0.73	1.56****	2.11****	1.50****
Refseq genes	1.47****	1.23****	1.20****	0.90****	0.58****
Satellite DNA	0.23****	1.09	1.23****	0.82****	1.06****
Simple repeats	0.78****	1.15****	1.09****	0.94****	0.00****
SINEs	1.62****	1.25****	0.85****	0.42****	0.04****
TSS	1.22	1.36****	1.65****	1.36****	0.70****
UCSC genes	1.45****	1.11**	1.16****	0.89****	0.60****

In vitro dataset compared to random

	Within	1-499	500-4,999	5,000-49,999	>50,000
CpG islands	0.10****	0.55****	2.22****	1.79****	0.59****
DNaseI HS	1.40****	1.19****	0.78****	0.16****	0.01****
ERVs	0.47****	0.52****	0.77****	1.91****	3.19****
LADs	0.31****	3.58****	0.89	1.38****	1.39****
LINEs	0.81****	1.02	1.04****	1.62****	0.00
LCR	1.02	1.16****	1.02	0.92****	0.13****
Oncogenes	1.48****	0.30	3.00****	1.93****	1.80****
Refseq genes	1.82****	1.70****	0.94	0.38****	0.17****
Satellite DNA	0.07****	0.19**	0.75	1.01	1.19****
Simple repeats	0.51****	1.04*	1.02***	0.90****	nan
SINEs	0.93**	1.44****	0.76****	0.20****	0.16***
TSS	4.95	1.74****	2.12****	1.51****	0.36****
UCSC genes	1.81****	1.65****	0.87**	0.39****	0.17****

In vivo compared to *in vitro* dataset

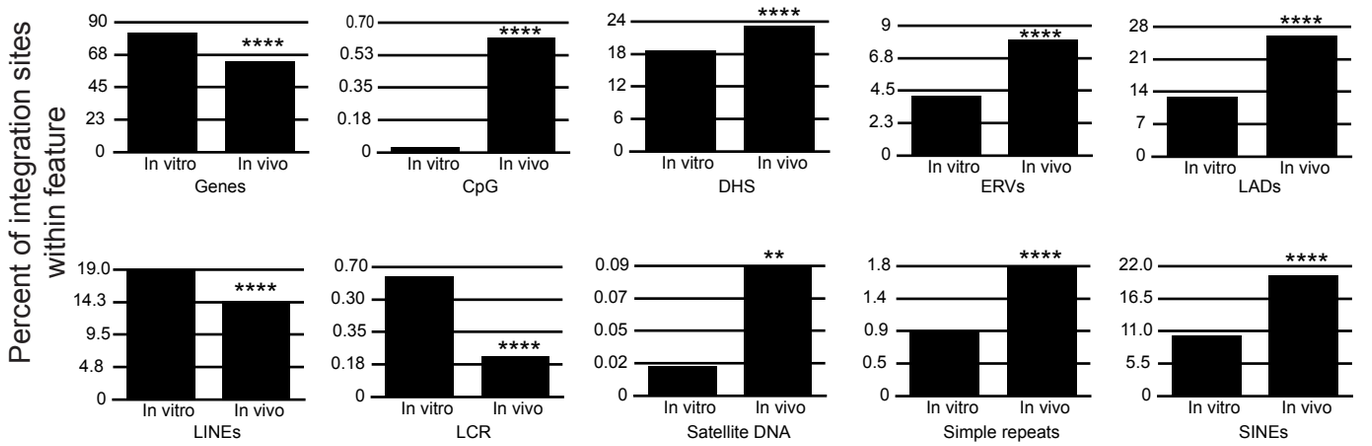
	Within	1-499	500-4,999	5,000-49,999	>50,000
CpG islands	21.06****	2.75****	0.89****	0.83****	1.25****
DNaseI HS	1.25****	1.01	0.83****	1.40**	2.64
ERVs	1.92****	1.60****	1.13****	0.72****	0.81*
LADs	2.01****	0.32***	0.73*	0.65****	0.81****
LINEs	0.76****	1.08****	1.06****	0.93	inf
LCR	0.34****	0.66****	0.97***	1.21****	3.72**
Oncogenes	0.92	2.01	0.56****	1.12*	0.89****
Refseq genes	0.75****	0.77**	1.30****	2.30****	3.04****
Satellite DNA	4.16**	6.75***	1.74***	0.85**	0.91****
Simple repeats	2.02****	1.07****	0.97****	0.98	nan
SINEs	2.01****	0.79****	0.99	2.21****	2.55
TSS	0.23	1.02	0.80****	0.86****	1.75****
UCSC genes	0.76****	0.71***	1.33****	2.21****	3.23****

