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

Virender Kumar Pal[1], Marie Canis[1], Emily Stone[2], Nathan L. Board[2], Klara Lenart[3], Marcilio Jorge Fumagalli[3], Colin Kovas[4], R. Brad Jones[2], Michel Nussenzweig[3,6], Frauke Muecksch[5], Paul D. Bieniasz[1,6], Guinevere Q. Lee[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): 147

Current word count (main text, excluding methods): 3171

Abstract

Retroviral integration into host genomes is central to both HIV-1 persistence and the safety and function of lentiviral vectors used in gene and cell therapies. However, 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-derived 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 underlies 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 (gDNA) 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 high due to multi-step library construction and sequencing depth requirements.

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

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 as a defined starting point for systematic assay characterization; however, the workflow can accommodate substantially higher input levels (~10,000 proviral templates pre-MDA[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, and the reads obtained were processed using our open-source computational pipeline, BulkIntSiteR, which was developed for the 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]. BulkIntSiteR normalizes junction coordinates to the 5′ LTR convention, accounting for the characteristic 4-bp offset between the

98 5′ and 3′ LTR junctions[29]. Using PRISM-seq, we identified a single viral integration site in
99 8E5/LAV cells on chromosome 13 at position 67,485,907 in the negative strand, consistent
100 with published data[28]. We did not observe any low-abundance secondary integration sites
101 (noise).

102

103 The single provirus in 8E5/LAV cells is associated with heterochromatin (Supplementary Table
104 1). However, viral integration can occur in diverse regions of the human genome, including
105 actively transcribing genes[30], and in repeat-rich, telomeric, and centromeric satellite
106 regions[3,6,31,32]. We generated 14 putative singly-infected Jurkat cell clones (Clones #1-14) using
107 low-MOI (0.3) infection with a replication-deficient HIV-1 reporter (V1/SBP-GFP), followed
108 by sorting of GFP/SBP-positive cells, limiting dilution, and expansion via culturing. Proviral
109 integration sites in these 14 putative clones were expected to be mostly in genes since HIV-1
110 integration favors actively transcribing genomic regions[30]. To evaluate PRISM-seq
111 performance within centromeric regions typically associated with difficult-to-map satellite
112 DNA repeats, we analyzed two additional HIV-1-infected primary cell clones (#15 and #16)
113 previously found to each contain a single provirus integrated within centromeric regions[32].

114

115 Without noise removal, our raw data showed 1-70 (min-max) unique proviral insertion sites in
116 clones #1-16 (Fig. S3a-c and S4a-c). As each putative clone is expected to harbor a single
117 proviral integration site, the detection of multiple unique integration sites per clone may
118 indicate assay-associated noise, which is consistent with the presence of noise reported in other
119 integration site sequencing assays[20]. We therefore used this clonal sample set as an internal
120 benchmark to systematically evaluate assay noise.

121

122 An initial quality-control filter designated each of our raw integration sites as high- or low-
123 confidence. High-confidence was defined if an integration site (i) contained precise viral DNA
124 5′ or 3′ ends linked to a human sequence without gaps/missing bases, and (ii) was detected by
125 both 5′ LTR and 3′ LTR viral-host junctions or (iii) had greater than 10 reads when identified
126 only by either 5′ LTR or 3′ LTR. All other integration sites were designated "low-confidence".

127

128 Among clones #1-14, 9/14 (64%) yielded a single high-confidence integration site (Fig. S3c).
129 Two of the 14 clones, #6 and #10, showed multiple high-confidence integration sites, consistent
130 with expectations from Poisson statistics for infections at MOI 0.3[33]. Clones #8, #11, and #12
131 had more than one high-confidence integration sites, and these integrations were found in the
132 same gene for the respective clone, which we explored in-depth in the next section for potential
133 assay-associated artifacts. We used two human reference genomes: hg38 and T2T-CHM13[34],
134 and their results were 100% concordant for clones #1-14. For clones 15 and 16, our data showed
135 multiple high-confidence integration sites with proviruses integrated within the centromeric
136 repeat regions of chromosomes 9 (8 sites) and 22 (2 sites), respectively (Fig. S4c), which were
137 only detected using T2T-CHM13 as the reference genome, while hg38 returned no maps. Due
138 to the clonal nature of these samples, we suspected repeat-region-driven ambiguity in mapping.
139 All ambiguities were further explored in the next section.

