PRISM-Seq: An Ultra-sensitive Sequencing Approach For Mapping Lentiviral Integration Sites

PAL VK[1], CANIS M[1], STONE E[2], BOARD N[2], LENART K[3], FUMAGALLI M[3], KOVACS C[4], JONES RB[2], NUSSENZWEIG M[3,6], HATZIIOANNOU T[1], MUECKSCH F[5], BIENIASZ PD[1,6], LEE GQ[2]

[1]Laboratory of Retrovirology, The Rockefeller University, New York.

[2]Infectious Diseases Division, Department of Medicine, Weill Cornell Medical College, New York.

[3]Laboratory of Molecular Immunology, The Rockefeller University, New York, NY 10065, USA.

[4]Maple Leaf Medical Clinic, Toronto, Ontario, Canada

[5]Department of Infectious Diseases, Virology, Heidelberg University, Heidelberg, Germany.

[6]Howard Hughes Medical Institute, The Rockefeller University, New York.

Current word count (abstract): 144

Current word count (main text): 2988

Abstract

Retroviral integration into host genomes underlies both HIV-1 persistence and the safety and function of lentiviral vectors used in gene and cell therapies. Existing integration site assays remain limited by sensitivity, input requirements, and analytical complexity, and none have been validated at the single-molecule detection limit. Here we introduce PRISM-seq, an ultra-sensitive workflow for genome-wide recovery of lentiviral-host junctions, paired with BulkIntSiteR, an open-source, fully automated pipeline for integration site annotation. We show that PRISM-seq accurately identifies proviral insertions across diverse genomic contexts, including euchromatin, heterochromatin, and repeat-rich centromeric regions, and detects high-confidence integration events down to a single input template molecule. By systematically characterizing assay- and amplification-associated noise, we developed a five-step quality control framework that removes PCR and sequencing artifacts. PRISM-seq also enables quantitative clonal tracking through replicate-based sampling and achieves performance comparable to or exceeding high-input assays at substantially reduced cost.

Introduction

Integration of retroviral DNA into host DNA is central to both HIV-1 pathogenesis and the use of lentiviral vectors in gene therapy. In HIV-1 infection, the site of viral integration predicts viral transcription and reactivation potential[1-3], while in lentiviral vector-based gene therapies, vector insertional profiles inform safety and clonal dynamics[4-10].

Existing proviral integration site sequencing assays and bioinformatics workflows (e.g., INSPIIRED[11,12], VISPA/VISPA2[13], QuickMap[14], VISA[15], MAVRIC[16], GeIST[17], IS-Seq[18], VSeq-Toolkit[19] face three practical barriers. First, many require relatively high template input, typically ~1-10 µg of genomic DNA per library[20,21], are incompatible with single cell sorting, and none were validated down to the single-molecule detection limit nor across distinct genome contexts, including difficult-to-map regions. Second, computational workflows are fragmented, platform-specific, and/or require bioinformatics expertise for read processing and annotation[12,20]. Third, per-sample per-replicate costs remain elevated due to multi-step library construction and sequencing depth requirements.

Here, we present proviral integration site mapping sequencing (PRISM-seq), an ultra-sensitive, single-copy, and bulk-input compatible lentiviral integration site sequencing assay, together with BulkIntSiteR, a fully automated and platform-agnostic bioinformatics analysis pipeline that converts with one-click raw sequencing data files into annotated proviral integration sites.

Results

PRISM-seq design

PRISM-seq detects single-molecule inputs and single-copy species in bulk samples. The core of the PRISM-seq design that enabled ultra-sensitive single-molecule compatible detection is bacteriophage phi29-mediated multiple displacement amplification (MDA), which is not commonly used in integration site identification (Fig. 1a and see methods for the origin and foundation assays for PRISM-seq). In this study, reactions were performed using 50-template inputs, but the assay also accommodates substantially higher input levels (~10,000 templates[22]). MDA-amplified DNA is processed by ligation-mediated PCR (LM-PCR) involving restriction digestion, adaptor ligation, and amplification of both 5′ and 3′ virus-host junctions using virus- and adaptor-specific primer pairs (Fig. 1a, Fig. S1 and S2a). The PRISM-seq primer sets were designed for broad compatibility across lentiviruses, including HIV and lentiviral vector delivery systems (Fig. S2b)[23-27]. PRISM-seq amplicons containing viral-host junctions can be sequenced on any next-generation sequencing (NGS) platform. Reads were processed using BulkIntSiteR, a one-click, open-source computational pipeline developed for automated mapping and annotation of integration sites (Fig. 1b).

Identification of proviral integration sites across the diverse human genome landscape

We first benchmarked PRISM-seq/BulkIntSiteR using the 8E5/LAV cell line, which harbors a single HIV-1 provirus at a known intergenic locus[28]. We identified a single viral integration

92 site on chromosome 13 at position 67,485,907 (5'LTR) and 67,485,903 (3'LTR) in the negative
93 strand, consistent with published data[28]. We did not observe any low-abundance secondary IS
94 (noise).

95

96 The single provirus in 8E5/LAV cells is associated with heterochromatin (supplementary table
97 1). However, viral integration can occur in diverse regions of the human genome, including
98 actively transcribing genes[29], and in repeat-rich, telomeric, and centromeric satellite
99 regions[3,6,30,31]. We generated 14 putative singly-infected Jurkat cell clones (Clones #1-14) using
100 low-MOI (0.3) infection with a replication-deficient HIV-1 reporter (V1/SBP-GFP), followed
101 by sorting of GFP/SBP-positive cells, limiting dilution, and expansion via culturing. Proviral
102 integration sites in these 14 clones were expected to be mostly in genes since HIV-1 integration
103 favors actively transcribing genomic regions[29]. To evaluate PRISM-seq performance within
104 centromeric regions typically associated with difficult-to-map satellite DNA repeats, we
105 analyzed two additional HIV-1-infected primary cell clones (#15 and #16) previously found to
106 each contain a single provirus within centromeric regions[31].

107

108 Without noise removal, our raw data showed 1-70 (min-max) unique proviral insertion sites in
109 clones #1-16 (Fig. S3a-c and S4a-c). The observation that some clones contained more than
110 one unique integration site suggested that methods for integration site determination should be
111 systematically evaluated for assay noise, for example, by leveraging the theoretically clonal
112 nature of our sample set.

113

114 An initial quality control filter designated each of our raw integration sites as either high- or
115 low-confidence. High-confidence was defined if an integration site (i) contained precise viral

116 DNA 5′ or 3′ ends linked to a human sequence without gaps/missing bases, and (ii) was detected
117 by both 5′ LTR and 3′ LTR viral-host junctions or (iii) had greater than 10 reads when identified
118 only by either 5′ LTR or 3′ LTR. All other integration sites were designated "low-confidence."

119

120 Among clones #1-14, 9/14 (64%) yielded a single high-confidence integration site (Fig. S3c).
121 Clone 6 and clone 10 showed multiple high-confidence integration sites, consistent with
122 expectations from Poisson statistics for infections at MOI 0.3[32]. Clone 8, clone 11, and clone
123 12 had more than one integrations, but all were found in the same gene, suggesting potential
124 assay-associated artifacts. We used two human reference genomes: hg38 and T2T-CHM13[33],
125 and their results were 100% concordant for clones #1-14. For clones 15 and 16, our data showed
126 multiple high-confidence integration sites with proviruses integrated within the centromeric
127 repeat regions of chromosomes 9 (8 sites) and 22 (2 sites), respectively (Fig. S4c), which were
128 only detected using T2T-CHM13 as the reference genome, while hg38 returned no maps. Due
129 to the clonal nature of these clones, we suspected repeat-driven ambiguity in mapping. All
130 ambiguities were further explored in the next section.

131

132 Across all 16 analyzed clones, high-confidence proviral integration sites were distributed
133 within 12 chromosomes spanning genic, intergenic, and repeat-rich centromeric satellite
134 regions, confirmed by the chromHMM model[34] based on distinct histone-mark ChIP-seq
135 datasets from the ENCODE database[35] (Supplementary Table 1). The average distance between
136 proviral integration sites and the nearest accessible chromatin region (ATAC-seq peak) varied
137 widely, ranging from 292 bp to 278,541 bp (Supplementary Table 1). These findings reflect
138 the diverse genomic landscape of proviral integration recovered by PRISM-seq.

139

140 **Analyses of low-confidence integration sites and implementation of additional quality**
141 **control filtering criteria**

142

143 In 15 of 16 analyzed clones, we detected 174 low-confidence secondary integration sites. We
144 next characterize this putative noise to design a quality-control filter. Of these low-confidence
145 integration sites, 57/174 (33%) mapped to the same gene or within ±10 kb relative to the
146 corresponding high-confidence integration sites (Fig. S3a-c, S4a-c, S5a-b, and supplementary
147 table 2). Given the clonal nature of most of the samples, we reasoned they were likely artifacts.
148 The remaining low-confidence sites were either in intergenic regions or in distinct genes.

149

150 To further investigate, we performed manual alignment of the host sequences of low-
151 confidence integration sites, and showed that many differed from the corresponding high-
152 confidence site only by short indels or 1–10 bp shifts in the junction sequence (e.g., clones 1,
153 2, 4, 8, 9, 10, 11, 13, 15, and 16) (Fig. S6-S12). These cases are likely the artifact signatures
154 previously highlighted to be found in integration site sequencing assays, mainly because of
155 PCR recombination, whole-genome amplification artifacts, sequencing errors, or ambiguous
156 mapping in repeat-rich chromosomal regions[20,36,37]. Collectively, the short indel cases, same
157 gene integration in a putative clone, and low read-count (≤10) when captured by only one
158 proviral end accounted for 174/174 (100%) of the low-confidence integration sites detected
159 among clones #1-16 (supplementary table 2).

160

161 Given these observations, we implemented a five-step quality control filter strategy to improve
162 the specificity of PRISM-seq (Fig. S13). Steps 1-4 are fully automated, and the accompanying
163 script is included with the BulkIntSiteR package; Step 5 requires manual curation and removes

164 5.7% (10/174) of the noise in our data. Step 5 is especially critical when proviruses are
165 integrated within repeat-rich genomic regions: without it, 7 of 8 (87.5%) and 1 of 2 (50%)
166 integration sites were called as high-confidence sites for the two clones (#15 and #16) with
167 centromeric proviral insertion site (Fig. S14a-e). Incorporating step 5 yielded a single high-
168 confidence integration site for each clone. All integration sites that passed the final filter are
169 designated "high confidence integration sites" and used for downstream analysis.

170

171 **PRISM-seq detection efficiency and cumulative recovery in an artificial mix**

172

173 Next, we examined whether PRISM-seq can accurately detect integration sites in samples with
174 more than one unique integration site, each with varying abundance. Variation in template input
175 copies, as well as multiple authentic proviral insertions, can limit detection efficiency and
176 accuracy due to PCR competition and recombination[12,20,38]. To model this scenario, we
177 generated two artificial pools composed of clones #1-14 (Pool 1 and Pool 2), each with
178 integration sites represented at different relative input frequencies, and used them to assess the
179 detection efficiency of PRISM-seq in a mixed template context (Fig. 2a-b).

180

181 With a 10-replicate setup, the abundant 40- and 10-copy-input integration site was detected at
182 100% efficiency in both pools (Fig. 2a-b and S15-17). For the single copy integration site
183 templates, after adjusting for Poisson sampling expectations, the probability of detection of all
184 high-confidence single copy integration sites increased with the number of replicates and
185 reached 100% at 6 and 5 replicates in pool 1 and pool 2, respectively (Fig. 2c-d), despite the
186 presence of a more abundant ~40 or ~10 copy template competitor. Moreover, the cumulative
187 recovery of all expected high-confidence integration sites was proportional to the number of

188 replicates in both pool 1 and pool 2 (Fig. 2e-f). Across 10 replicates, we recovered 14/17 (82%)
189 and 19/20 (95%) of all expected high-confidence integration sites in pool 1 and pool 2,
190 respectively. Overall, these findings demonstrate that PRISM-seq can reliably recover
191 integration sites in a sample expected to contain both rare and abundant templates, with an
192 average recovery of 89% integration sites across pools, and can detect down to a single-
193 molecule input at a 100% detection efficiency with 5-6 replicates.

194

195 **Estimation of the extent of clonal expansion within a sample using PRISM-seq**

196

197 Across replicate analyses, repeated detection of the same integration site signifies that the
198 original sample contains multiple copies of that viral-host junction (Fig. 3a). In a clinical
199 context, such an observation would indicate clonal expansion of infected cells. We reasoned
200 that the frequency with which an integration site appears across PRISM-seq replicates would
201 be an indirect quantitative measure of the relative size of the clone harboring that integration.
202 As proof of principle, we applied this framework to the post-filtered datasets for our artificial
203 pools. As anticipated, the dominant clones, present at ~400 copies in pool 1 and ~100 copies
204 in pool 2, were detected in the largest number of replicates and therefore showed the highest
205 clonal abundance (Fig. 3b-e).

206

207 **Application to longitudinal clinical samples from a person with HIV-1 on ART**

208

209 To demonstrate the clinical utility of PRISM-seq, we longitudinally sampled $CD4^+$ T cells from
210 a study participant living with HIV-1 before ART initiation (pre-ART), and 3 months, 1 year,

211 and 7 years post-ART (Supplementary Table 3). Each sample was expected to contain multiple
212 unique integration sites originating from infected cells exhibiting variable levels of clonal
213 expansion[39-42]. Using 2-5 replicates and after only sampling 100-250 HIV-containing templates
214 per time point, we recovered 69, 67, 11, and 17 unique and high-confidence integration sites
215 from each time point, respectively (Fig. 4a). The proportion of proviruses integrated into genic
216 or intergenic regions remained stable over time (Fig. 4b). The fraction of proviruses integrated
217 into ZNF genes increased from 2%-3% (pre-ART and 3 months) to 9% (1 year) and 18% (7
218 years) post-ART time points (Fig. 4c), consistent with previous reports indicating that
219 proviruses persisting in ZNF genes and transcriptionally repressive genomic contexts are more
220 likely to be transcriptionally silent and capable of long-term persistence[6,30].

221

222 With longitudinal sampling and the replicate setup, we tracked clonally expanded infected
223 cells: we observed identical viral integration sites in BUB1B-PAK6(chr15:40226639(+)) and
224 intergenic region (chr13:105139021(+)) across all 4 time points sampled over 7 years (Fig. 4d),
225 an indicator that HIV-1 infected cells with integrations at these loci were expanded *in vivo* and
226 persisted during treatment over time. Overall, our findings show that PRISM-seq could be
227 directly applied to clinical samples that had an extremely low target concentration, and
228 provided a wealth of integration site data and biological insights with minimal sample input.

229

230 **Discussion**

231

232 In this study, we present PRISM-seq and its accompanying software BulkIntSiteR. The assay
233 recovers integration sites from single-molecule input, captures diverse genomic contexts,
234 quantifies clonal expansion, and is compatible with *ex vivo* clinical samples.

Existing bulk integration site assays, including assays INSPIIRED[11,12], VISPA/VISPA2[13], QuickMap[14], VISA[15], MAVRIC[16], GeIST[17], IS-Seq[18], VSeq-Toolkit[19], do not support reliable single-molecule detection or broad performance across challenging genomic contexts. PRISM-seq overcomes these limitations through its MDA-enabled single-molecule capture workflow, which efficiently recovers rare viral-host junctions from minimal DNA input. Although MDA has been used previously in MIP-seq, that approach requires limiting dilution, is not designed for bulk integration site analysis, and is costly. PRISM-seq provides a bulk-input compatible, sensitive, and scalable platform for high-confidence integration site profiling.