>2.0 1.0 >2.0

Fold depleted compared to random — Fold enriched compared to random

Fold depleted compared to in vitro — Fold enriched compared to in vitro

B



C

In vivo dataset compared to random

	Within	1-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499
A Phased motifs	0.72****	0.85****	0.92***	0.86****	0.90****	0.94*	0.96	1.03	1.00	0.93**	0.98
Cruciform motifs	0.65****	0.92	1.04	1.17***	0.87**	0.89*	1.10*	1.16***	1.13**	1.15**	1.22****
Direct repeats	0.81****	0.90****	0.99	1.05***	1.26****	1.19****	1.17****	1.22****	1.32****	1.32****	1.18****
G4 motifs	1.25****	1.00	1.06*	1.01	1.19****	1.32****	1.40****	1.43****	1.22****	1.29****	1.30****
Inverted repeats	0.86****	0.97****	1.03***	1.08****	1.09****	1.14****	1.12****	1.16****	1.19****	1.14****	1.24****
Mirror repeats	0.78****	0.84****	0.92****	1.01	1.22****	1.16****	1.10****	1.17****	1.18****	1.17****	1.14****
Short tandem repeats	1.13****	1.06****	1.08****	1.09****	1.27****	1.18****	1.16****	1.21****	1.18****	1.18****	1.09****
Slipped motifs	0.73****	0.87****	0.95	0.97	1.22****	1.22****	1.19****	1.32****	1.37****	1.28****	1.10**
Triplex motifs	0.46****	0.67****	0.84**	0.97	1.37****	1.22****	1.21***	1.18**	1.49****	1.45****	1.51****
Z-DNA motifs	1.07	1.11****	1.28****	1.12****	1.29****	1.16****	1.24****	1.17****	1.11****	1.04	1.04

In vitro dataset compared to random

	Within	1-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499
A Phased motifs	1.01	1.00	0.93	0.88	0.96	0.95	1.10	0.91	0.89	0.98	0.88
Cruciform motifs	0.92	0.96	1.15	0.86	1.12	0.83	0.97	1.19	0.95	1.41**	1.15
Direct repeats	0.62****	1.03	1.06	0.99	0.99	1.10	0.99	1.20***	1.16**	1.19***	1.06
G4 motifs	0.68	0.56****	0.76*	0.67***	0.51****	0.81	0.71**	0.57****	0.58****	0.53****	0.52****
Inverted repeats	1.07	1.06**	1.05*	0.98	1.01	1.04	0.91*	0.99	0.89*	0.99	0.88*
Mirror repeats	0.75****	1.05	1.13**	0.98	1.09*	1.19****	1.08	1.19***	1.19***	1.27****	0.98
Short tandem repeats	0.83*	1.11***	1.22****	1.13***	0.99	1.11**	1.16****	1.11**	1.20****	1.23****	1.07
Slipped motifs	0.40***	1.04	1.08	0.90	0.98	0.92	0.94	1.35***	1.18	1.47****	0.93
Triplex motifs	0.09***	0.72	0.59*	0.45**	0.56*	1.25	0.32****	0.64*	1.06	0.95	0.50**
Z-DNA motifs	0.39***	0.95	1.06	1.05	0.87	0.85	0.69***	0.92	0.87	0.79*	0.83

>2.0 1.0 >2.0

Fold depleted compared to random — Fold enriched compared to random

In vivo compared to *in vitro* dataset

	Within	1-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499
A Phased motifs	0.63**	0.73****	0.84*	0.81**	0.80**	0.87	0.77***	1.00	0.99	0.85	0.93
Cruciform motifs	0.67	0.88	0.81	1.13	0.72*	0.94	1.02	0.81	1.02	0.75*	0.91
Direct repeats	1.54****	0.90*	0.93	1.06	1.23****	1.03	1.11*	0.94	1.06	0.98	0.99
G4 motifs	2.72****	2.52****	1.79****	1.85****	2.65****	1.86****	2.23****	2.74****	2.37****	2.45****	2.61****
Inverted repeats	0.72****	0.81****	0.88****	0.99	1.01	1.03	1.15****	1.14***	1.27****	1.13*	1.31****
Mirror repeats	1.01	0.75****	0.79****	0.99	1.03	0.92	0.95	0.91*	0.89*	0.82****	1.07
Short tandem repeats	1.47****	0.94*	0.85****	0.94	1.23****	0.99	0.91**	0.98	0.90**	0.86***	0.93
Slipped motifs	2.40***	0.91	0.92	1.13	1.30**	1.31**	1.22*	0.92	1.16	0.80**	1.17
Triplex motifs	7.48*	1.16	1.45	2.33***	2.64****	1.03	3.86****	1.71*	1.37	1.42	3.00****
Z-DNA motifs	2.83****	1.18	1.18*	1.08	1.43****	1.21*	1.63****	1.17	1.22*	1.17	1.19