140

141 Across all 16 analyzed clones, high-confidence proviral integration sites were distributed
142 within 12 chromosomes spanning genic, intergenic, and repeat-rich centromeric satellite
143 regions, confirmed by the chromHMM model[35] based on distinct histone-mark ChIP-seq
144 datasets from the ENCODE database[36] (Supplementary Table 1). The average distance between
145 proviral integration sites and the nearest accessible chromatin region (ATAC-seq peak) varied

146 widely, ranging from 292 bp to 278,541 bp (Supplementary Table 1). These findings reflect
147 the diverse genomic landscape of proviral integration recovered by PRISM-seq.

148

149 **Implementation of additional quality control filtering criteria**

150 In 15 of 16 analyzed clones, we detected 174 low-confidence secondary integration sites in
151 addition to the described high-confidence integration sites. We next characterized this putative
152 noise. Of these low-confidence integration sites, 57/174 (33%) mapped to the same gene or
153 within ±10 kb relative to the corresponding high-confidence integration sites (Fig. S3a-c, S4a-
154 c, S5a-b, and Supplementary Table 2). The remaining low-confidence sites were either in
155 intergenic regions or in distinct genes. Given the clonal nature of most of the samples, these
156 sites may constitute artifacts. All junction sequences were subjected to manual examination as
157 described below.

158

159 To further investigate these potential artificats, we performed manual alignment of the host
160 sequences of low-confidence integration sites, and showed that many differed from the
161 corresponding high-confidence site of the clone only by short indels or 1-10 bp shifts in the
162 junction sequence (e.g., clones 1, 2, 4, 8, 9, 10, 11, 13, 15, and 16) (Fig. S6-S12). These cases
163 are likely artifact signatures previously highlighted to be found in integration site sequencing
164 assays, caused mainly by PCR recombination, whole-genome amplification artifacts,
165 sequencing errors, or ambiguous mapping in repeat-rich chromosomal regions[20,37,38].
166 Collectively, the short indel cases, same gene integration in a putative clone, and low read-
167 count (≤10) when captured by only one proviral end accounted for 174/174 (100%) of the low-
168 confidence integration sites detected among clones #1-16 (Supplementary Table 2). Together,

these observations demonstrated that low-confidence integration sites arose from predictable, assay-associated artifacts, providing a rational basis for systematic quality-control filtering.

We therefore implemented a five-step quality control filter strategy to improve the specificity of PRISM-seq (Fig. S13). Steps 1-4 are fully automated, and the accompanying script is included with the BulkIntSiteR package; Step 5 requires manual curation. In this study, steps 1-4 removed 94% (164/174) and step 5 removed an additional 6% (10/174) of the noise in our data. Step 5 is especially critical when proviruses are integrated within repeat-rich genomic regions: without it, 7 of 8 (88%) and 1 of 2 (50%) integration sites were called as high-confidence sites for the two clones (#15 and #16) with centromeric proviral insertion site (Fig. S14a-e). Incorporating step 5 yielded a single high-confidence integration site for each clone. All integration sites that pass the final filter are designated as "high confidence integration sites" and are used for downstream analysis.

PRISM-seq detection efficiency and cumulative recovery in samples containing multiple integration sites

Next, we examined whether PRISM-seq can accurately detect integration sites in bulk input samples with more than one unique integration site, each with varying abundance. Variation in template input copies, as well as multiple authentic proviral insertions, can limit detection efficiency and accuracy due to PCR competition and recombination[12,20,39]. To model this scenario, we generated two artificial gDNA pools composed of clones #1-14 (Pool 1 and Pool 2), each with integration sites represented at different relative input frequencies, and used them to assess the detection efficiency of PRISM-seq in a mixed template context (Fig. 2a-b).

193 We first estimated the number of molecules containing viral integration sites for each clone
194 based on gDNA input mass, assuming a clonal population structure and using an established
195 approximation of ~150,000 diploid human cells per microgram of genomic DNA[40]. Using these
196 estimates, we constructed two master pools: Pool 1 consisted of ~400 copies of clone #1
197 combined with ten additional clones, each represented at ~10 copies, whereas Pool 2 consisted
198 of ~100 copies each of clones #1-4 combined with ten additional clones, each represented at
199 ~10 copies. These two mixes were each distributed into 10 replicate reactions. Therefore,
200 effectively, Pool 1 contained clone #1 at an average of ~40 copies per reaction, while clones
201 #2-11 were present at approximately single-copy input per reaction, whereas Pool 2 contained
202 clones #1-4 at ~10 copies per reaction and clones #5-14 at approximately single-copy input per
203 reaction. The automated quality-control filter steps 1-4 outlined in the previous section were
204 applied to all results. Note that this experimental design represents a limiting-dilution strategy,
205 such that the actual number of template molecules distributed into each reaction followed
206 Poisson statistics, with mean (λ, lamda) values of 40, 10, or 1 copy per reaction.