Another key advance of PRISM-seq is the systematic characterization and mitigation of assay-associated noise, an aspect that has not been quantitatively defined in prior proviral integration site sequencing methods[12,20] nor formalized into a standardized filtering framework. Across biologically pure samples with a single provirus integration site, we identified recurring artifact classes arising from whole-genome amplification, PCR recombination, sequencing errors, and mapping ambiguity in repeat-rich regions. These artifacts manifested as displaced junctions, small indels, low-read viral-host chimeras detected by only one LTR end, and clusters of look-alike integration sites mapping within the same gene or genomic window as the true high-confidence site. Using clonal standards, including centromere-integrated proviruses resolvable only using T2T-CHM13, we were able to distinguish genuine proviral insertions from amplification- or sequencing-derived noise and develop a five-step quality control pipeline that removes >85–90% of raw artifacts while preserving bona fide integration sites.

258 Application of PRISM-seq to longitudinal samples from a study participant living with HIV-1
259 revealed persistent clones integrated within transcriptionally repressive, ZNF-rich genomic
260 regions. These findings reinforce prior observations that proviruses can integrate and persist in
261 epigenetically repressed environments, including KRAB-ZNF gene clusters and other
262 heterochromatic regions[2,6,30,43]. The compatibility of PRISM-seq with low sample input, and
263 its ability to capture low-abundance, genomically diverse proviruses directly from clinically-
264 derived material highlights its utility for high-resolution reservoir characterization.
265 Furthermore, our companion study (Pal et al., 2025, preprint available in BioRivX [cite])
266 applied PRISM-seq to a novel mouse model engrafted with HIV-infected cells, demonstrating
267 that proviral integration sites play a key determinant role in latency fate decisions in
268 immunocompromised settings, underscoring the biological insight gained through single-
269 template-level mapping of proviral integrations.

270

271 PRISM-seq has several limitations. First, clonal expansion estimates currently require replicate
272 reactions. Future iterations incorporating unique molecular identifiers (UMIs) could enable
273 direct quantification of clonal abundance from single reactions. Second, increasing input
274 concentration beyond ≥50 templates per reaction is possible[22] but means users should turn off
275 the same-gene filter (step 3 in Fig. S9 and see Monte Carlo stimulation Fig S5b) with modest
276 compromise to specificity. Third, while our quality-filtering framework effectively reduces
277 artifacts, it may occasionally exclude rare authentic sites detected by only one proviral end with
278 low read counts - manual check is always recommended. Finally, PRISM-seq provides precise
279 integration sites mapping but does not assess proviral genome intactness, which remains
280 specific to HIV-1 reservoir studies and should be complemented with assays that evaluate
281 proviral sequence integrity [5,44,45].

Overall, we have shown that PRISM-seq is an ultra-sensitive, streamlined, and economical method for detecting lentiviral integration sites with single-template resolution. Coupled with its automated bioinformatics pipeline BulkIntSiteR and noise-mitigating steps, the assay delivers a fully integrated workflow for high-confidence integration site identification, filtering, and annotation, especially for rare template detection.

Method

Viral-host junction amplification and sequencing

Genomic DNA was extracted (QIAGEN DNeasy kit), quantified for HIV-1 gag by droplet digital PCR (Bio-Rad), subjected to whole-genome amplification using multiple displacement amplification (MDA; QIAGEN REPLI-g Single Cell Kit), and processed using an adaptor-ligation PCR workflow adapted from the Clontech Lenti-X Integration Site Analysis Kit[5]. Final nested PCR was used to selectively enrich viral-host junction products and sequenced on an Illumina MiSeq platform (2 × 150 bp). Full protocol details, origin of PRISM-seq, reagent specifications, and cost are provided in the Extended Methods. Technical replicates were defined as independent MDA reactions from the same DNA extraction; 10 replicates were used for pooling experiments, 2–5 for clinical samples, and none for 8E5/LAV or clones #1–16.

Identification of HIV-1 IS using BulkIntSiteR

Illumina sequencing-derived FASTQ files were processed using BulkIntSiteR, a one-click R-based, fully automated pipeline designed to detect, map, and annotate HIV-1 integration sites

307 in the human genome. BulkIntSiteR is cross-platform and can handle non-Illumina data. All
308 scripts are available freely on GitHub (https://github.com/guineverelee/BulkIntSiteR/).
309 Briefly, each small read was independently assessed for the presence of high-quality viral ends
310 (containing HXB2 1 or 9719 using BLAST+ suite[46]). For each read positively identified to
311 contain a viral end, the non-viral portion of the read was mapped to the human genome using
312 BLAT[47]. See GitHub documentation for the full algorithm.
313
314 **Generation of 14 putative clones with a single integrated provirus**
315 To generate the V1/SBP-GFP HIV-1 reporter virus, 7.5×10^7 293T cells (ATCC #CRL-3216;
316 RRID: CVCL_0063) were seeded in 875 cm^2 5-layer flasks 1 day prior to transfection. 293T
317 cells were transfected with 75 µg pV1/SBP-GFP, 75 µg pCRV1/NLGagPol, and 15 µg pVSV-
318 G (Yee et al., 1994) using polyethyleneimine (PEI). After 12 hours, the medium was replaced
319 with fresh growth medium. Supernatants were collected 48 hours post-transfection, filtered,
320 and concentrated by ultracentrifugation over a 20% sucrose cushion. The pellet containing the
321 virus was resuspended in PBS + 10% BSA, aliquoted, and stored at -80 °C. The viral stock titre
322 (IU/ml) was determined by infecting MT4-LTR-GFP cells and using flow cytometry to assess
323 GFP expression.
324
325 For infection, Jurkat and MOLT-4 cells were infected with V1/SBP-GFP at a multiplicity of
326 infection (MOI) of 0.3. Infected cells were enriched using streptavidin bead-based magnetic
327 selection. Single-cell clones of infected Jurkat and MOLT-4 cells were isolated by limiting
328 dilution by diluting to 0.5 cells per well in RPMI growth medium, and 100 µL of the suspension
329 was seeded into each well of 96-well plates. The clonal outgrowth was monitored
330 microscopically, and wells containing single-cell-derived colonies were transferred to fresh
331 wells for expansion.

332

333 **Clinical samples**

334 The HIV-1-infected study participant was recruited at the Maple Leaf Clinic in Toronto,
335 Canada. PBMC samples were used according to protocols approved by the respective
336 Institutional Review Boards. $CD4^+$ T cells were enriched from total PBMCs using a $CD4^+$ T
337 Cell Isolation Kit (STEMCELL Technologies, catalog 17952). Clinical and demographical
338 characteristics of the study participant are summarized in Supplemental Table 3.

339

340

341

342

343

344 **Extended methods (Supplementary)**

345

346 **Cost and origin**

347 PRISM-seq is an evolution of our previously published matched integration site and proviral
348 sequencing (MIP-seq) protocol[5], which itself was adapted from the Clontech LentiX
349 integration site analysis method (catalog #631263). MIP-seq was designed to co-capture HIV
350 integration sites together with full-length proviral genomes at single-input-template resolution.
351 In contrast, PRISM-seq is optimized for multiple-template input and focuses specifically on
352 capturing viral integration sites, enabling substantially higher throughput at dramatically
353 reduced cost. At the time of writing, while MIP-seq was prohibitively expensive, with an
354 estimated cost of $240,000 to capture 100 integration sites and their associated proviral
355 genomes. PRISM-seq offers an economical solution at $280 per sample with detection
356 sensitivity down to a single template, making the assay broadly accessible. In addition, PRISM-
357 seq is fully supported by a publicly available bioinformatics (BulkIntSiteR) and quality control
358 pipeline, which automates integration site calling, annotation, and quality assessment.

359

360 **Input template quantification by Droplet digital PCR (ddPCR)**

361 DdPCR quantifications were performed to estimate the HIV-1 *gag*-containing proviral
362 template concentration[48]. $CD4^+$ T cells were enriched from total PBMCs using a $CD4^+$ T Cell
363 Isolation Kit (STEMCELL Technologies, catalog 17952) and subjected to DNA extraction
364 using commercial kits (QIAGEN DNeasy, 69504). We amplified total HIV-1 DNA using
365 ddPCR (Bio-Rad), with primers and probes described previously (127-bp 5′-LTR-gag
366 amplicon; HXB2 coordinates 684-810). PCR was performed using the following program:
367 95°C for 10 minutes; 45 cycles (95°C for 30 seconds, 60°C for 1 minute); 98°C for 10 minutes.

368 The droplets were subsequently read by a QX100 droplet reader, and data were analyzed using
369 Quanta- Soft software (BIO-RAD).
370
371 **Multiple Displacement Amplification**
372 Genomic DNA was isolated from the V1/SBP-GFP infected single cell clone using commercial
373 kits (QIAGEN DNeasy Blood & Tissue Kit, catalog 69504), according to the manufacturer's
374 instructions. To maximize the likelihood of capturing a single HIV-1 proviral DNA copy, the
375 genomic DNA was amplified using multiple displacement amplification (MDA) with phi29
376 polymerase (QIAGEN REPLI-g Single Cell Kit, catalog 150345). Buffers DLB, D1
377 (denaturation buffer), and N1 (neutralization buffer) were prepared according to the
378 manufacturer's protocol. The genomic DNA (2.5 µL) was denatured by the addition of 2.5 µL
379 of buffer D1, followed by an incubation at room temperature for 3 mins. The mix was
380 neutralized by adding 5 µL of buffer N1. 40 µL of the MDA master mix, consisting of nuclease-
381 free water, REPLI-g single cell Reaction Buffer, and REPLI-g single cell DNA Polymerase,
382 was added to the 10 µL denatured genomic DNA sample. The final 50 µL reaction was
383 incubated at 30°C for 4 h, followed by heat inactivation at 65°C for 3 mins. The amplified
384 product was purified using AMPure XP beads (Beckman Coulter) and was used for
385 downstream integration site analysis.
386
387 **Enzymatic digestion of genomic DNA**
388 This protocol was adapted and modified from the Clontech Lenti-X Integration Site Analysis
389 Kit (catalog #631263), which, in its standard form without MDA, lacks sufficient sensitivity to
390 detect single-copy proviral DNA in a sample. Purified MDA products were digested using three
391 blunt-end restriction enzymes - HpaI (NEB #R0105L), SspI (NEB #R3132L), and DraI (NEB
392 #R0129L) - to fragment genomic DNA and facilitate subsequent ligation-based integration site

393 analysis. Briefly, 5 μL of purified MDA product was added to a master mix containing
394 nuclease-free water, 10× restriction enzyme buffer, and either DraI or SspI (20 U μL^{-1}) or HpaI
395 (5 U μL^{-1}), to a final reaction volume of 50 μL. Reactions were incubated at 37 °C for 18 hours
396 to ensure complete digestion. The digested genomic DNA was then purified using AMPure XP
397 beads and carried forward to adaptor ligation.
398
399 **Ligation of the genome-walker adaptor to the digested genomic DNA**
400 The genomic DNA fragments are ligated with double-stranded T-linker DNA (genome walker)
401 (see section below for exact sequence), in which the shorter strand is an oligonucleotide with
402 a 5' end phosphorylated (to enable efficient linker ligation) and the 3' end is modified with an
403 amino modification, which limits the extension of the short strand by DNA polymerase and
404 restricts the linker self-ligation amplicons amplification**.** Digested genomic DNA (4.8 μL) was
405 ligated to adaptors in an 8 μL reaction containing 1.9 μL adaptor (25 μM), 0.8 μL 10× ligation
406 buffer, and 0.5 μL T4 DNA ligase (6 U μL^{-1}; NEB #M0202L). Ligation was performed at 16
407 °C overnight in a thermal cycler to ensure temperature stability. The reaction was terminated
408 by heat inactivation at 70 °C for 5 min, followed by dilution with 32 μL of DEPC-treated water
409 (Ambion #4387937) to a final volume of 40 μL.
410
411 **Viral-host junction PCR amplification**
412 The adaptor ligated genomic DNA fragments were used as a template for nested PCR
413 amplification of viral-host junctions. The 5′LTR HIV-1 junction was amplified using LSP1
414 (GCTTCAGCAAGCCGAGTCCTGCGTCGAG) and LSP2
415 (GCTCCTCTGGTTTCCCTTTCGCTTTCAA) as forward primers, both derived directly from
416 the CloneTech LentiX kit. The 3′LTR HIV-1 junction was amplified using LestralLTR1
417 (CTTAAGCCTCAATAAAGCTTGCCTTGAG) and LestralLTR2

418 (AGACCCTTTTAGTCAGTGTGGAAAATC) [5]. These were paired with adaptor-specific
419 reverse primers AP1 (GTAATACGACTCACTATAGGGC) and AP2
420 (ACTATAGGGCACGCGTGGT) for the first and second PCR reactions, respectively. The
421 AP1 primer is specific to the 5' single-stranded region of the T-linker to avoid unwanted
422 amplification of the host DNA. The adaptor sequence was adapted from the CloneTech LentiX
423 kit (catalog #631263).
424 Each 25 µL PCR reaction contained 1× reaction buffer, 1× dNTP mix, 0.2 µM of each primer,
425 and Advantage 2 polymerase (Clontech Advantage® 2 PCR Kit, Cat. #639206). For the first
426 PCR, 15 µL of master mix was combined with 10 µL of diluted adaptor-ligated genomic DNA,
427 using AP1/LSP1 for 5′LTR junctions and AP1/LestralLTR1[5] for 3′LTR junctions. Cycling
428 conditions were: 7 cycles of 94 °C for 25 s and 72 °C for 3 min, followed by 32 cycles of 94
429 °C for 25 s and 67 °C for 3 min, and a final extension at 67 °C for 7 min.
430 For the second (nested) PCR, 24 µL of master mix was combined with 1 µL of first-round PCR
431 product, using AP2/LSP2 for 5′LTR junctions and AP2/LestralLTR2[5] for 3′LTR junctions.
432 Cycling conditions were: 5 cycles of 94 °C for 25 s and 72 °C for 3 min, followed by 20 cycles
433 of 94 °C for 25 s and 67 °C for 3 min, with a final extension at 67 °C for 7 min. Both 5′LTR
434 and 3′LTR viral-host junctions were successfully amplified under these conditions. The
435 resulting PCR products were subjected to Illumina MiSeq paired-end 150 base pairs (bp)
436 sequencing.
437
438 **Cell culture**
439 Jurkat cells and MOLT-4 cells were cultured at 37 °C and 5% CO_2 in RPMI supplemented with
440 10 % FCS and Gentamycin. All cell lines were monitored by the SG-PERT assay[49] regularly
441 to ensure the absence of retroviral contamination. Cell lines used in this study were devoid of
442 mycoplasma contamination.

443 **Availability**

444 BulkIntSiteR is available under an open-source license at:

445 https://github.com/guineverelee/BulkIntSiteR/

446

447 **Acknowledgements**

448 This work was supported by NIH grant Research Enterprise to Advance a Cure for HIV
449 (REACH) 5UM1AI164565-05 and the Center for the Structural Biology of HIV RNA (CRNA)
450 U54 AI170660. We acknowledge Dr. Douglas Barrows, Bioinformatics Analyst,
451 Bioinformatics Resource Center, The Rockefeller University, New York. The authors gratefully
452 acknowledge the MGH DNA core facility, Massachusetts, USA. This article is subject to
453 HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a
454 nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their
455 research articles. Pursuant to those licenses, the author-accepted manuscript of this article can
456 be made freely available under a CC BY 4.0 license immediately upon publication.