>2.0 1.0 >2.0

Fold depleted compared to in vitro — Fold enriched compared to in vitro

D

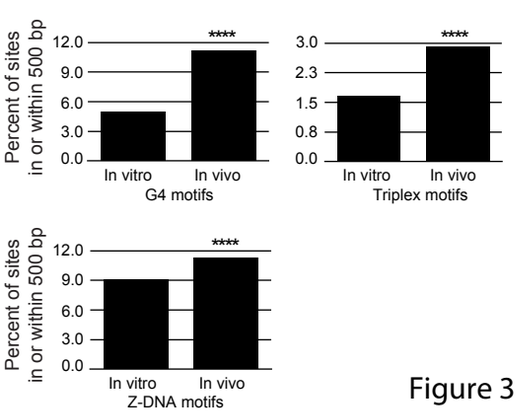


Figure 3

Figure 3: Integration site profiles differ between *in vitro*- and *in vivo*-derived datasets. A, Heatmaps illustrating the fold-enrichment or depletion of unique integration sites (compared to the matched random control) near common genomic features from *in vivo*-derived datasets (n= 167,098 sites) and *in* vitro-derived datasets (n= 13,601 sites). Inset numbers represent the fold-change in the percentage of integration sites. **B,** Comparison of the percentage of integration sites within 5,000 bp of common genomic features between *in vitro*- and *in vivo*-derived datasets. **C,** Heatmaps illustrating the fold-enrichment of unique integration sites compared to the matched random control near non-B DNA from *in vivo*- and *in vitro*-derived datasets. **D,** Comparisons of the percentage of integration sites within 500 bp of G4, triplex and Z-DNA between *in vitro*- and *in vivo*-derived datasets. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001) (Fisher's Exact test).