207

208 With this 10-replicate setup, the abundant 40- and 10-copy integration sites were detected with
209 100% efficiency in both pools (Fig. 2a-b and S15-17). As for single-copy integration site
210 templates, each replicate reaction contained an average of one HIV template. Under a Poisson
211 sampling model, the probability that a given well received at least one template molecule is
212 63%; however, when multiple replicate reactions were performed, the cumulative probability
213 of observing at least one positive well increased with the number of replicates. Therefore,
214 detection efficiency for single-copy integration sites was evaluated in the context of replicate-
215 level cumulative recovery rather than single-well sampling probability. After excluding clones
216 with multiple integration sites (clones #6 and #10), three single-copy integration sites (clones
217 #3, #5, and #7) were not detected in Pool 1 across all replicate reactions (Fig. 2a). Under a

218 Poisson sampling model, assuming an average of 10 template copies per 10 cumulative
219 reactions ($\lambda = 10$), the probability of observing zero molecules in a given set of 10-replicate is
220 <0.1%, corresponding to a >99.9% probability that at least one molecule is present. This
221 indicates that stochastic loss at the level of individual reactions is unlikely to explain the
222 observed lack of signal in Pool 1. Therefore, the absence of these clones in Pool 1 is most
223 consistent with either (i) their under-representation or absence during pool construction, or (ii)
224 clone-specific amplification or detection failure. In contrast, all three clones #3, #5, and #7
225 were also input as a single template per reaction in Pool 2 but were robustly detected (Fig. 2b),
226 ruling out integration-site-specific amplification or detection limitations and strongly
227 implicating upstream sampling during Pool 1 preparation as the primary cause of their loss.
228 Consequently, detection efficiency calculations for rare (single-copy) integration sites were
229 based on Pool 2 data. Using replicate-level cumulative recovery and correcting for Poisson
230 sampling expectations, 75% of single-copy integration sites were detected by six replicates,
231 with complete (100%) recovery achieved by seven replicates, despite the presence of abundant
232 competitor templates (Fig. 2c). Overall, these findings demonstrate that PRISM-seq can
233 reliably recover integration sites in a sample expected to contain both rare and abundant
234 templates.

235

236 **Estimation of the extent of clonal expansion using PRISM-seq**

237 Repeated detection of the same integration site across pre-MDA replicates signifies that the
238 original sample contains multiple copies of a given viral-host junction (Fig. 3a). In a clinical
239 context, such an observation would indicate clonal expansion of infected cells. We reasoned
240 that the frequency with which an integration site appears across PRISM-seq replicates would
241 be an indirect quantitative measure of the relative size of the clone harboring that integration.

242 As proof of principle, we applied this framework to the post-filtered datasets for pools
243 containing multiple integration sites. For each integration site, we counted the number of
244 replicate reactions in which it was recovered. Integration sites detected in multiple replicates
245 are highly likely to represent true biological events, whereas those observed only once are more
246 likely to reflect residual noise. As anticipated, the dominant clones, present at ~400 copies in
247 Pool 1 and ~100 copies in Pool 2, were detected in the largest number of replicate reactions
248 and therefore exhibited the highest inferred clonal abundance (Fig. 3b-3e). In Pool 1, the high-
249 copy integration site was recovered in all 10 replicate wells, corresponding to 20% of the total
250 relative clonal abundance, whereas lower-copy integration sites were detected in fewer wells
251 (1-6 wells; 2-12% clonal abundance) (Fig. 3b and 3d). Similarly, in Pool 2, high-copy
252 integration sites were recovered in 7-10 of 10 replicate wells, accounting for 10-14% of the
253 total relative clonal abundance, while low-copy integration sites were detected in fewer
254 replicates (1-5 wells; 1-7% clonal abundance) (Fig. 3c and 3e).