457

458

459

460

461

462

463

464

465 **References**

466 1 Janssens, J., De Wit, F., Parveen, N. & Debyser, Z. Single-Cell Imaging Shows That
467 the Transcriptional State of the HIV-1 Provirus and Its Reactivation Potential
468 Depend on the Integration Site. *mBio* **13**, e00007–00022 (2022).
469 https://doi.org/10.1128/mbio.00007-22
470 2 Teixeira, A. R. *et al.* Transcription of HIV-1 at sites of intact latent provirus
471 integration. *Journal of Experimental Medicine* **221** (2024).
472 https://doi.org/10.1084/jem.20240391
473 3 Jiang, C. *et al.* Distinct viral reservoirs in individuals with spontaneous control of
474 HIV-1. *Nature* **585**, 261–267 (2020). https://doi.org/10.1038/s41586-020-2651-8
475 4 Jadlowsky, J. K. *et al.* Long-term safety of lentiviral or gammaretroviral gene-
476 modified T cell therapies. *Nature Medicine* **31**, 1134–1144 (2025).
477 https://doi.org/10.1038/s41591-024-03478-6
478 5 Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal
479 positions during prolonged antiretroviral therapy. *J Clin Invest* **129**, 988–998
480 (2019). https://doi.org/10.1172/jci124291
481 6 Huang, A. S. *et al.* Integration features of intact latent HIV-1 in CD4+ T cell clones
482 contribute to viral persistence. *Journal of Experimental Medicine* **218** (2021).
483 https://doi.org/10.1084/jem.20211427
484 7 Chun, T.-W. *et al.* Presence of an inducible HIV-1 latent reservoir during highly
485 active antiretroviral therapy. *Proceedings of the National Academy of Sciences*
486 **94**, 13193–13197 (1997). https://doi.org/doi:10.1073/pnas.94.24.13193
487 8 Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong
488 persistence of HIV-1, even in patients on effective combination therapy. *Nature*
489 *Medicine* **5**, 512–517 (1999). https://doi.org/10.1038/8394
490 9 Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged
491 suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
492 https://doi.org/10.1126/science.278.5341.1291
493 10 Ronen, K. *et al.* Distribution of lentiviral vector integration sites in mice following
494 therapeutic gene transfer to treat β-thalassemia. *Mol Ther* **19**, 1273–1286 (2011).
495 https://doi.org/10.1038/mt.2011.20
496 11 Berry, C. C. *et al.* INSPIIRED: Quantification and Visualization Tools for Analyzing
497 Integration Site Distributions. *Mol Ther Methods Clin Dev* **4**, 17–26 (2017).
498 https://doi.org/10.1016/j.omtm.2016.11.003
499 12 Sherman, E. *et al.* INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New
500 DNA Integration in Cellular Genomes. *Mol Ther Methods Clin Dev* **4**, 39–49
501 (2017). https://doi.org/10.1016/j.omtm.2016.11.002
502 13 Calabria, A. *et al.* VISPA: a computational pipeline for the identification and
503 analysis of genomic vector integration sites. *Genome Medicine* **6**, 67 (2014).
504 https://doi.org/10.1186/s13073-014-0067-5
505 14 Appelt, J. U. *et al.* QuickMap: a public tool for large-scale gene therapy vector
506 insertion site mapping and analysis. *Gene Ther* **16**, 885–893 (2009).
507 https://doi.org/10.1038/gt.2009.37
508 15 Hocum, J. D. *et al.* VISA - Vector Integration Site Analysis server: a web-based
509 server to rapidly identify retroviral integration sites from next-generation

sequencing. *BMC Bioinformatics* **16**, 212 (2015). https://doi.org/10.1186/s12859-015-0653-6

16 Huston, M. W. *et al.* Comprehensive Investigation of Parameter Choice in Viral Integration Site Analysis and Its Effects on the Gene Annotations Produced. *Human Gene Therapy* **23**, 1209–1219 (2012). https://doi.org/10.1089/hum.2011.037

17 LaFave, M. C. *et al.* MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Research* **42**, 4257–4269 (2014). https://doi.org/10.1093/nar/gkt1399

18 Yan, A. *et al.* IS-Seq: a bioinformatics pipeline for integration sites analysis with comprehensive abundance quantification methods. *BMC Bioinformatics* **24**, 286 (2023). https://doi.org/10.1186/s12859-023-05390-1

19 Afzal, S., Fronza, R. & Schmidt, M. VSeq-Toolkit: Comprehensive Computational Analysis of Viral Vectors in Gene Therapy. *Mol Ther Methods Clin Dev* **17**, 752–757 (2020). https://doi.org/10.1016/j.omtm.2020.03.024

20 Wells, D. W. *et al.* An analytical pipeline for identifying and mapping the integration sites of HIV and other retroviruses. *BMC Genomics* **21**, 216 (2020). https://doi.org/10.1186/s12864-020-6647-4

21 Niederer, H. A. *et al.* HTLV-1 proviral integration sites differ between asymptomatic carriers and patients with HAM/TSP. *Virology Journal* **11**, 172 (2014). https://doi.org/10.1186/1743-422X-11-172

22 Hsiao, F. *et al.* Tissue memory CD4+ T cells expressing IL-7 receptor-alpha (CD127) preferentially support latent HIV-1 infection. *PLoS Pathog* **16**, e1008450 (2020). https://doi.org/10.1371/journal.ppat.1008450

23 Milone, M. C. & O'Doherty, U. Clinical use of lentiviral vectors. *Leukemia* **32**, 1529–1541 (2018). https://doi.org/10.1038/s41375-018-0106-0

24 Kane, M. *et al.* MX2 is an interferon-induced inhibitor of HIV-1 infection. *Nature* **502**, 563–566 (2013). https://doi.org/10.1038/nature12653

25 Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nature Methods* **11**, 783–784 (2014). https://doi.org/10.1038/nmeth.3047

26 Lajoie, M. J. *et al.* Designed protein logic to target cells with precise combinations of surface antigens. *Science* **369**, 1637–1643 (2020). https://doi.org/doi:10.1126/science.aba6527

27 Kim, G. B. *et al.* High-affinity mutant Interleukin-13 targeted CAR T cells enhance delivery of clickable biodegradable fluorescent nanoparticles to glioblastoma. *Bioact Mater* **5**, 624–635 (2020). https://doi.org/10.1016/j.bioactmat.2020.04.011

28 Deichmann, M., Bentz, M. & Haas, R. Ultra-sensitive FISH is a useful tool for studying chronic HIV-1 infection. *J Virol Methods* **65**, 19–25 (1997). https://doi.org/10.1016/s0166-0934(96)02164-7

29 Schröder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002). https://doi.org/10.1016/s0092-8674(02)00864-4

30 Lian, X. *et al.* Progressive transformation of the HIV-1 reservoir cell profile over two decades of antiviral therapy. *Cell Host Microbe* **31**, 83–96.e85 (2023). https://doi.org/10.1016/j.chom.2022.12.002

31 Bittar, C. *et al.* HIV-1 expression is heterogeneous among clones of CD4+ T cells carrying authentic intact latent proviruses. *bioRxiv*, 2025.2008.2005.668730 (2025). https://doi.org/10.1101/2025.08.05.668730

32 Mistry, B. A., D'Orsogna, M. R. & Chou, T. The Effects of Statistical Multiplicity of Infection on Virus Quantification and Infectivity Assays. *Biophys J* **114**, 2974–2985 (2018). https://doi.org/10.1016/j.bpj.2018.05.005

33 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022). https://doi.org/doi:10.1126/science.abj6987

34 Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12**, 2478–2492 (2017). https://doi.org/10.1038/nprot.2017.124

35 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). https://doi.org/10.1038/nature14248

36 Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research* **43**, e143–e143 (2015). https://doi.org/10.1093/nar/gkv717

37 McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol Biol Int* **2014**, 287430 (2014). https://doi.org/10.1155/2014/287430

38 Kalle, E., Kubista, M. & Rensing, C. Multi-template polymerase chain reaction. *Biomol Detect Quantif* **2**, 11–29 (2014). https://doi.org/10.1016/j.bdq.2014.11.002

39 Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014). https://doi.org/10.1126/science.1254194

40 Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014). https://doi.org/10.1126/science.1256304

41 Lee, G. Q. *et al.* Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *J Clin Invest* **127**, 2689–2696 (2017). https://doi.org/10.1172/jci93289

42 Hiener, B. *et al.* Identification of Genetically Intact HIV-1 Proviruses in Specific CD4+ T Cells from Effectively Treated Participants. *Cell Reports* **21**, 813–822 (2017). https://doi.org/https://doi.org/10.1016/j.celrep.2017.09.081

43 Dragoni, F. *et al.* Proviral location affects cognate peptide-induced virus production and immune recognition of HIV-1-infected T cell clones. *J Clin Invest* **133** (2023). https://doi.org/10.1172/jci171097

44 Artesi, M. *et al.* PCIP-seq: simultaneous sequencing of integrated viral genomes and their insertion sites with long reads. *Genome Biology* **22**, 97 (2021). https://doi.org/10.1186/s13059-021-02307-0

45 Sadri, G. *et al.* Development and validation of HIV SMRTcap for the characterization of HIV-1 reservoirs across tissues and subtypes. *bioRxiv*, 2025.2005.2001.651657 (2025). https://doi.org/10.1101/2025.05.01.651657

46 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990). https://doi.org/https://doi.org/10.1016/S0022-2836(05)80360-2

47 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656–664 (2002). https://doi.org/10.1101/gr.229202

48 Bruner, K. M. *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019). https://doi.org/10.1038/s41586-019-0898-8

49 Pizzato, M. *et al.* A one-step SYBR Green I-based product-enhanced reverse transcriptase assay for the quantitation of retroviruses in cell culture supernatants. *J Virol Methods* **156**, 1–7 (2009). https://doi.org/10.1016/j.jviromet.2008.10.012

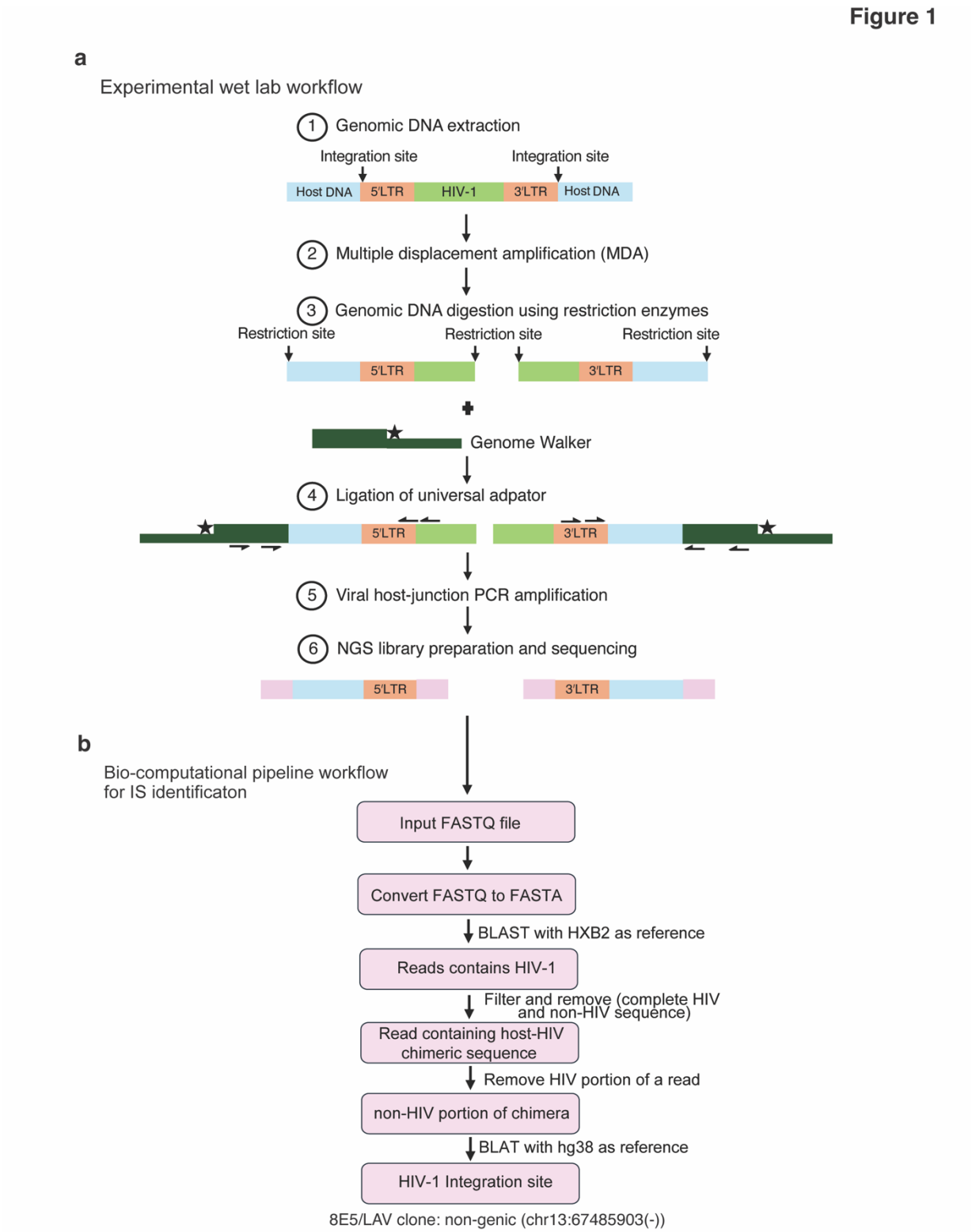


Figure 1. Experimental and bioinformatic workflow for proviral integration sites identification using PRISM-seq. (a) Experimental workflow of PRISM-seq. The viral-host junction is specifically amplified using the gDNA samples of HIV-1 infected or lentiviral vector-

transduced cell samples and sequenced. **(b)** The bioinformatics steps involved in the BulkIntSiteR pipeline to process the raw sequencing data and generate an output containing the list of proviral integration sites. BulkIntSiteR normalizes junction coordinates to the 5′ LTR convention, accounting for the characteristic 4-bp offset between the 5′ and 3′ LTR junctions (Craigie and Bushman, 2012). The large disparity in junction-containing reads between the 5′ and 3′ LTR integration sites libraries (e.g. 8E5/LAV benchmarking experiment) likely reflects the differences in amplicon composition, relative to Illumina tagmentation and read length. The 5′ LTR amplicon contains ~681 nt of viral sequence, whereas the 3′ LTR amplicon contains only ~34 nt. Because tagmentation generates fragments of ~300 bp on average and the sequencing read length was 150 bp, most 5′ LTR reads fall entirely within the viral portion of the amplicon and never reach the host sequence, which would markedly reducing the number of recoverable viral-host chimeras.