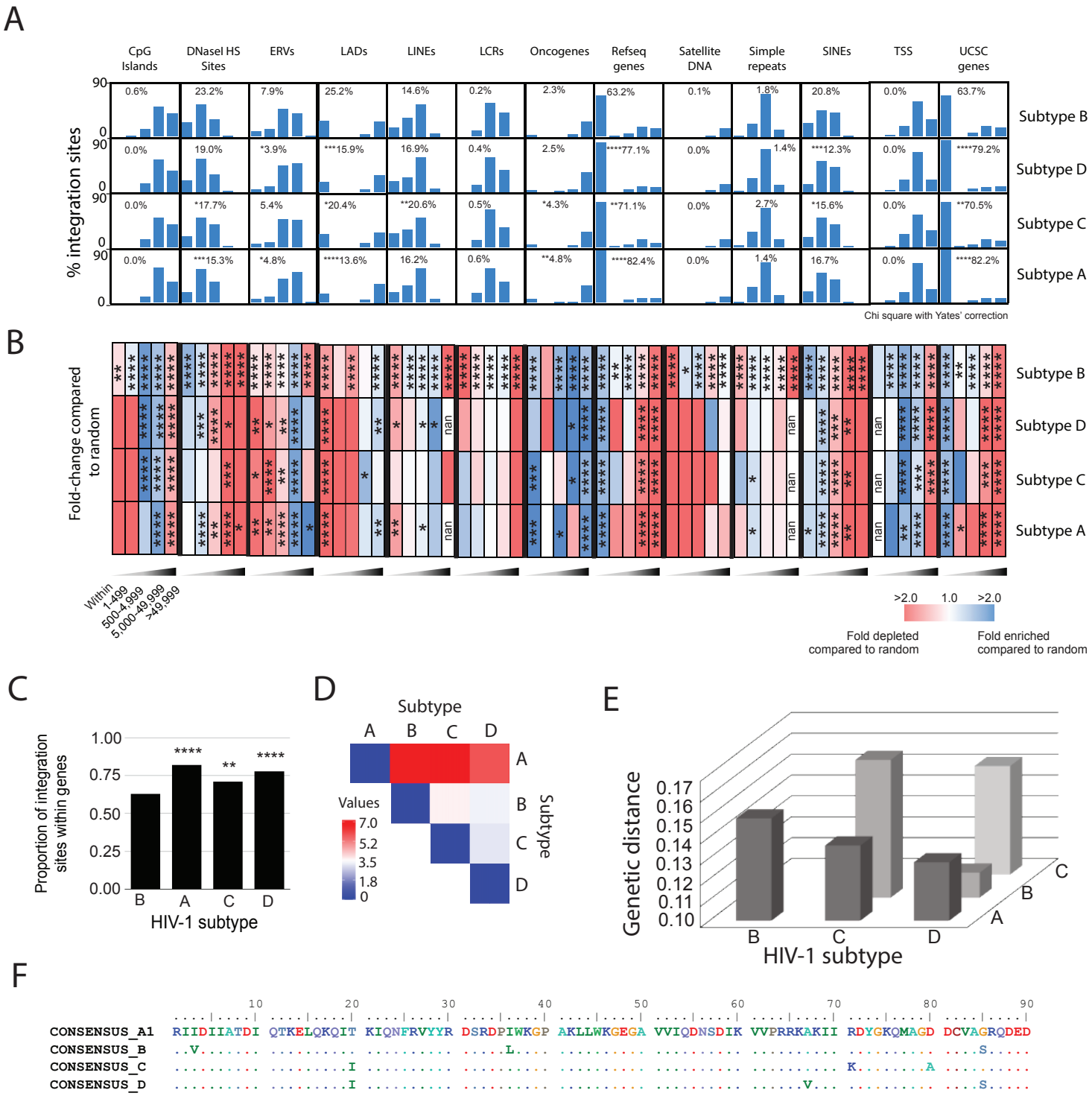


Figure 4: HIV-1 subtype A, B, C and D have different integration site preferences for genomic features. A, Comparison of the percentage of integration sites near common genomic features between HIV-1 subtypes A, B, C and D. **B**, Heatmaps depicting the fold enrichment or depletion of integration sites near common genomic features compared to the matched random control. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Bins in A and B represent the distance of the integration sites from the genomic feature. **C**, Comparison of the percentage of integration sites within genes from HIV-1 subtypes A, B, C and D. **D**, Pairwise analysis was performed on the retroviral integration site profile preferences using Euclidean distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site profiles are indicated by darker blue color in the pairwise distance matrix. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001 (Fisher's Exact test). **E**, Estimates of Evolutionary Divergence over Sequence Pairs between Groups. The number of amino acid substitutions per site from averaging over all sequence pairs between groups are shown. Analyses were conducted using the Poisson correction model. This analysis involved 486 amino acid sequences. The coding data was translated assuming a Standard genetic code table. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 265 positions in the final dataset. Evolutionary analyses were conducted in MEGA X. F, Amino acid alignment of the C-terminal domain of HIV-1 integrase from subtypes A, B, C and D.