255

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

257

258 To demonstrate the clinical utility of PRISM-seq, we longitudinally sampled $CD4^+$ T cells from
259 a study participant living with HIV-1 before antiretroviral therapy initiation (pre-ART), and 3
260 months, 1 year, and 7 years post-ART (Supplementary Table 3). Each sample was expected to
261 contain multiple unique integration sites originating from infected cells exhibiting variable
262 levels of clonal expansion[41-44]. Using 2-5 replicates each containing 50 proviral copies, and
263 thus only sampling 100-250 HIV-containing templates per time point, we recovered 69, 67, 11,
264 and 17 high-confidence integration sites from each time points, respectively (Fig. 4a). The
265 proportion of proviruses integrated into genic or intergenic regions remained stable over time

266 (Fig. 4b). The fraction of proviruses integrated into ZNF genes increased from 2%-3% (pre-
267 ART and 3 months) to 9% (1 year) and 18% (7 years) post-ART time points (Fig. 4c), consistent
268 with previous reports indicating that proviruses persisting in ZNF genes and transcriptionally
269 repressive genomic contexts are more likely to be transcriptionally silent and capable of long-
270 term persistence[6,31].

271

272 Considering that repeated detection of the same integration site in multiple replicates indicates
273 a higher chance of clonal expansion of the respective cell clone, we used this metric across the
274 longitudinal samples to track persistent clones: we observed identical viral integration sites in
275 BUB1B-PAK6(chr15:40226639(+)) and intergenic region (chr13:105139021(+)) across all 4
276 time points sampled over 7 years (Fig. 4d), an indicator that HIV-1 infected cells with provirus
277 integrations at these loci were expanded *in vivo* and persisted during treatment over time in this
278 study participant. Overall, our findings show that PRISM-seq could be directly applied to
279 clinical samples that had an extremely low target concentration, and provided a wealth of
280 integration site data and biological insights with minimal sample input.

281

282 **Discussion**

283

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

287

288 Existing bulk integration site assays, including assays INSPIIRED[11,12], VISPA/VISPA2[13],
289 QuickMap[14], VISA[15], MAVRIC[16], GeIST[17], IS-Seq[18], VSeq-Toolkit[19], were not validated for
290 single-molecule detection and/or broad performance across challenging genomic contexts.
291 PRISM-seq overcomes these limitations through its MDA-enabled single-molecule capture
292 workflow, which efficiently recovers rare viral-host junctions from minimal DNA input.
293 Although MDA has been used previously in MIP-seq[5] that approach requires limiting dilution,
294 it is costly, and is not designed for bulk integration site analysis. PRISM-seq provides an
295 economic, bulk-input compatible, sensitive, and scalable platform for high-confidence
296 integration site profiling.

297

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

311

312 Application of PRISM-seq to longitudinal samples from a study participant with HIV-1
313 revealed persistent clones integrated within transcriptionally repressive, ZNF-rich genomic
314 regions. These findings reinforce prior observations that proviruses can integrate and persist in
315 epigenetically repressed environments, including KRAB-ZNF gene clusters and other
316 heterochromatic regions[2,6,31,45]. The compatibility of PRISM-seq with low sample input and its
317 ability to capture low-abundance, genomically diverse proviruses directly from clinically-
318 derived material highlights its utility for high-resolution reservoir characterization.
319 Furthermore, our companion study[46] applied PRISM-seq to flow-cytometry-sorted cells
320 derived from a novel mouse model of HIV latency, demonstrating both that proviral integration
321 site is a deterministic factor in proviral latency fate decisions in the absence of external
322 immunological or anti-viral drug selection pressure and that PRISM-seq is compatible with
323 post-sort samples.

324

325 PRISM-seq has several limitations. First, clonal expansion estimates currently require replicate
326 reactions and yield relative, rather than absolute, quantification. Second, increasing input
327 concentration beyond ≥50 templates per reaction is possible[22] but means users should turn off
328 the same-gene filter, as the frequency of true same-gene integrations within a replicate will be
329 higher with increasing template concentrations (step 3 in Fig. S9 and see Monte Carlo
330 simulation Fig S5b). This will result in a modest compromise to specificity. Third, while our
331 quality-filtering framework effectively reduces artifacts, it may occasionally exclude rare
332 authentic sites detected by only one proviral end with low read counts - manual check is always
333 recommended. Finally, PRISM-seq provides precise integration sites mapping but does not
334 assess proviral genome intactness, which remains specific to HIV-1 reservoir studies and
335 should be complemented with assays that evaluate proviral sequence integrity[5,47,48].