Figure 2

a Pool 1 post-filter

Sample	High-confidence integration site	Average copies per replicate	Number of replicates									
			1	2	3	4	5	6	7	8	9	10
1	REV3L (chr6:111352829(-))	40	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
2	ESYT2 (chr7:158757908(-))	1	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
3	NPHP4 (chr1:5984755(+))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	CCDC88A (chr2:55417185(+))	1	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
5	EIF4G3 (chr1:20971026(-))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	non-genic (chr2:1783182(+))	1	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
6	ZNHIT6 (chr1:85705560(-))	1	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
6	CARD8 (chr19:48220395(-))	1	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE
7	LOC105375433 (chr7:102416320(-))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	KATNA1 (chr6:149598014(-))	1	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
9	ANKRD11 (chr16:89469214(-))	1	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
10	RBM10 (chrX:47154026(+))	1	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	POLE (chr12:132654892(+))	1	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
10	WASF1 (chr6:110137385(-))	1	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
10	VPS54 (chr2:63945292(+))	1	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
10	ZNF823 (chr19:11726895(+))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
11	TAPBP (chr6:4496473(-))	1	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE

b Pool 2 post-filter

Sample	High-confidence integration site	Average copies per replicate	Number of replicates									
			1	2	3	4	5	6	7	8	9	10
1	REV3L (chr6:111352829(-))	10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
2	ESYT2 (chr7:158757908(-))	10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
3	NPHP4 (chr1:5984755(+))	10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
4	CCDC88A (chr2:55417185(+))	10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE
5	EIF4G3 (chr1:20971026(-))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
6	non-genic (chr2:1783182(+))	1	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
6	ZNHIT6 (chr1:85705560(-))	1	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
6	CARD8 (chr19:48220395(-))	1	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
7	LOC105375433 (chr7:102416320(-))	1	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	KATNA1 (chr6:149598014(-))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE
9	ANKRD11 (chr16:89469214(-))	1	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
10	RBM10 (chrX:47154026(+))	1	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
10	POLE (chr12:132654892(+))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	WASF1 (chr6:110137385(-))	1	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
10	VPS54 (chr2:63945292(+))	1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
10	ZNF823 (chr19:11726895(+))	1	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	TAPBP (chr6:4496473(-))	1	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
12	PLCB1-IT1 (chr20:8251940(+))	1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
13	AKAP13 (chr15:85640065(-))	1	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
14	GIMAP8 (chr7:150471878(+))	1	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE

5′LTR and 3′LTR 5′LTR 3′LTR

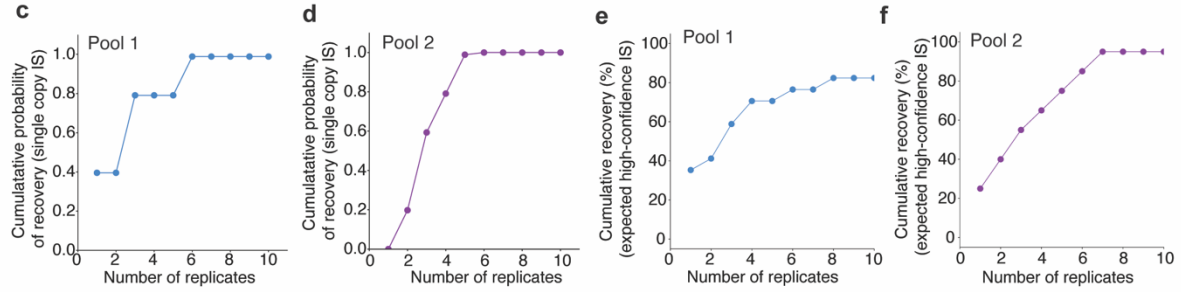


Figure 2. Single-molecule detection efficiency of PRISM-seq in artificial mixes. (a and b) Two artificial pools were constructed to test PRISM-seq's ability to recover integration sites in samples with multiple unique integration sites at varying relative abundance. Pool 1 contained 11 of the 14 clones #1-14 encompassing at least 17 unique high confidence integration sites, whereas pool 2 contained all 14 clones #1-14 with at least 20 unique high confidence integration sites (designs described in Methods). To model differences in template clonal abundance, pool 1 included one dominant clone (Clone #1) at ~400 genome copies and ten minor clones at 10 copies each; pool 2 included four abundant clones at 100 copies each and ten minor clones at 10 copies each. Each pool was divided into ten independent genomic DNA sampling replicates resulting in each reaction containing an average of ~50 total proviral copies per reaction, equivalent to an average of ~40

copies of the dominant clone plus ten single-copy clones for pool 1, and four 10-copy clones plus ten single-copy clones for pool 2. Each of the 10 replicates was independently subjected to PRIMS-seq. The Table shows the number of high-confidence integration sites present in the artificial pool 1 (a) and pool 2 (b), respectively. The colored matrix shows the recovery of expected high-confidence integration sites across 10 replicates. The color highlights if a high-confidence integration site was recovered by both 5′LTR and 3′LTR (blue), or exclusively by either 5′LTR (red), or 3′LTR (yellow) viral-host junction. The true/false denotes if a high-confidence integration site was recovered by either proviral end. **(c and d)** Cumulative recovery of single-copy clone integration sites after adjusting for Poisson sampling expectation across 10 replicates. Clones #6 and #10 were excluded from this analysis. **(e and f)** Percent cumulative recovery of all expected high-confidence integration sites from pool 1 and pool 2, respectively.

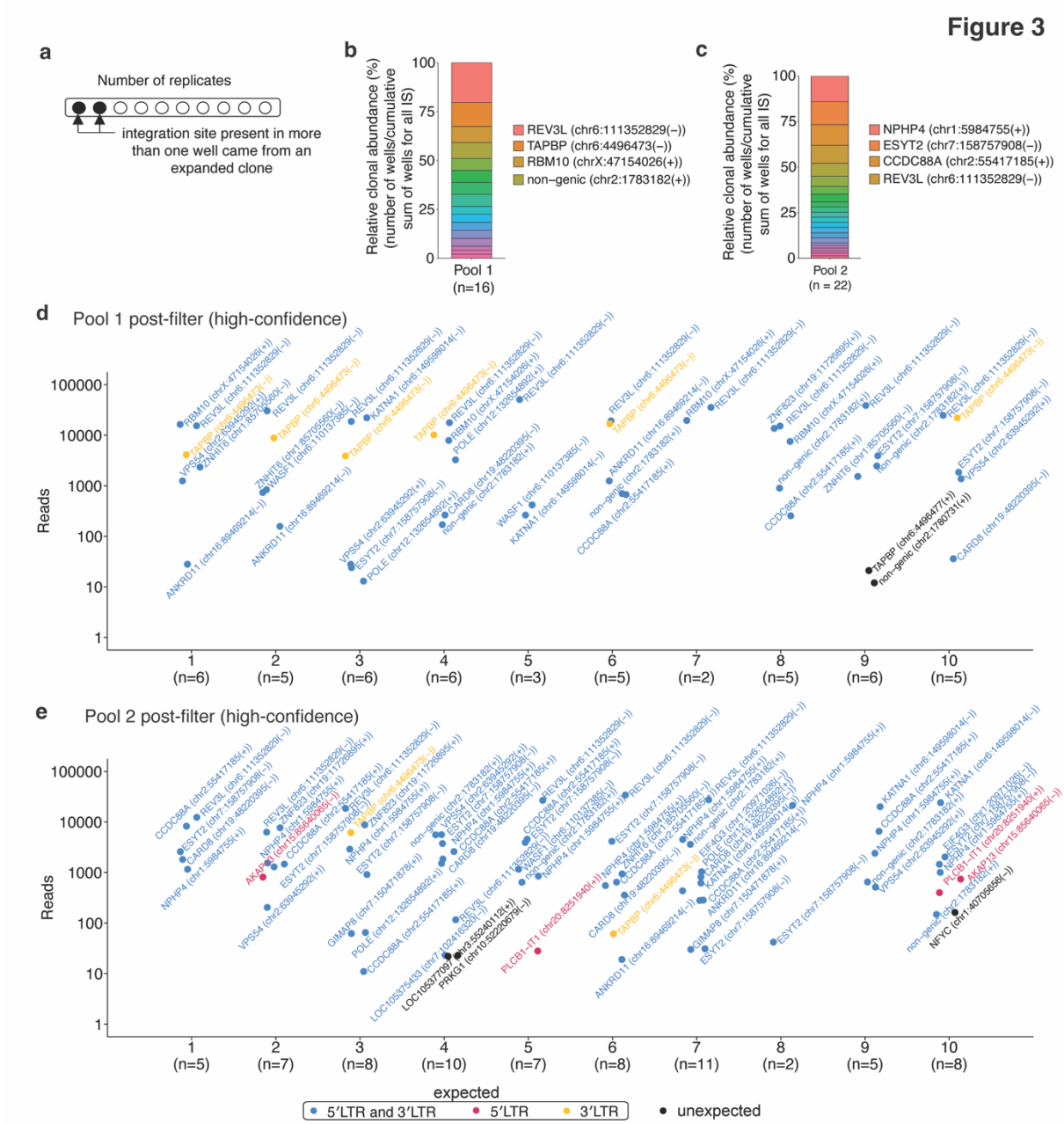


Figure 3. Post-quality filter recovery of high-confidence integration sites and estimation of clonal abundance. (a) Schematics showing the identification of expanded clones using PRISM-seq. We used the post-filtered data to evaluate integration site authenticity and to estimate clonal abundance in mixed, non-clonal samples, Pool 1 and Pool 2. The post-filtered datasets resulted in the recovery of a total of 16 high-confidence unique integration sites in pool 1 and 22 in pool 2, closely matching the expected input sites (17 for pool 1 and 20 for pool 2), with false-positive calls limited to 2 of 16 (12.5%) and 3 of 22 (13.6%), respectively. For each integration site, we counted the number of replicate reactions in which it was recovered. Integration sites detected in multiple replicates are highly likely to represent true biological events, whereas those observed only once

are more likely to reflect residual noise. **(b and c)** post-filter estimation of the relative clonal abundance of individual clones found in pool 1 and pool 2 after 10 replicates, respectively. The top 4 abundant clones are labelled in the legend. **(d and e)** The dot plots show the total number of integration sites retrieved from individual sequencing replicates of pool 1 (a) and pool 2 (b), respectively. The expected high-confidence integration sites are highlighted if recovered by both 5′LTR and 3′LTR (blue), or exclusively by either 5′LTR (red), or 3′LTR (yellow) viral-host junction. The unexpected integration sites are colored in black. The quality filter led to 2 of 16 (12.5% noise) and 3 of 22 (13.6% noise) unexpected integration sites identification as noise in pool 1 and pool 2, respectively.

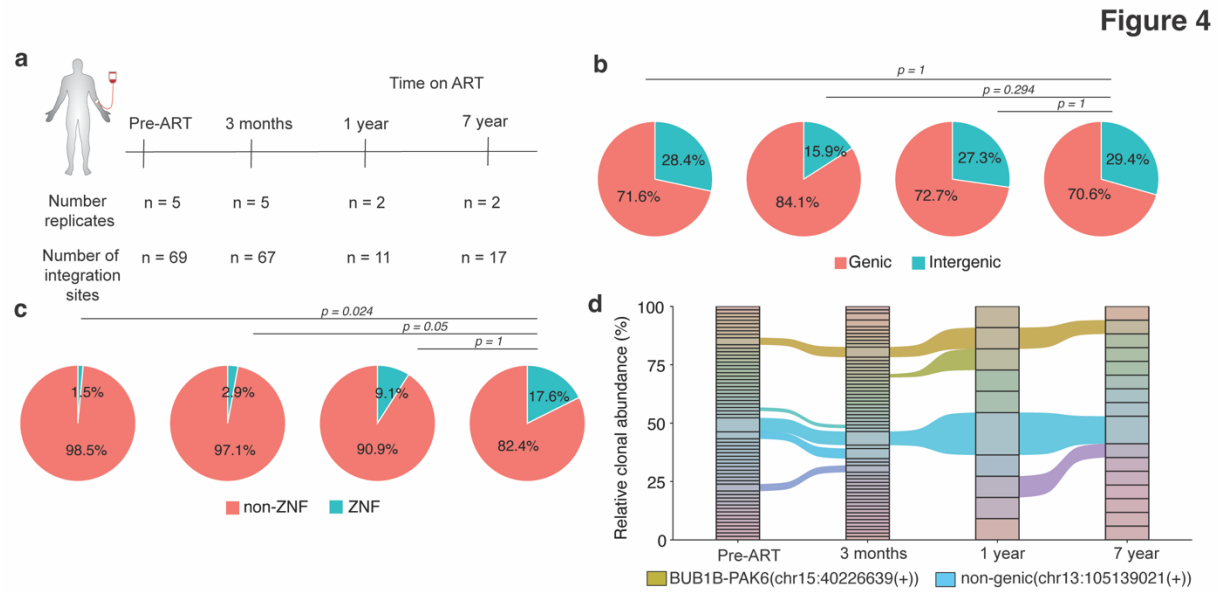


Figure 4. Longitudinal clonal tracking and identification of proviral integration sites in a person with HIV-1 infection. **(a)** Number of unique integration sites identified using PRISM-seq from an HIV-1 infected person under long-term ART treatment. Across the number of replicates performed, we sampled in total 250, 250, 100, and 100 HIV-containing templates at each indicated time point. **(b)** The precent genic and intergenic distribution of proviral integration sites. **(c)** The distribution of proviruses in ZNF and non-ZNF genes in the genome. P-value denotes Fisher's exact test. **(d)** Clonal tracking of HIV-1 infected cell clones over time under ART. The shared clones identified at all the time points are labelled with their respective proviral integration sites.

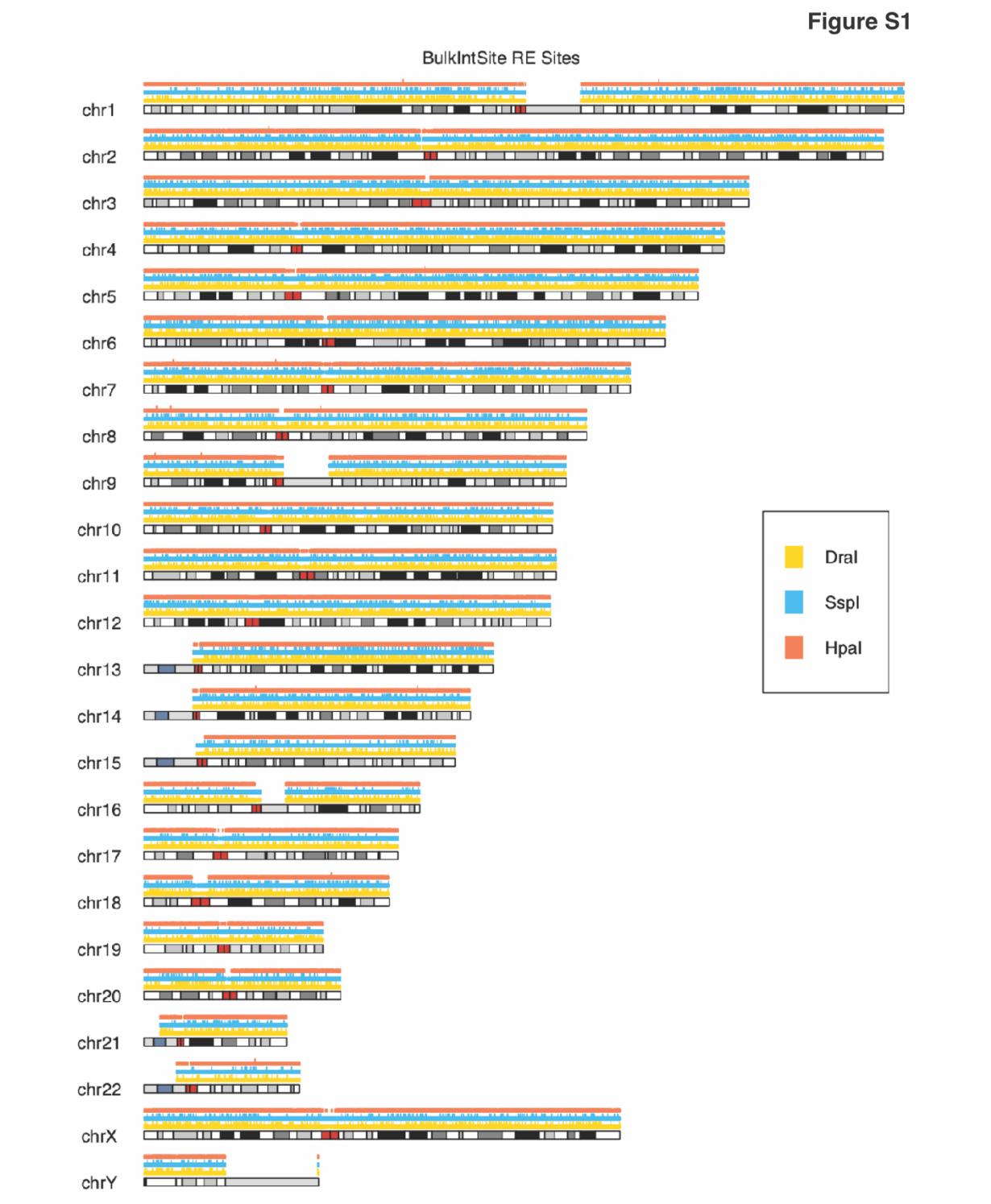


Figure S1. Genome-wide distribution of the restriction enzyme recognition site. The karyoplot shows the DraI (yellow), SspI (cyan), and HpaI (orange) restriction sites distribution genome-wide on the human reference genome assembly hg38. RE, restriction enzymes.