Figure 4

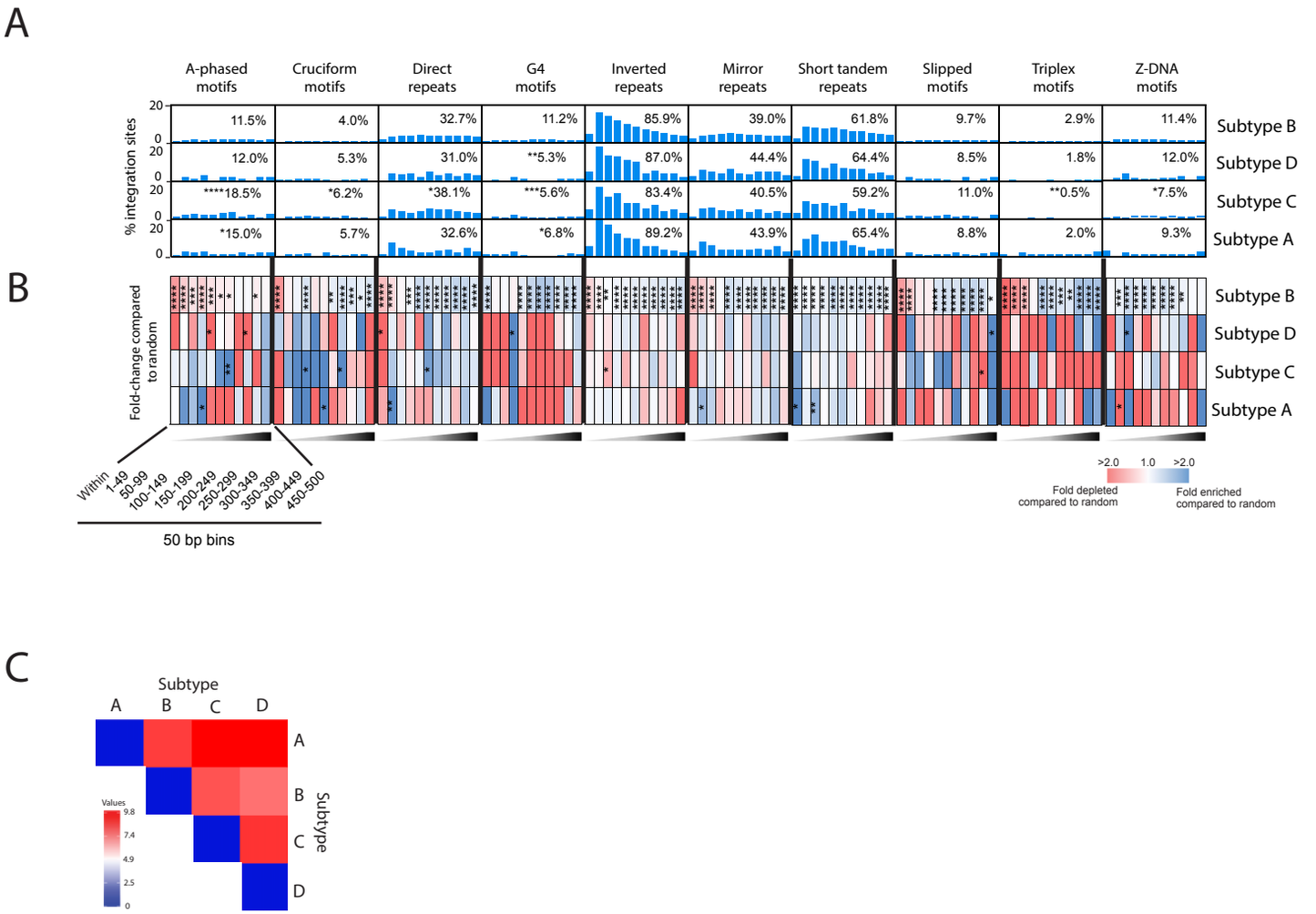


Figure 5: HIV-1 subtype A, B, C and D have different integration site preferences for non-B DNA. A, Comparison of the percentage of integration sites near non-B DNA features between HIV-1 subtypes A, B, C and D. **B**, Heatmaps depicting the fold enrichment or depletion of integration sites near non-B DNA compared to the matched random control. Darker shades represent higher fold-changes in the ratio of integration sites to matched random control sites. Bins in A and B represent the distance of the integration sites from the non-B DNA feature. Comparison of the percentage of integration sites within genes from HIV-1 subtypes A, B, C and D. **C**, Pairwise analysis was performed on the retroviral integration site profile preferences using Euclidean distance as the measurement method (Heatmapper, Babicki et al., 2016). Stronger relationships between retroviral integration site profiles are indicated by darker blue color in the pairwise distance matrix. Significant differences are denoted by asterisks *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001 (Fisher's Exact test).