336

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

342

343 **Method**

344

345 **Viral-host junction amplification and sequencing**

346

347 Genomic DNA was extracted (QIAGEN DNeasy kit), quantified for HIV-1 *gag* by droplet
348 digital PCR (Bio-Rad), subjected to whole-genome amplification using multiple displacement
349 amplification (MDA; QIAGEN REPLI-g Single Cell Kit) at 50 copies per reaction, and
350 processed using an adaptor-ligation PCR workflow adapted from the Clontech Lenti-X
351 Integration Site Analysis Kit[5]. Final nested PCR was used to selectively enrich viral-host
352 junction products and sequenced on an Illumina MiSeq platform (2 × 150 bp). Full protocol
353 details, origin of PRISM-seq, reagent specifications, and cost are provided in the Extended
354 Methods. Technical replicates were defined as independent sampling of each DNA extraction
355 (pre-MDA); 10 replicates were used for pooling experiments, 2-5 for clinical samples, and
356 none for 8E5/LAV or clones #1-16.

357

358 **Identification of HIV-1 integration site using BulkIntSiteR**

359

360 Illumina sequencing-derived FASTQ files were processed using BulkIntSiteR, a one-click R-
361 based, fully automated pipeline designed to detect, map, and annotate HIV-1 integration sites
362 in the human genome. BulkIntSiteR is cross-platform and is not restricted to Illumina data. All
363 scripts are available freely on GitHub (https://github.com/guineverelee/BulkIntSiteR/).
364 Briefly, each small read was independently assessed for the presence of high-quality viral ends
365 (containing HXB2 1 or 9719 using BLAST+ suite[49]). For each read positively identified to
366 contain a viral end, the non-viral portion of the read was mapped to the human genome using
367 BLAT[50]. Additionally, we observed a large disparity in junction-containing reads between the
368 5′ and 3′ LTR integration sites libraries (e.g., for the 8E5/LAV benchmarking experiment),
369 likely reflecting the differences in amplicon composition, relative to Illumina tagmentation and
370 read length. The 5′ LTR amplicon contains ~681 nt of viral sequence, whereas the 3′ LTR
371 amplicon contains only ~34 nt. Because tagmentation generates fragments of ~300 bp on
372 average and the sequencing read length was 150 bp, most 5′ LTR reads fall entirely within the
373 viral portion of the amplicon and never reach the host sequence, which would markedly reduce
374 the number of recoverable viral-host chimeras. See the GitHub BulkIntSiteR documentation
375 for the full algorithm.
376
377 **Generation of 14 putative clones with a single integrated provirus**
378 To generate the V1/SBP-GFP HIV-1 reporter virus, 7.5×10^7 293T cells (ATCC #CRL-3216;
379 RRID: CVCL_0063) were seeded in 875 cm^2 5-layer flasks 1 day prior to transfection. 293T
380 cells were transfected with 75 µg pV1/SBP-GFP, 75 µg pCRV1/NLGagPol, and 15 µg pVSV-
381 G (Yee et al., 1994) using polyethyleneimine (PEI). After 12 hours, the medium was replaced
382 with fresh growth medium. Supernatants were collected 48 hours post-transfection, filtered,
383 and concentrated by ultracentrifugation over a 20% sucrose cushion. The pellet containing the
384 virus was resuspended in PBS + 10% BSA, aliquoted, and stored at -80 °C. The viral stock titre

385 (IU/ml) was determined by infecting MT4-LTR-GFP cells and using flow cytometry to assess
386 GFP expression.
387
388 For infection, Jurkat and MOLT-4 cells were infected with V1/SBP-GFP[46] at a multiplicity of
389 infection (MOI) of 0.3. Infected cells were enriched using streptavidin bead-based magnetic
390 selection. Single-cell clones of infected Jurkat and MOLT-4 cells were isolated by limiting
391 dilution by diluting to 0.5 cells per well in RPMI growth medium, and 100 µL of the suspension
392 was seeded into each well of 96-well plates. The clonal outgrowth was monitored
393 microscopically, and wells containing single-cell-derived colonies were transferred to fresh
394 wells for expansion.
395
396 **Clinical samples**
397 The study participant with HIV-1 was recruited at the Maple Leaf Clinic in Toronto, Canada.
398 PBMC samples were used according to protocols approved by the respective Institutional
399 Review Boards. $CD4^+$ T cells were enriched from total PBMCs using a $CD4^+$ T Cell Isolation
400 Kit (STEMCELL Technologies, catalog #17952). Clinical and demographic characteristics of
401 the study participant are summarized in Supplemental Table 3.
402
403
404
405
406
407
408
409