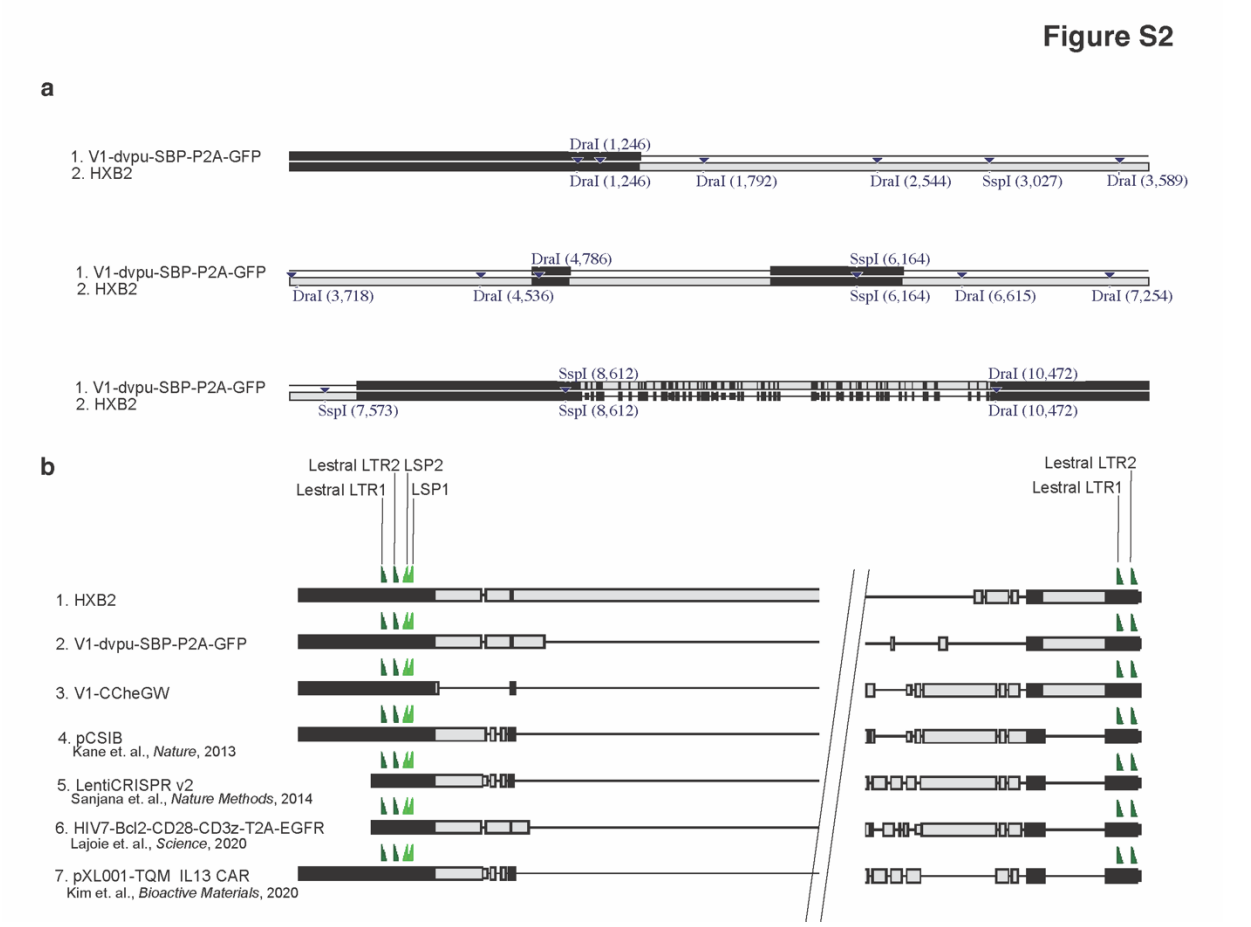


Figure S2. Distribution of restriction enzyme sites within the provirus genome and PRISM-seq primer binding sites compatibility across HIV and lentiviral vector delivery systems. (a) Location of DraI and SspI restriction sites in the genome of the full-length HIV-1 reference sequence (HXB2) and replication-deficient HIV-1 reporter (V1-SBP-GFP) from 5′ to 3′ direction. The restriction sites coordinate in the proviral genome are mentioned within brackets next to each restriction enzyme. The black and gray colors represent the identical and non-identical sequences, respectively, between the V1-SBP-GFP and HXB2 genome sequence. **(b)** Multiple sequence alignment of HXB2, replication-deficient single-round HIV-1 reporters (V1- SBP-GFP and V1-CCheGW), pCSIB (lentivirus-based overexpression construct to generate stable cell lines), LentiCRISPRv2 (CRISPR/Cas9-based genome editing lentiviral vector), HIV7-Bcl2-CD28-CD3z-T2A-EGFR, and pXL001-TQM IL13 CAR lentiviral vectors. The primer binding sites for 5′LTR (LSP1 and LSP2) and 3′LTR (LestralLTR1 and LestralLTR2) viral-host junction amplification are conserved across these sequences.

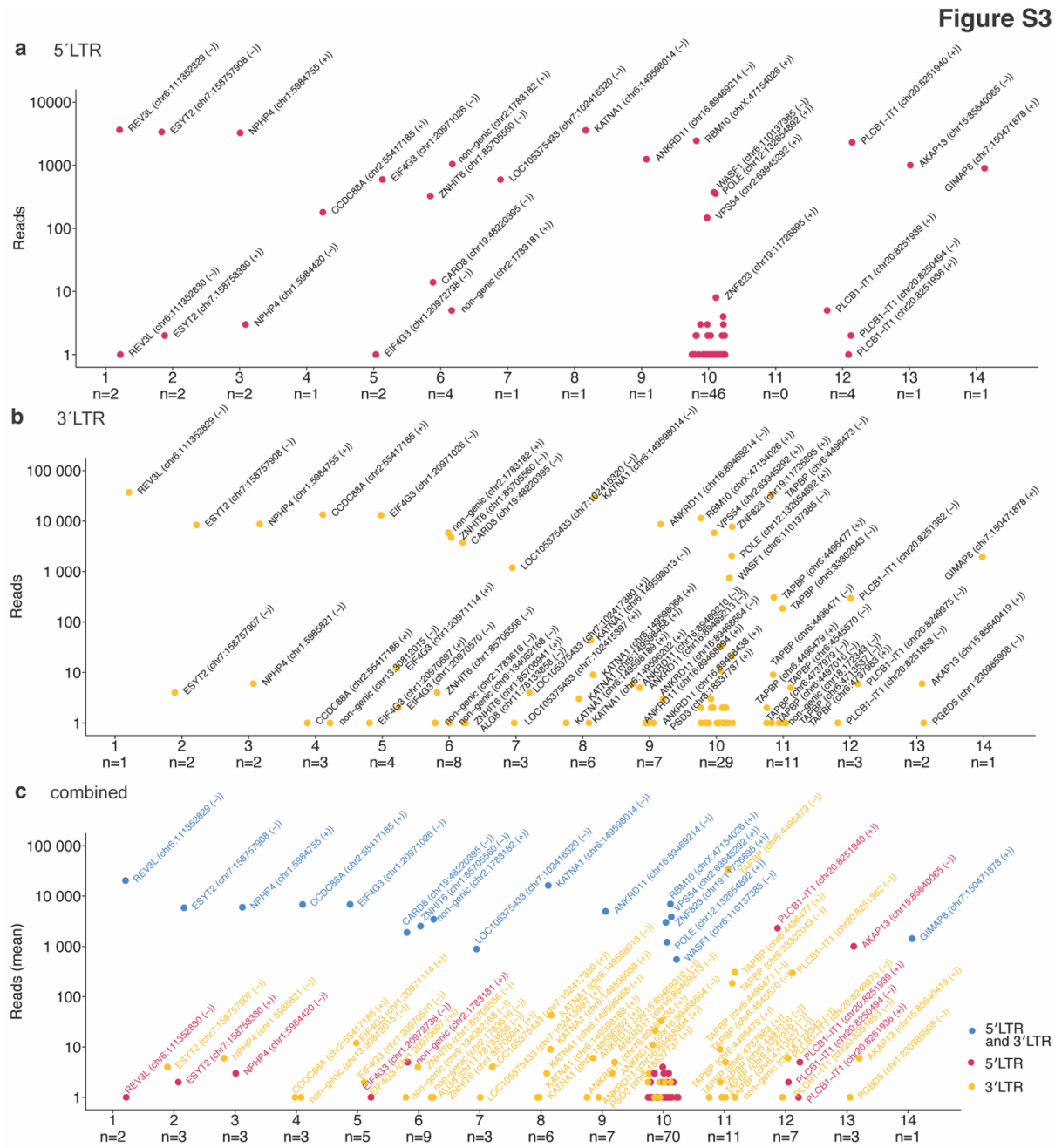


Figure S3. Identification of proviral integration sites using PRISM-seq in putative single-cell clones #1-14, raw data. (a-c) The proviral integration sites retrieved from individual putative clones, either from the 5′LTR (a) or 3′LTR (b) or combined (c) viral host junction. The blue, red, and yellow dots represent the integration sites identified by both 5′LTR and 3′LTR, exclusively by 5′LTR or 3′LTR, respectively. LTR; long terminal repeat of lentiviruses.

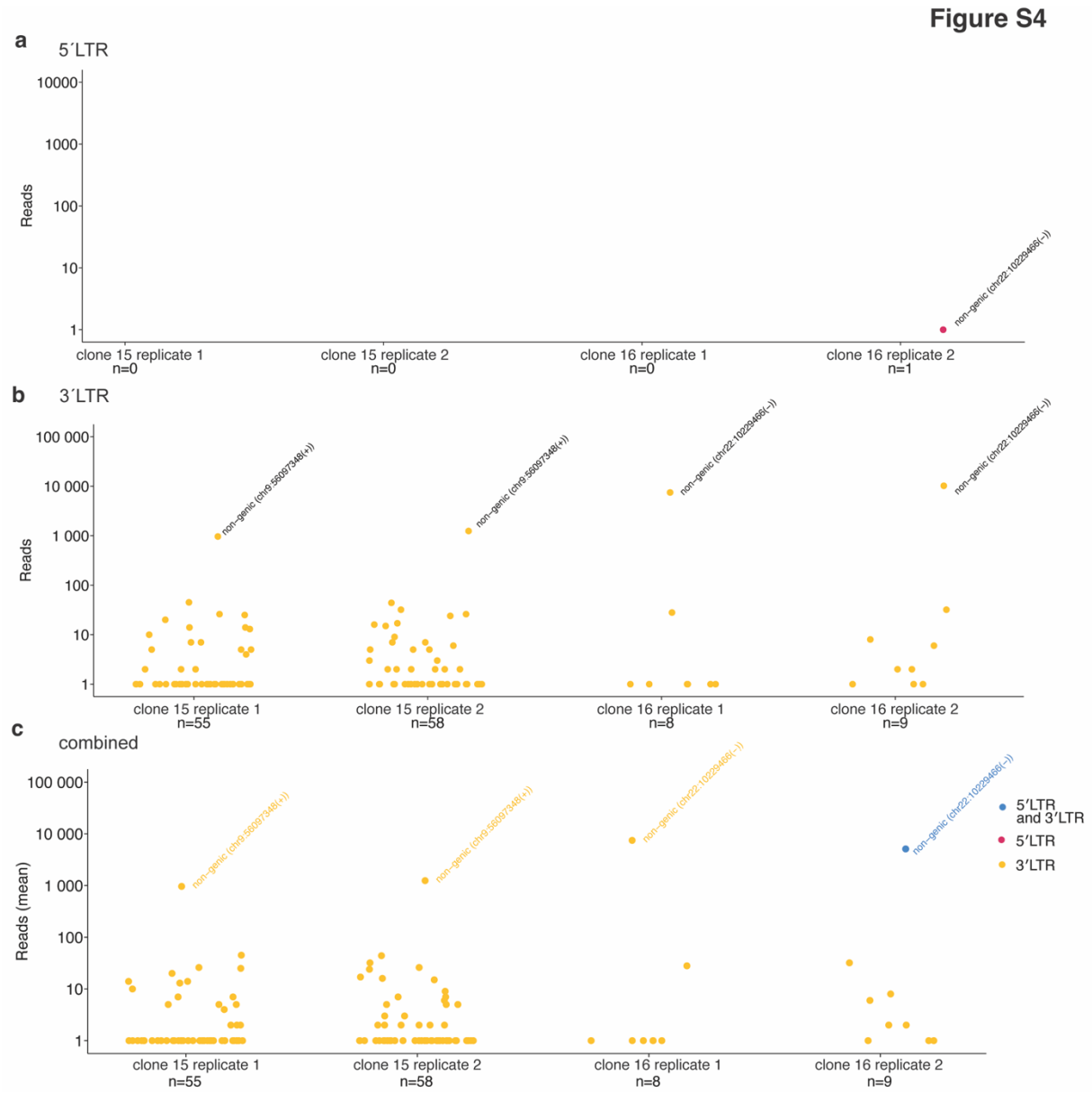


Figure S4. Identification of proviral integration sites using PRISM-seq in clones with known centromeric integrations #15 and #16, raw data. (a-c) The proviral integration sites were identified in clones #15 and #16, either from the 5′LTR (a) or 3′LTR (b) or combined (c) viral host junction using PRISM-seq. Results for clones #15 and #16 are shown for both technical duplicate reactions.

Figure S5

a

Number of cases	Types of staggers	examples
1	Integration sites 1 base pair apart (1 bp stagger)	REV3L (chr6:111352829 in negative strand) and REV3L (chr6:111352830 in negative strand)
2	Integration sites 1-10 base pair apart (1-10 bp stagger)	PLCB1-IT1 (chr20:8251940 in positive strand) and PLCB1-IT1 (chr20: 8251936 in positive strand)
3	Integration sites 10-100 bp pair apart (10-100 bp stagger)	EIF4G3 (chr1:20971026 in negative strand and EIF4G3 (chr1:20971114 in positive strand)
4	Integration sites 100 - 3000 bp apart (100-3000 bp stagger)	KATNA1(chr6: 149598014 in negative strand and KATNA1 (chr6: 149596202 in positive strand)

b

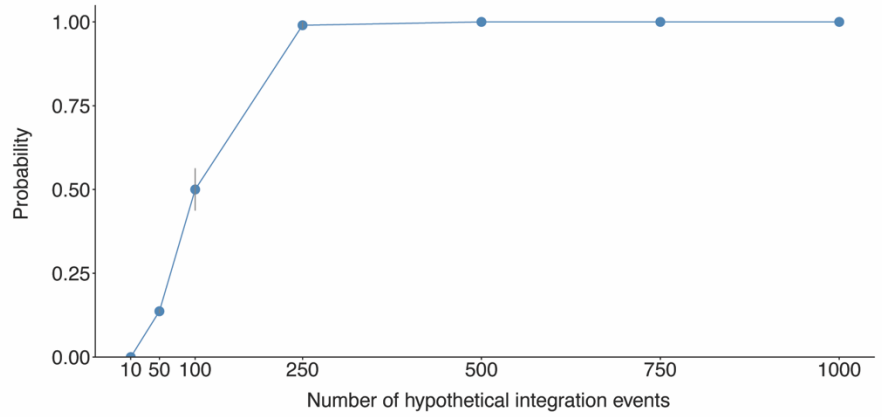
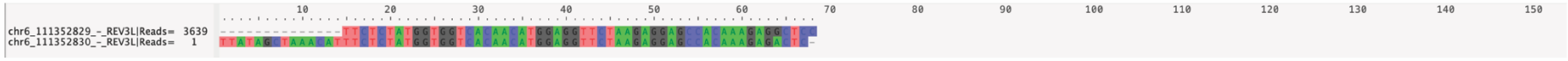


Figure S5. Identification of look-alike integration site artifact cases and the relationship between integration events and the probability of having true multiple integrations into the same gene(s). (a) Characterization of artifacts generated during the integration sites assay. Each staggered case is distinguished based on its genomic distance (bp) from the high-confidence integration sites coordinates of a putative clone. The examples shown here were observed in distinct putative clones. Bp: base pairs. **(b)** Regarding Step 3 in our proposed quality filter which involves collapsing integration sites that mapped to the same gene if found within a single reaction: We recognize that when PRISM-seq is applied to non-clonal clinical samples, there will be a genuine possibility that proviruses had integrated into distinct locations within the same gene. The likelihood of detecting genuine gene integration depends on the number of integration events sampled within a PRISM-seq reaction. We performed a Monte Carlo simulation and confirmed that the likelihood of detecting genuine same-gene integration was 0%, 0%, 14%, 50% and 99% when 1, 10, 50, 100, and 250 unique integration events were sampled. Therefore, we suggest that users should quantify targeting viral DNA copies by techniques such as Droplet Digital PCR (ddPCR), and turn off Step 3 if more than 50 target copies were input into each PRISM-seq reaction, with the caveat that specificity would be compromised. In this study, all reactions were performed at 50 copies per reaction input.