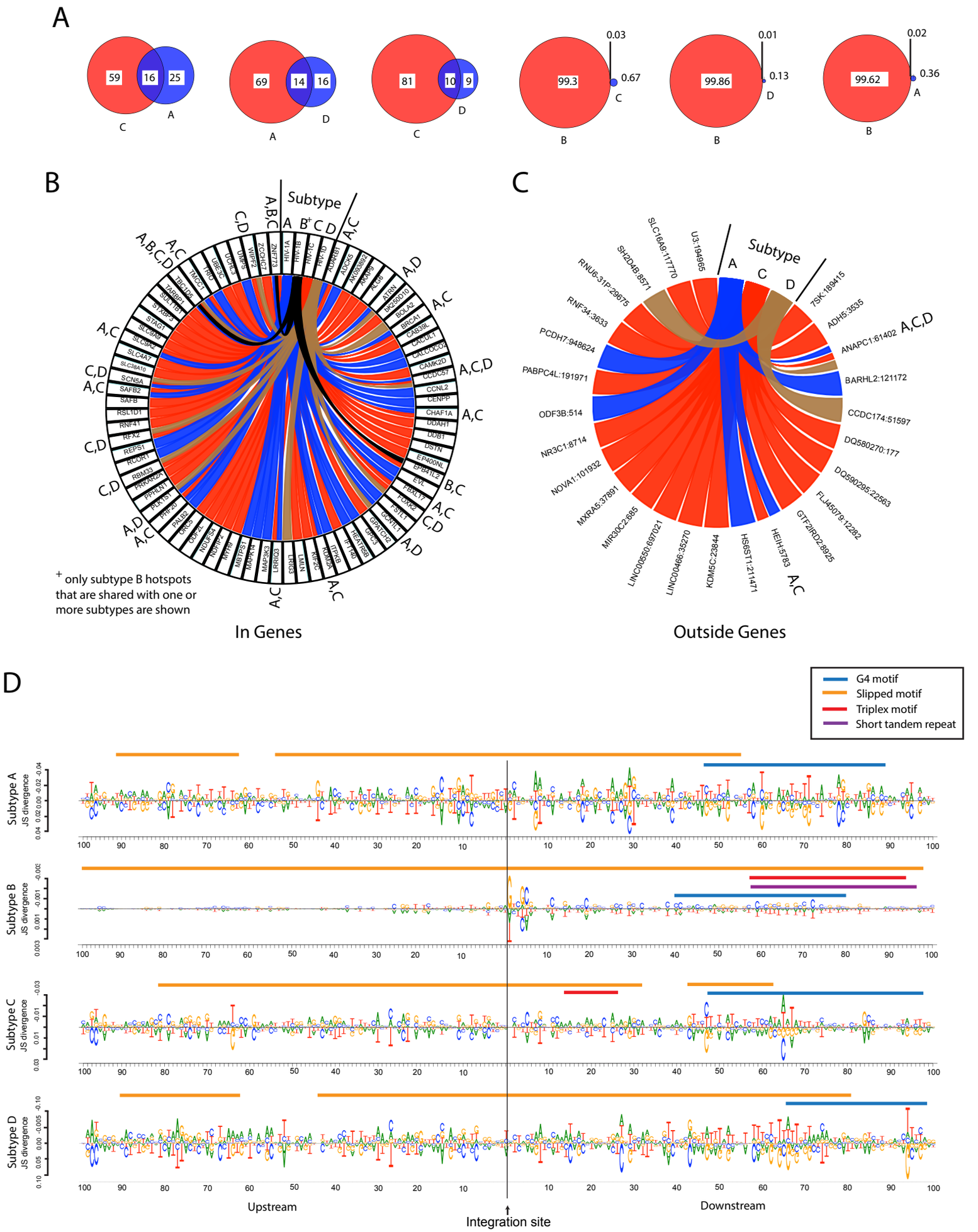

A
59
16
25
C
A
69
14
16
A
D
81
10
9
C
D
99.3
0.03
0.67
C
B
99.86
0.01
0.13
D
B
99.62
0.02
0.36
A
B
B
Subtype
A
B+
C
D
+ only subtype B hotspots that are shared with one or more subtypes are shown
In Genes
C
Subtype
A
C
D
Outside Genes
D
G4 motif
Slipped motif
Triplex motif
Short tandem repeat
Subtype A
Subtype B
Subtype C
Subtype D
JS divergence
Upstream
Integration site
Downstream

Figure 6

Figure 6: Integration hotspots identified from individuals infected with different HIV-1 subtypes. A, Venn diagram showing the percentage of unique (colored red or light blue) and shared (colored dark blue) integration site hotspots among the different HIV-1 subtypes. **B** and **C,** circos plots showing integration site hotspots inside (B) or outside (C) genes. The subtypes whose hotspots were shared is denoted in capital letters on the outside of the circos plots. For integration site hotspots outside genes (C), the closest gene and the distance from that gene are shown on the outside of the circos plots separated by a colon. Hotspots were defined as a 1 kb window in the genome hosting 2 or more integration sites. Note: (+) Due to the large number of subtype B integration site hotspots (10,762), only those hotspots shared between subtypes A, C and/or D are shown on the circos plots. No integration site hotspots outside of genes were shared between subtype B and any of the other subtypes. **D**, Consensus sequences were generated from a window of 100 nucleotides upstream and 100 nucleotides downstream of each integration site. Consensus sequences from integration sites located in hotspots were compared to consensus sequences from sites not located in hotspots using DiffLogo. Consensus sequences were analyzed for the presence of non-B DNA motifs (solid colored lines).