410 **Extended methods (Supplementary)**

411

412 **Cost and origin**

413 PRISM-seq is an evolution of our previously published matched integration site and proviral
414 sequencing (MIP-seq) protocol[5], which itself was adapted from the Clontech LentiX
415 integration site analysis method (catalog #631263). MIP-seq was designed to co-capture HIV
416 integration sites together with full-length proviral genomes at single-input-template resolution.
417 In contrast, PRISM-seq is optimized for both single and multiple-template input and focuses
418 specifically on capturing viral integration sites, enabling substantially higher throughput at
419 dramatically reduced cost. At the time of writing, PRISM-seq offers an economical solution at
420 $280 per sample with detection sensitivity down to a single template and supports up to ~10000
421 pre-MDA template input[22], making the assay broadly accessible. In addition, PRISM-seq is
422 fully supported by a publicly available bioinformatics (BulkIntSiteR) and quality control
423 pipeline, which automates integration site calling, annotation, and quality assessment.

424

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

426 DdPCR quantifications were performed to estimate the HIV-1 *gag* and *env*-containing proviral
427 template concentration[51]. $CD4^+$ T cells were enriched from total PBMCs using a $CD4^+$ T Cell
428 Isolation Kit (STEMCELL Technologies, catalog 17952) and subjected to DNA extraction
429 using commercial kits (QIAGEN DNeasy, 69504). We amplified total HIV-1 DNA using
430 ddPCR (Bio-Rad), with primers and probes previously described (127-bp 5′-LTR-gag
431 amplicon; HXB2 coordinates 684-810[43,52,53] and a modified version of the intact proviral DNA
432 assay (IPDA)[51]we published previously[54]. The droplets were subsequently read by a QX600
433 droplet reader, and data were analyzed using Quanta-Soft software (BIO-RAD).

434

435 **Multiple Displacement Amplification**

436 Genomic DNA was isolated using commercial kits (QIAGEN DNeasy Blood & Tissue Kit,
437 catalog 69504), according to the manufacturer's instructions. Then, the genomic DNA was
438 amplified using multiple displacement amplification (MDA) with phi29 polymerase (QIAGEN
439 REPLI-g Single Cell Kit, catalog #150345). Buffers DLB, D1 (denaturation buffer), and N1
440 (neutralization buffer) were prepared according to the manufacturer's protocol. The genomic
441 DNA (2.5 μL) was denatured by the addition of 2.5 μL of buffer D1, followed by an incubation
442 at room temperature for 3 mins. The mix was neutralized by adding 5 μL of buffer N1. 40 μL
443 of the MDA master mix, consisting of nuclease-free water, REPLI-g single cell Reaction
444 Buffer, and REPLI-g single cell DNA Polymerase, was added to the 10 μL denatured genomic
445 DNA sample. The final 50 μL reaction was incubated at 30°C for 4 h, followed by heat
446 inactivation at 65°C for 3 mins. The amplified product was purified using AMPure XP beads
447 (Beckman Coulter) and was used for downstream integration site analysis.

448

449 **Enzymatic digestion of genomic DNA**

450 This protocol was adapted and modified from the Clontech Lenti-X Integration Site Analysis
451 Kit (catalog #631263), which, in its standard form without MDA, lacks sufficient sensitivity to
452 detect single-copy proviral DNA in a sample. Purified MDA products were digested using three
453 blunt-end restriction enzymes - HpaI (NEB #R0105L), SspI (NEB #R3132L), and DraI (NEB
454 #R0129L) - to fragment genomic DNA and facilitate subsequent ligation-based integration site
455 analysis. Briefly, 5 μL of purified MDA product was added to a master mix containing
456 nuclease-free water, 10× restriction enzyme buffer, and either DraI or SspI (20 U μL^{-1}) or HpaI
457 (5 U μL^{-1}), to a final reaction volume of 50 μL. Reactions were incubated at 37 °C for 18 hours