Figure S6

1_5´LTR

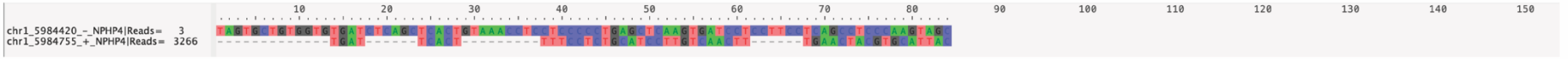


1_3´LTR

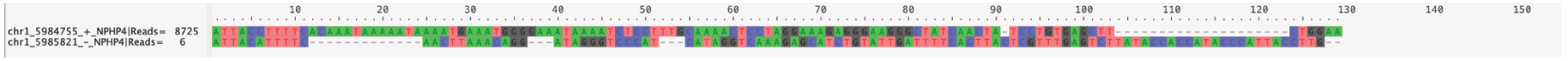
2_5´LTR

2_3´LTR

3_5´LTR



3_3´LTR



4_5´LTR

4_3´LTR

5_5´LTR

5_3´LTR

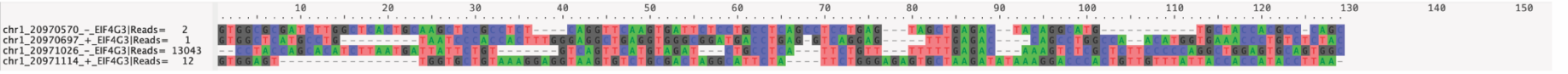
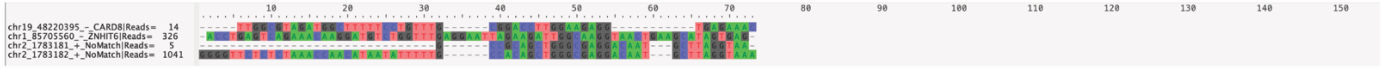


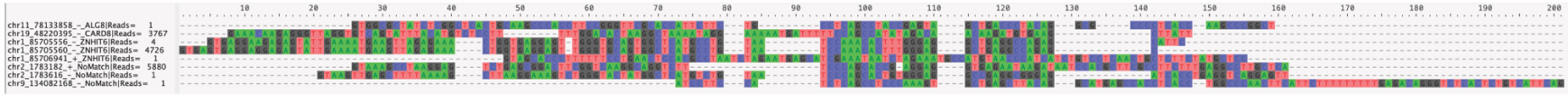
Figure S6. Comparison of host sequences for each unique integration site found within the putative clones. The host sequence derived from the viral-host chimeric junction read for individual integration sites retrieved from putative clones 1-5 by the 5′LTR and 3′LTR viral-host junction. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clones 1, 2, and 4.

Figure S7

6_5´LTR

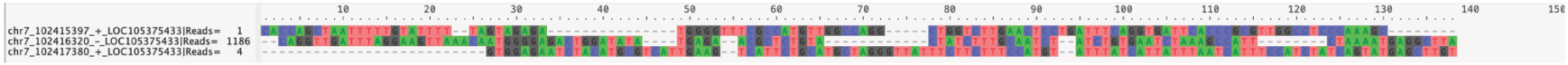


6_3´LTR



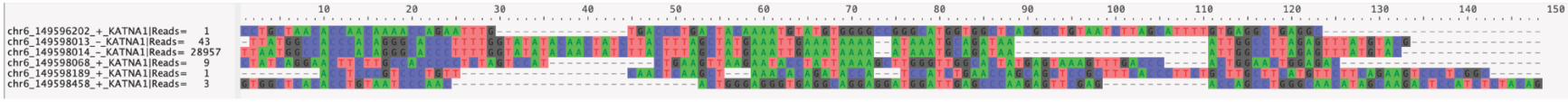
7_5´LTR

7_3´LTR



8_5´LTR

8_3´LTR



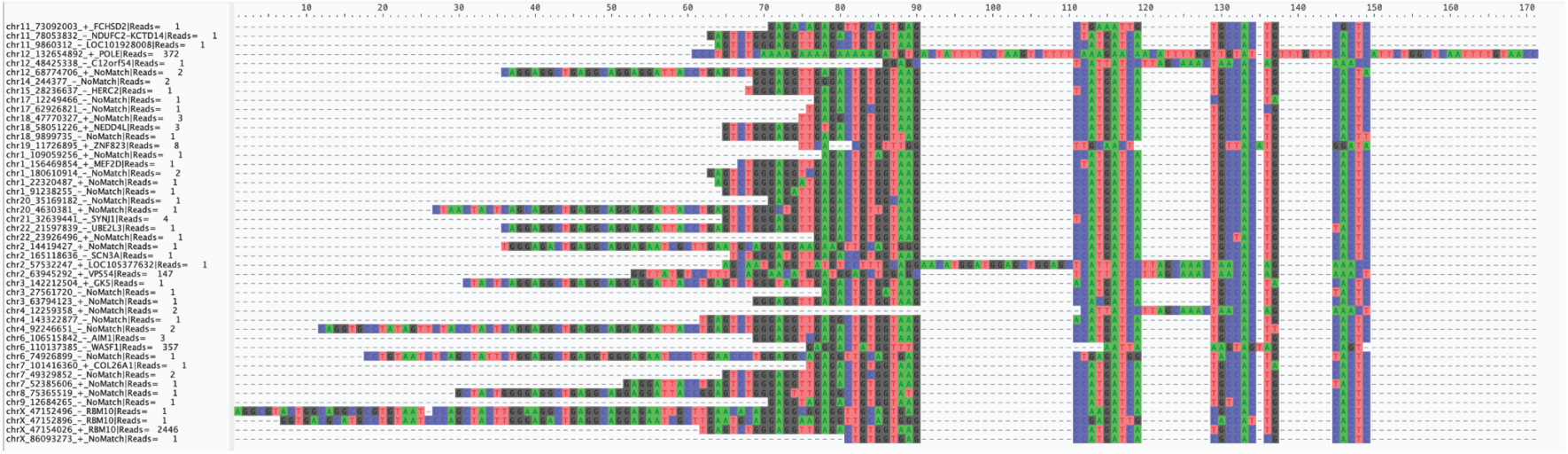
9_5´LTR

9_3´LTR

Figure S7. Comparison of host sequences for each unique integration site found within the putative clones. The host sequence derived from the viral-host chimeric junction read for individual integration sites retrieved from putative clones 6-9 by the 5′LTR and 3′LTR viral-host junction. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clones 6, 8, and 9.

Figure S8

10_5´LTR



10_3´LTR

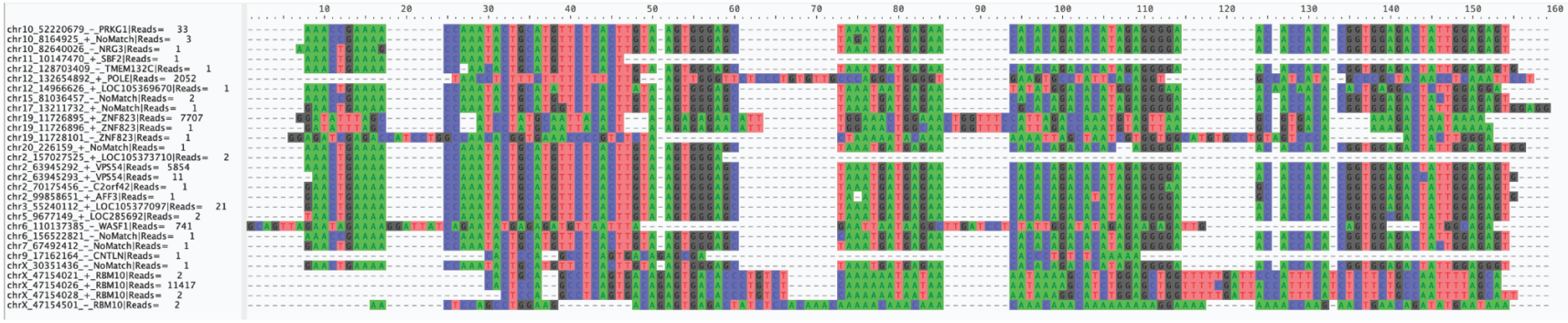
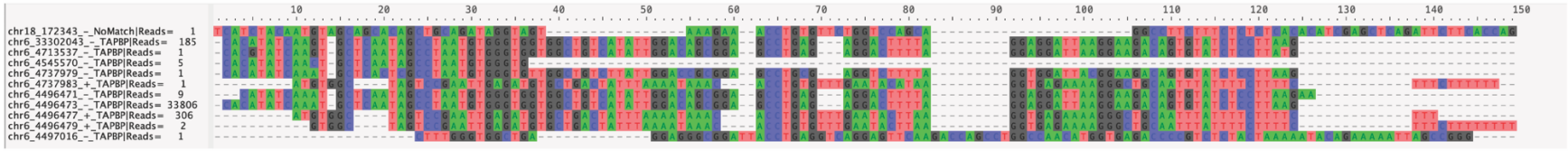


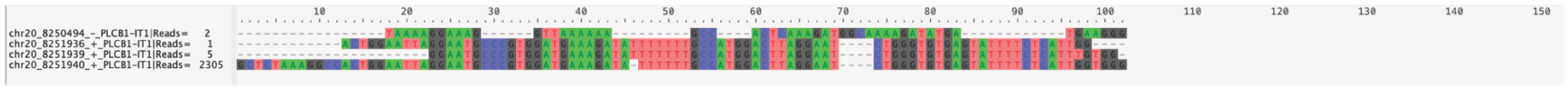
Figure S8. Comparison of host sequences for each unique integration site found within the putative clones. The host sequence derived from the viral-host chimeric junction read for individual integration sites retrieved from putative clone 10 by the 5′LTR and 3′LTR viral-host junction. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clone 10.

Figure S9

11_3´LTR



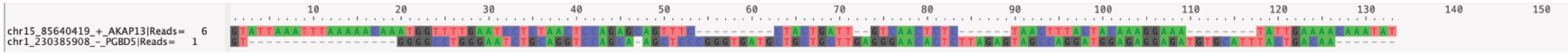
12_5´LTR



12_3´LTR

13_5´LTR

13_3´LTR



14_5´LTR

14_3´LTR

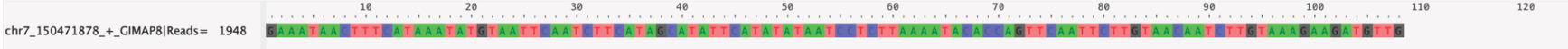


Figure S9. Comparison of host sequences for each unique integration site found within the putative clones. The host sequence derived from the viral-host chimeric junction read for individual integration site retrieved from putative clones 11-14 by the 5′LTR and 3′LTR viral-host junction. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clones 11 and 12.

Figure S10

clone 15 replicate 1_3´LTR

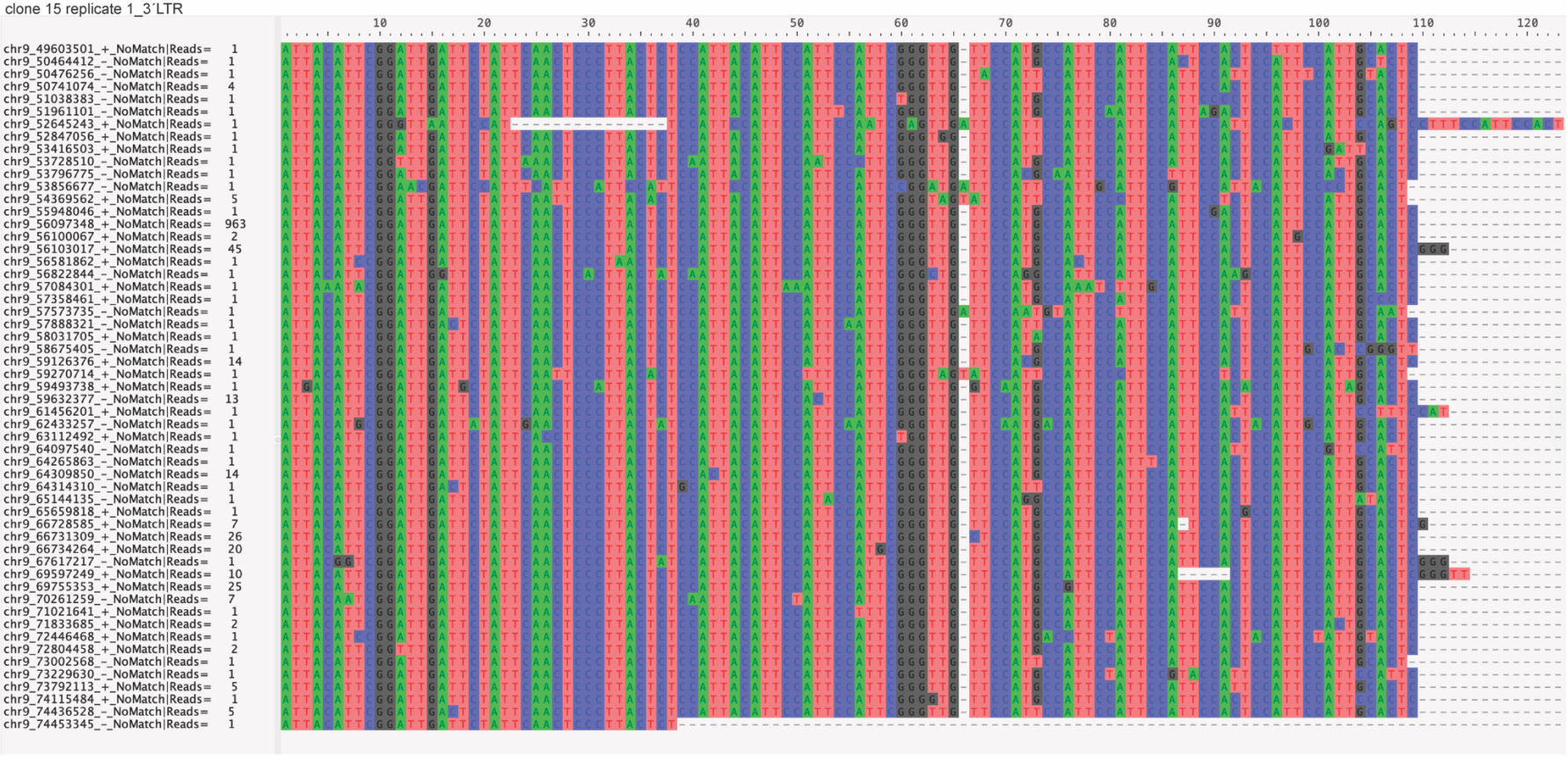


Figure S10. Comparison of host sequences for each unique integration site found in clones isolated from HIV-1-infected individuals. The host sequence derived from the 3′ LTR viral-host chimeric junction read for all unique integration sites identified for clone #15 replicate 1. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clone 15.

Figure S11

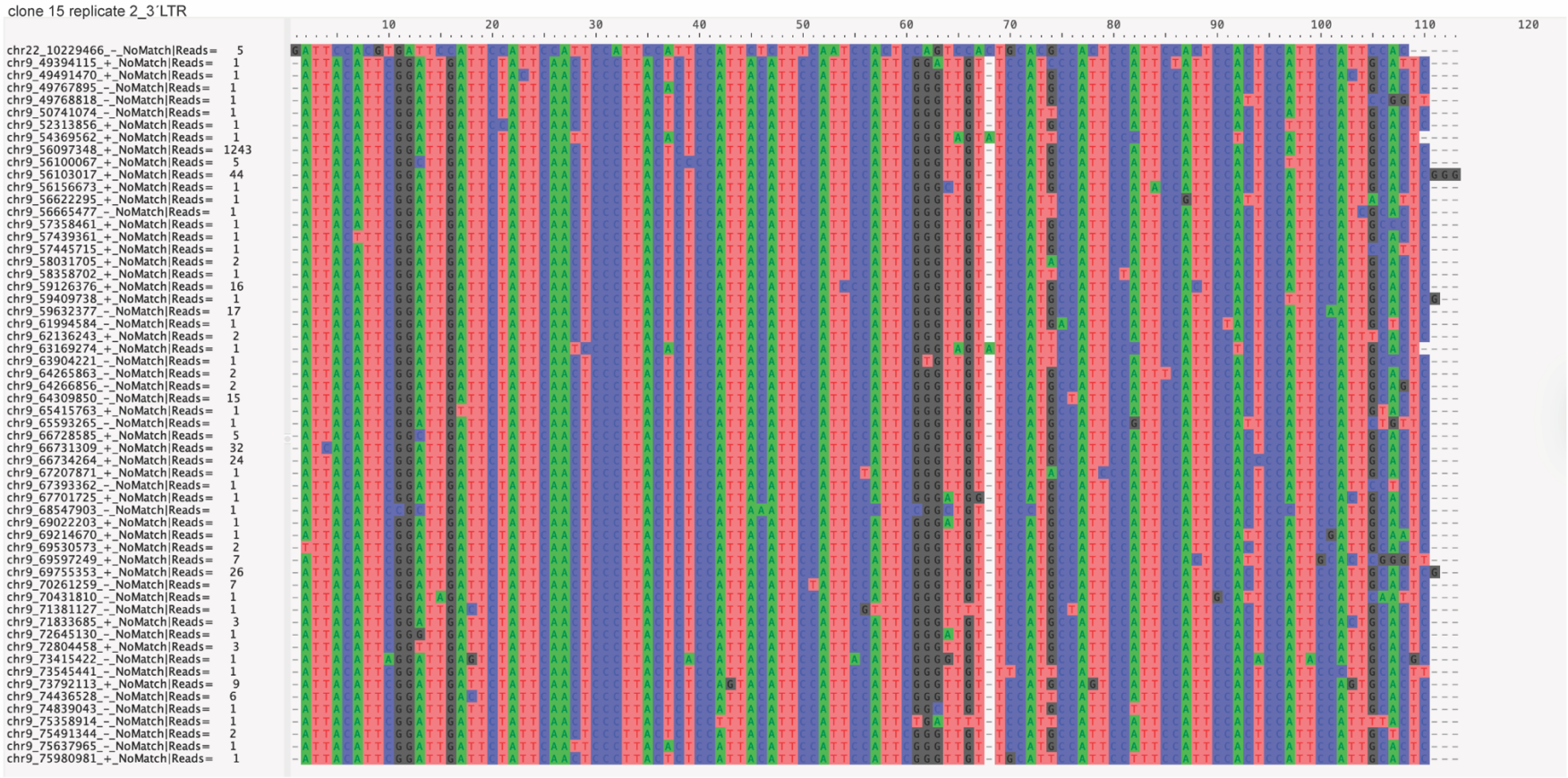
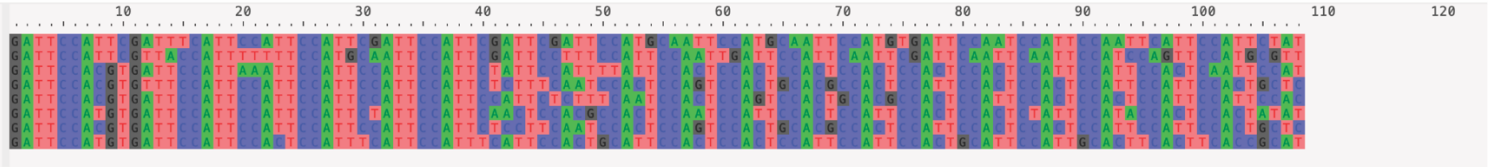

clone 15 replicate 2_3´LTR
10
20
30
40
50
60
70
80
90
100
110
120
chr22_10229466_-_NoMatch|Reads= 5
chr9_49394115_+_NoMatch|Reads= 1
chr9_49491470_+_NoMatch|Reads= 1
chr9_49767895_-_NoMatch|Reads= 1
chr9_49768818_-_NoMatch|Reads= 1
chr9_50741074_-_NoMatch|Reads= 1
chr9_52313856_+_NoMatch|Reads= 1
chr9_54369562_+_NoMatch|Reads= 1
chr9_56097348_+_NoMatch|Reads= 1243
chr9_56100067_+_NoMatch|Reads= 5
chr9_56103017_+_NoMatch|Reads= 44
chr9_56156673_+_NoMatch|Reads= 1
chr9_56622295_+_NoMatch|Reads= 1
chr9_56665477_-_NoMatch|Reads= 1
chr9_57358461_+_NoMatch|Reads= 1
chr9_57439361_+_NoMatch|Reads= 1
chr9_57445715_+_NoMatch|Reads= 1
chr9_58031705_+_NoMatch|Reads= 2
chr9_58358702_+_NoMatch|Reads= 1
chr9_59126376_+_NoMatch|Reads= 16
chr9_59409738_+_NoMatch|Reads= 1
chr9_59632377_-_NoMatch|Reads= 17
chr9_61994584_-_NoMatch|Reads= 1
chr9_62136243_+_NoMatch|Reads= 2
chr9_63169274_+_NoMatch|Reads= 1
chr9_63904221_-_NoMatch|Reads= 1
chr9_64265863_-_NoMatch|Reads= 2
chr9_64266856_-_NoMatch|Reads= 2
chr9_64309850_-_NoMatch|Reads= 15
chr9_65415763_+_NoMatch|Reads= 1
chr9_65593265_-_NoMatch|Reads= 1
chr9_66728585_+_NoMatch|Reads= 5
chr9_66731309_+_NoMatch|Reads= 32
chr9_66734264_+_NoMatch|Reads= 24
chr9_67207871_+_NoMatch|Reads= 1
chr9_67393362_-_NoMatch|Reads= 1
chr9_67701725_+_NoMatch|Reads= 1
chr9_68547903_-_NoMatch|Reads= 1
chr9_69022203_+_NoMatch|Reads= 1
chr9_69214670_+_NoMatch|Reads= 1
chr9_69530573_+_NoMatch|Reads= 2
chr9_69597249_+_NoMatch|Reads= 7
chr9_69755353_+_NoMatch|Reads= 26
chr9_70261259_-_NoMatch|Reads= 7
chr9_70431810_-_NoMatch|Reads= 1
chr9_71381127_-_NoMatch|Reads= 1
chr9_71833685_+_NoMatch|Reads= 3
chr9_72645130_-_NoMatch|Reads= 1
chr9_72804458_+_NoMatch|Reads= 3
chr9_73415422_-_NoMatch|Reads= 1
chr9_73545441_-_NoMatch|Reads= 1
chr9_73792113_+_NoMatch|Reads= 9
chr9_74436528_-_NoMatch|Reads= 6
chr9_74839043_-_NoMatch|Reads= 1
chr9_75358914_-_NoMatch|Reads= 1
chr9_75491344_-_NoMatch|Reads= 2
chr9_75637965_-_NoMatch|Reads= 1
chr9_75980981_+_NoMatch|Reads= 1

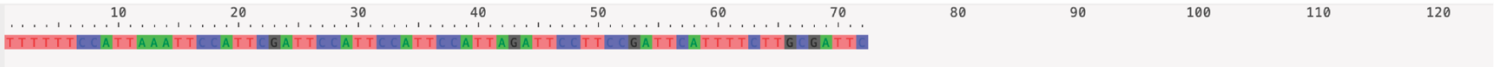
Figure S11. Comparison of host sequences for each unique integration site found in clones isolated from HIV-1-infected individuals. The host sequence derived from the 3′ LTR viral-host chimeric junction read for all unique integration sites identified for clone #15 replicate 2. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clone 15.

Figure S12

clone 16 replicate 1_3´LTR



clone 16 replicate 2_5´LTR



clone 16 replicate 2_3´LTR

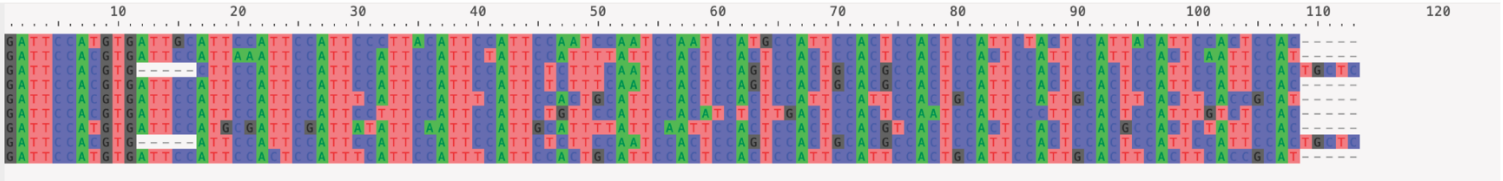


Figure S12. Comparison of host sequences for each unique integration site found in clones isolated from HIV-1-infected individuals. The host sequence derived from the viral-host chimeric junction read for all unique integration sites identified for clone #16 replicates 1 and 2, respectively. The low-confidence integration sites share the same junction sequence as the high-confidence integration site for clone 16.

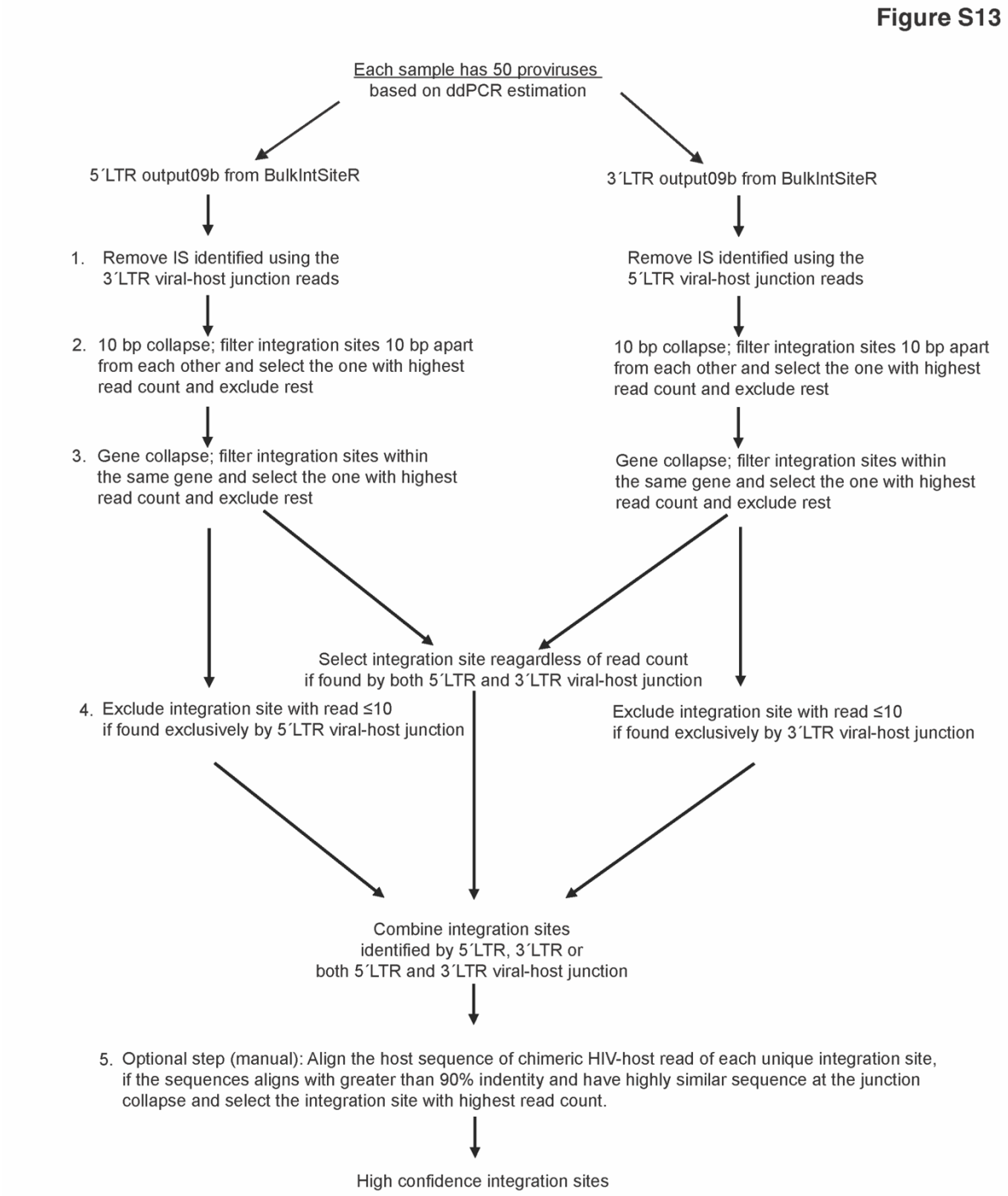


Figure S13. Data-guided noise removal approach. Leveraging the clonal nature of Clones #1-16, we devised a final quality control filter to remove assay noise. This set of quality filters is applied to the BulkIntSiteR output; (1) filter and remove the spillover noise of integration site identified by 3′ LTR in the 5′ LTR library and vice versa. (2) integration sites within ≤10 bp were collapsed by retaining the site with the highest read count; (3) integration sites located within the same gene were filtered to retain only one site with the highest read count, excluding all others if the input template count is ≤50 copies; (4) integration sites detected exclusively by either 5′ LTR or 3′ LTR junctions were excluded if supported by ≤10 read counts. Integration sites detected by

both 5′ LTR and 3′ LTR virus-host junctions, regardless of the raw read counts, and the sites that passed step 4 were classified as high-confidence. An optional Step (5), involving user-guided homology clustering of host sequences derived from the viral-host chimeric reads, removes look-alike integration sites that share identical junction sequences and co-occur within the same reaction, retaining only the site with the highest read count (Fig. S9). Steps 1-4 are fully automated, and the accompanying script is included with the BulkIntSiteR package; Step 5 requires user input and manual curation. If Step 5 is not performed, PRISM-seq results in 5.7% (10/174) noise in our data. This final step 5 is particularly critical when proviruses are expected to be integrated within repeat-rich genomic regions. Without Step 5, BulkIntSiteR identified 7 of 8 (87.5%) and 1 of 2 (50%) integration sites as low-confidence (noise) for the two clones (#15 and #16) with provirus integrated in centromeric regions (Fig. S14a-e). Incorporating step 5 and removing redundant low-confidence sites sharing the same junction sequence yielded one high-confidence integration site for each clone (0% noise), underscoring its importance for accurate proviral insertion site detection in repetitive loci.