458 to ensure complete digestion. The digested genomic DNA was then purified using AMPure XP
459 beads and carried forward to adaptor ligation.
460
461 **Ligation of the genome-walker adaptor to the digested genomic DNA**
462 The genomic DNA fragments are ligated with double-stranded T-linker DNA (genome walker)
463 (see section below for exact sequence), in which the shorter strand is an oligonucleotide with
464 a 5' end phosphorylated (to enable efficient linker ligation) and the 3' end is modified with an
465 amino modification, which limits the extension of the short strand during PCR in case of linker
466 self-ligation**.** Digested genomic DNA (4.8 μL) was ligated to adaptors in an 8 μL reaction
467 containing 1.9 μL adaptor (25 μM), 0.8 μL 10× ligation buffer, and 0.5 μL T4 DNA ligase (6
468 U μL^{-1}; NEB #M0202L). Ligation was performed at 16 °C overnight in a thermal cycler to
469 ensure temperature stability. The reaction was terminated by heat inactivation at 70 °C for 5
470 min, followed by dilution with 32 μL of DEPC-treated water (Ambion #4387937) to a final
471 volume of 40 μL.
472
473 **Viral-host junction PCR amplification**
474 The adaptor ligated genomic DNA fragments were used as a template for nested PCR
475 amplification of viral-host junctions. The 5′LTR HIV-1 junction was amplified using LSP1
476 (GCTTCAGCAAGCCGAGTCCTGCGTCGAG) and LSP2
477 (GCTCCTCTGGTTTCCCTTTCGCTTTCAA) as forward primers, both derived directly from
478 the CloneTech LentiX kit. The 3′LTR HIV-1 junction was amplified using LestralLTR1
479 (CTTAAGCCTCAATAAAGCTTGCCTTGAG) and LestralLTR2
480 (AGACCCTTTTAGTCAGTGTGGAAAATC)[5]. These were paired with adaptor-specific
481 reverse primers AP1 (GTAATACGACTCACTATAGGGC) and AP2
482 (ACTATAGGGCACGCGTGGT) for the first and second PCR reactions, respectively. The

483 AP1 primer is specific to the 5' single-stranded region of the T-linker to avoid unwanted
484 amplification of the host DNA. The adaptor sequence was adapted from the CloneTech LentiX
485 kit (catalog #631263).
486 Each 25 µL PCR reaction contained 1× reaction buffer, 1× dNTP mix, 0.2 µM of each primer,
487 and Advantage 2 polymerase (Clontech Advantage® 2 PCR Kit, Cat. #639206). For the first
488 PCR, 15 µL of master mix was combined with 10 µL of diluted adaptor-ligated genomic DNA,
489 using AP1/LSP1 for 5′LTR junctions and AP1/LestralLTR1[5] for 3′LTR junctions. Cycling
490 conditions were: 7 cycles of 94 °C for 25 s and 72 °C for 3 min, followed by 32 cycles of 94
491 °C for 25 s and 67 °C for 3 min, and a final extension at 67 °C for 7 min.
492 For the second (nested) PCR, 24 µL of master mix was combined with 1 µL of first-round PCR
493 product, using AP2/LSP2 for 5′LTR junctions and AP2/LestralLTR2[5] for 3′LTR junctions.
494 Cycling conditions were: 5 cycles of 94 °C for 25 s and 72 °C for 3 min, followed by 20 cycles
495 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
496 and 3′LTR viral-host junctions were successfully amplified under these conditions. The
497 resulting PCR products were subjected to Illumina MiSeq paired-end 150 base pairs (bp)
498 sequencing.
499
500 **Cell culture**
501 Jurkat cells and MOLT-4 cells were cultured at 37 °C and 5% CO_2 in RPMI supplemented with
502 10 % FCS and Gentamycin. All cell lines were monitored regularly by the SG-PERT assay[55]
503 to ensure the absence of retroviral contamination. Cell lines used in this study were devoid of
504 mycoplasma contamination.
505
506 **Availability**

BulkIntSiteR is available under an open-source license at:

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

Acknowledgements

We acknowledge the Massachusetts General Hospital Center for Computational and Integrative Biology DNA Core for Illumina sequencing support. This work was supported by NIH grants Research Enterprise to Advance a Cure for HIV (REACH) UM1AI164565 (RBJ), R21AI150398 (GQL), R01AI162221 (GQL), and the Center for the Structural Biology of HIV RNA U54 AI170660 (PB). We acknowledge Dr. Douglas Barrows, Bioinformatics Analyst, Bioinformatics Resource Center, The Rockefeller University, New York. The authors gratefully acknowledge the MGH DNA core facility, Massachusetts, USA. This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