Figure S14

a
Reads (mean)
100 000
10 000
1 000
100
10
1
clone 15 replicate 1
n=8
clone 15 replicate 2
n=8
clone 16 replicate 1
n=2
clone 16 replicate 2
n=2
5′LTR and 3′LTR
5′LTR
3′LTR
b
clone 15 replicate 1_3´LTR
chr9_56097348_+_NoMatch|Reads= 963
chr9_56103017_+_NoMatch|Reads= 45
chr9_59126376_+_NoMatch|Reads= 14
chr9_59632377_-_NoMatch|Reads= 13
chr9_64309850_-_NoMatch|Reads= 14
chr9_66731309_+_NoMatch|Reads= 26
chr9_66734264_+_NoMatch|Reads= 20
chr9_69755353_+_NoMatch|Reads= 25
c
clone 15 replicate 2_3´LTR
chr9_56097348_+_NoMatch|Reads= 1243
chr9_56103017_+_NoMatch|Reads= 44
chr9_59126376_+_NoMatch|Reads= 16
chr9_59632377_-_NoMatch|Reads= 17
chr9_64309850_-_NoMatch|Reads= 15
chr9_66731309_+_NoMatch|Reads= 32
chr9_66734264_+_NoMatch|Reads= 24
chr9_69755353_+_NoMatch|Reads= 26
d
clone 16 replicate 1_3´LTR
chr22_10229466_-_NoMatch|Reads= 7478
chr22_8860560_-_NoMatch|Reads= 28
e
clone 16 replicate 2_5´LTR
chr22_10229466_-_NoMatch|Reads= 10208
clone 16 replicate 2_3´LTR
chr22_10229466_-_NoMatch|Reads= 10208
chr22_8860560_-_NoMatch|Reads= 32

Figure S14. Post-filter recovery of proviral integration sites found within centromeric clones. **(a)** The BulkIntSiteR integration sites output data was processed using the QC filtering pipeline, and results for clones #15 and #16 are shown for technical duplicate reactions. **(b-e)** Comparison of the host sequence for all unique integration sites after QC filtering for clones #15 and #16. The unexpected low-confidence integration sites share the same junction sequence as the high-confidence integration site for clones 15 and 16.

Figure S15

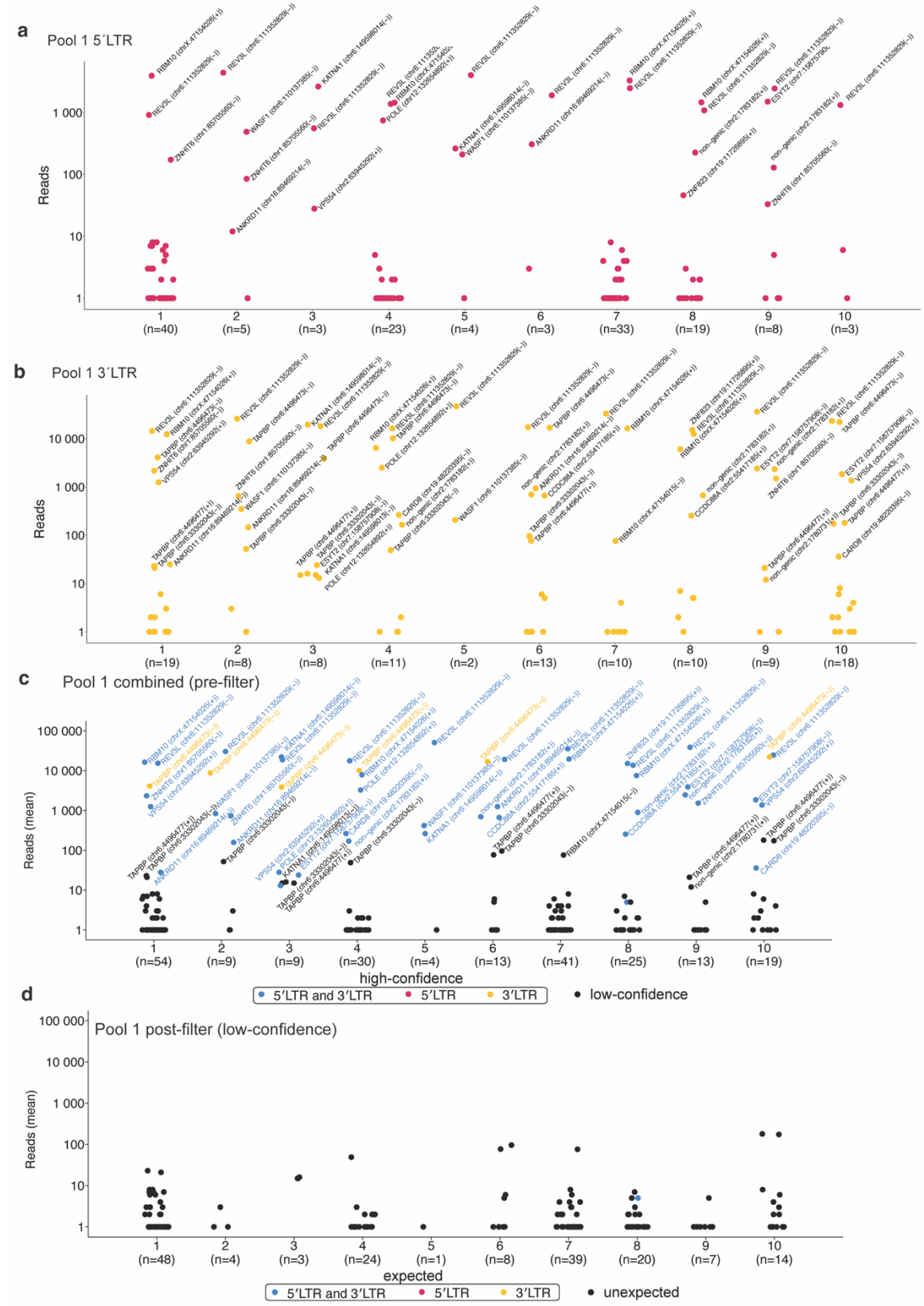

a
Pool 1 5´LTR
Reads
1 (n=40)
2 (n=5)
3 (n=3)
4 (n=23)
5 (n=4)
6 (n=3)
7 (n=33)
8 (n=19)
9 (n=8)
10 (n=3)
b
Pool 1 3´LTR
Reads
1 (n=19)
2 (n=8)
3 (n=8)
4 (n=11)
5 (n=2)
6 (n=13)
7 (n=10)
8 (n=10)
9 (n=9)
10 (n=18)
c
Pool 1 combined (pre-filter)
Reads (mean)
1 (n=54)
2 (n=9)
3 (n=9)
4 (n=30)
5 (n=4)
6 (n=13)
7 (n=41)
8 (n=25)
9 (n=13)
10 (n=19)
high-confidence
5´LTR and 3´LTR
5´LTR
3´LTR
low-confidence
d
Pool 1 post-filter (low-confidence)
Reads (mean)
1 (n=48)
2 (n=4)
3 (n=3)
4 (n=24)
5 (n=1)
6 (n=8)
7 (n=39)
8 (n=20)
9 (n=7)
10 (n=14)
expected
5´LTR and 3´LTR
5´LTR
3´LTR
unexpected

Figure S15. Proviral integration sites identification in artificial pool 1. (a-c) The dot plot represents the total number of unique integration sites identified in each replicate of pool 1 by either the 5′LTR (a) or 3′LTR (b) or combined (c) viral-host junction. The total number of unique integration sites retrieved in an individual replicate is represented as an n value. The blue, red, and yellow dots represent the integration site identified by both 5′LTR and 3′LTR, exclusively by 5′LTR or 3′LTR, respectively. (d) Post-filter identification of low-confidence expected and unexpected integration sites.

Figure S16

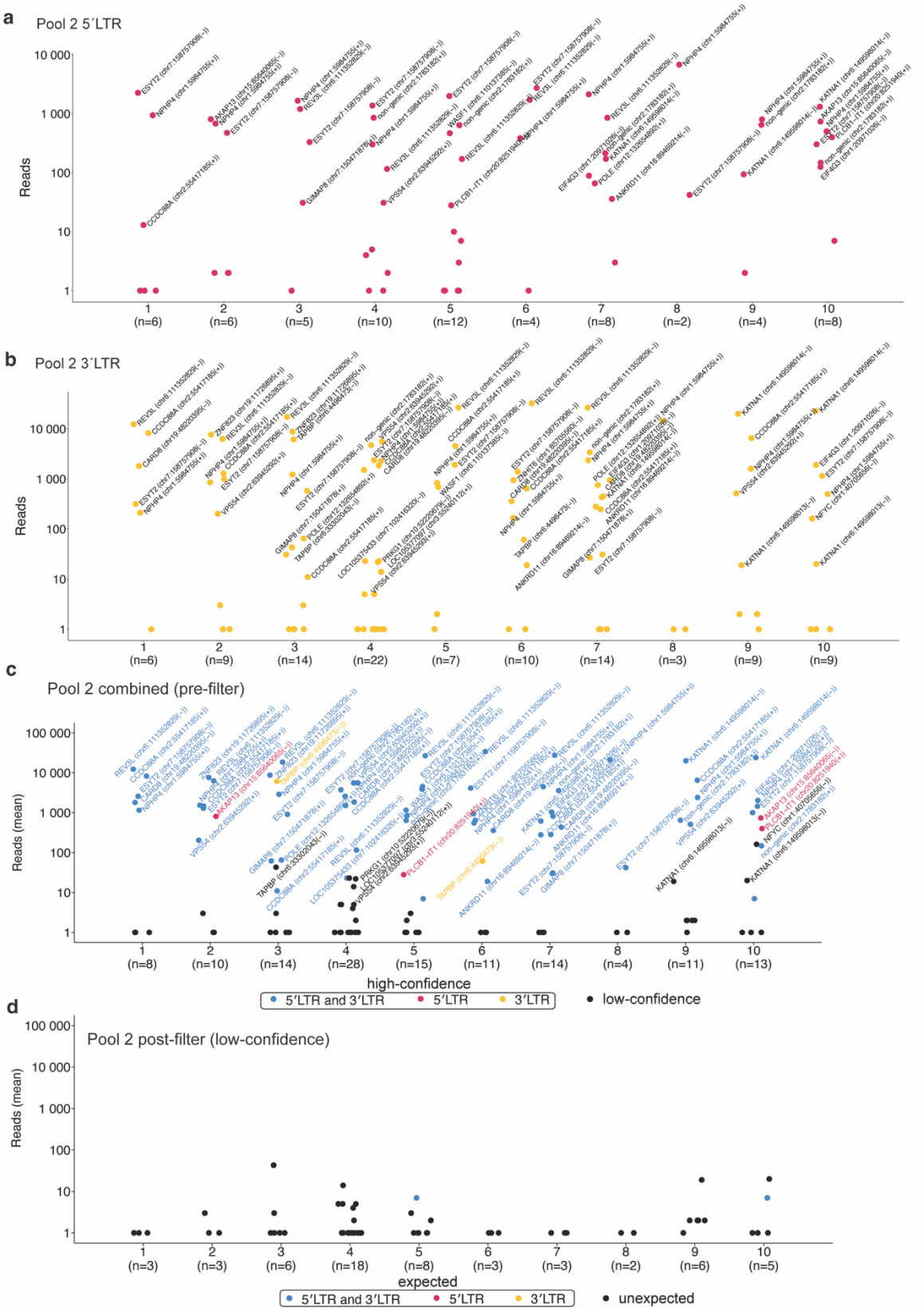

a
Pool 2 5´LTR
Reads
10 000
1 000
100
10
1
1 (n=6)
2 (n=6)
3 (n=5)
4 (n=10)
5 (n=12)
6 (n=4)
7 (n=8)
8 (n=2)
9 (n=4)
10 (n=8)
b
Pool 2 3´LTR
Reads
10 000
1 000
100
10
1
1 (n=6)
2 (n=9)
3 (n=14)
4 (n=22)
5 (n=7)
6 (n=10)
7 (n=14)
8 (n=3)
9 (n=9)
10 (n=9)
c
Pool 2 combined (pre-filter)
Reads (mean)
100 000
10 000
1 000
100
10
1
1 (n=8)
2 (n=10)
3 (n=14)
4 (n=28)
5 (n=15)
6 (n=11)
7 (n=14)
8 (n=4)
9 (n=11)
10 (n=13)
high-confidence
5´LTR and 3´LTR
5´LTR
3´LTR
low-confidence
d
Pool 2 post-filter (low-confidence)
Reads (mean)
100 000
10 000
1 000
100
10
1
1 (n=3)
2 (n=3)
3 (n=6)
4 (n=18)
5 (n=8)
6 (n=3)
7 (n=3)
8 (n=2)
9 (n=6)
10 (n=5)
expected
5´LTR and 3´LTR
5´LTR
3´LTR
unexpected

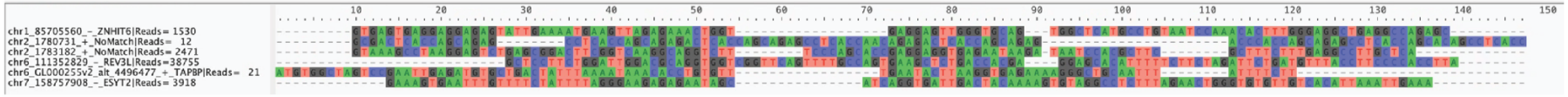
Figure S16. Proviral integration sites identification in artificial pool 2. (a-c) The dot plot represents the total number of unique integration sites identified in each replicate of pool 2 by either the 5′LTR (a) or 3′LTR (b) or combined (c) viral-host junction. The total number of unique integration sites retrieved in an individual replicate is represented as an n value. The blue, red, and yellow dots represent the integration sites identified by both 5′LTR and 3′LTR, exclusively by 5′LTR or 3′LTR, respectively. (d) Post-filter identification of low-confidence expected and unexpected integration sites.

Figure S17

a

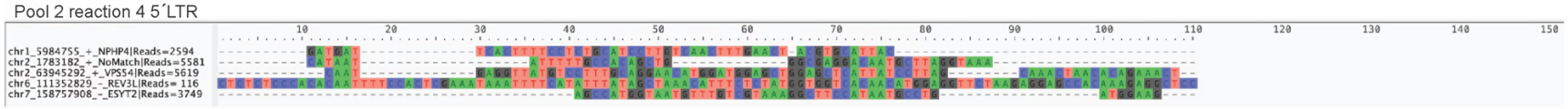
Pool 1 reaction 9 5´LTR

Pool 1 reaction 9 3´LTR



b

Pool 2 reaction 4 5´LTR



Pool 2 reaction 4 3´LTR

Figure S17. Comparison of host sequences of unexpected integration sites observed in pool 1 and pool 2, respectively (a and b). The host sequence derived from unexpected integration sites found in pool 1 replicate 9 (a) and pool 2 replicate 4 (b) by 5′LTR and 3′LTR viral-host junction. The unexpected integration sites sharing the same junction sequence or ±10 kb away from the expected high-confidence integration site are labelled with an asterisk (*).