References

1 Janssens, J., De Wit, F., Parveen, N. & Debyser, Z. Single-Cell Imaging Shows That the Transcriptional State of the HIV-1 Provirus and Its Reactivation Potential Depend on the Integration Site. *mBio* **13**, e00007–00022 (2022). https://doi.org/10.1128/mbio.00007-22

2 Teixeira, A. R. *et al.* Transcription of HIV-1 at sites of intact latent provirus integration. *Journal of Experimental Medicine* **221** (2024). https://doi.org/10.1084/jem.20240391

3 Jiang, C. *et al.* Distinct viral reservoirs in individuals with spontaneous control of HIV-1. *Nature* **585**, 261–267 (2020). https://doi.org/10.1038/s41586-020-2651-8

4 Jadlowsky, J. K. *et al.* Long-term safety of lentiviral or gammaretroviral gene-modified T cell therapies. *Nature Medicine* **31**, 1134–1144 (2025). https://doi.org/10.1038/s41591-024-03478-6

5 Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest* **129**, 988–998 (2019). https://doi.org/10.1172/jci124291

6 Huang, A. S. *et al.* Integration features of intact latent HIV-1 in CD4+ T cell clones contribute to viral persistence. *Journal of Experimental Medicine* **218** (2021). https://doi.org/10.1084/jem.20211427

7 Chun, T.-W. *et al.* Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences* **94**, 13193–13197 (1997). https://doi.org/doi:10.1073/pnas.94.24.13193

8 Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Medicine* **5**, 512–517 (1999). https://doi.org/10.1038/8394

9 Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997). https://doi.org/10.1126/science.278.5341.1291

10 Ronen, K. *et al.* Distribution of lentiviral vector integration sites in mice following therapeutic gene transfer to treat β-thalassemia. *Mol Ther* **19**, 1273–1286 (2011). https://doi.org/10.1038/mt.2011.20

11 Berry, C. C. *et al.* INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions. *Mol Ther Methods Clin Dev* **4**, 17–26 (2017). https://doi.org/10.1016/j.omtm.2016.11.003

12 Sherman, E. *et al.* INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes. *Mol Ther Methods Clin Dev* **4**, 39–49 (2017). https://doi.org/10.1016/j.omtm.2016.11.002

13 Calabria, A. *et al.* VISPA: a computational pipeline for the identification and analysis of genomic vector integration sites. *Genome Medicine* **6**, 67 (2014). https://doi.org/10.1186/s13073-014-0067-5

14 Appelt, J. U. *et al.* QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis. *Gene Ther* **16**, 885–893 (2009). https://doi.org/10.1038/gt.2009.37

15 Hocum, J. D. *et al.* VISA - Vector Integration Site Analysis server: a web-based 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 Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harb Perspect Med* **2**, a006890 (2012). https://doi.org/10.1101/cshperspect.a006890

30 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

31 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

32 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

33 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

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

35 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
36 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). https://doi.org/10.1038/nature14248
37 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
38 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
39 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
40 Christopherson, C. *et al.* PCR-Based assay to quantify human immunodeficiency virus type 1 DNA in peripheral blood mononuclear cells. *J Clin Microbiol* **38**, 630–634 (2000). https://doi.org/10.1128/jcm.38.2.630-634.2000
41 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
42 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
43 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
44 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
45 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
46 Pal, V. *et al.* HIV-1 Integration Site Determines the Transcriptional Fate and Persistence of Integrated Proviruses. *bioRxiv*, 2025.2012.2026.696579 (2025). https://doi.org/10.64898/2025.12.26.696579
47 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
48 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
49 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
50 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656–664 (2002). https://doi.org/10.1101/gr.229202
51 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
52 Malnati, M. S. *et al.* A universal real-time PCR assay for the quantification of group-M HIV-1 proviral load. *Nat Protoc* **3**, 1240–1248 (2008). https://doi.org/10.1038/nprot.2008.108

53 Hindson, B. J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* **83**, 8604–8610 (2011). https://doi.org/10.1021/ac202028g

54 Kinloch, N. N. *et al.* HIV-1 diversity considerations in the application of the Intact Proviral DNA Assay (IPDA). *Nature Communications* **12**, 165 (2021). https://doi.org/10.1038/s41467-020-20442-3

55 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

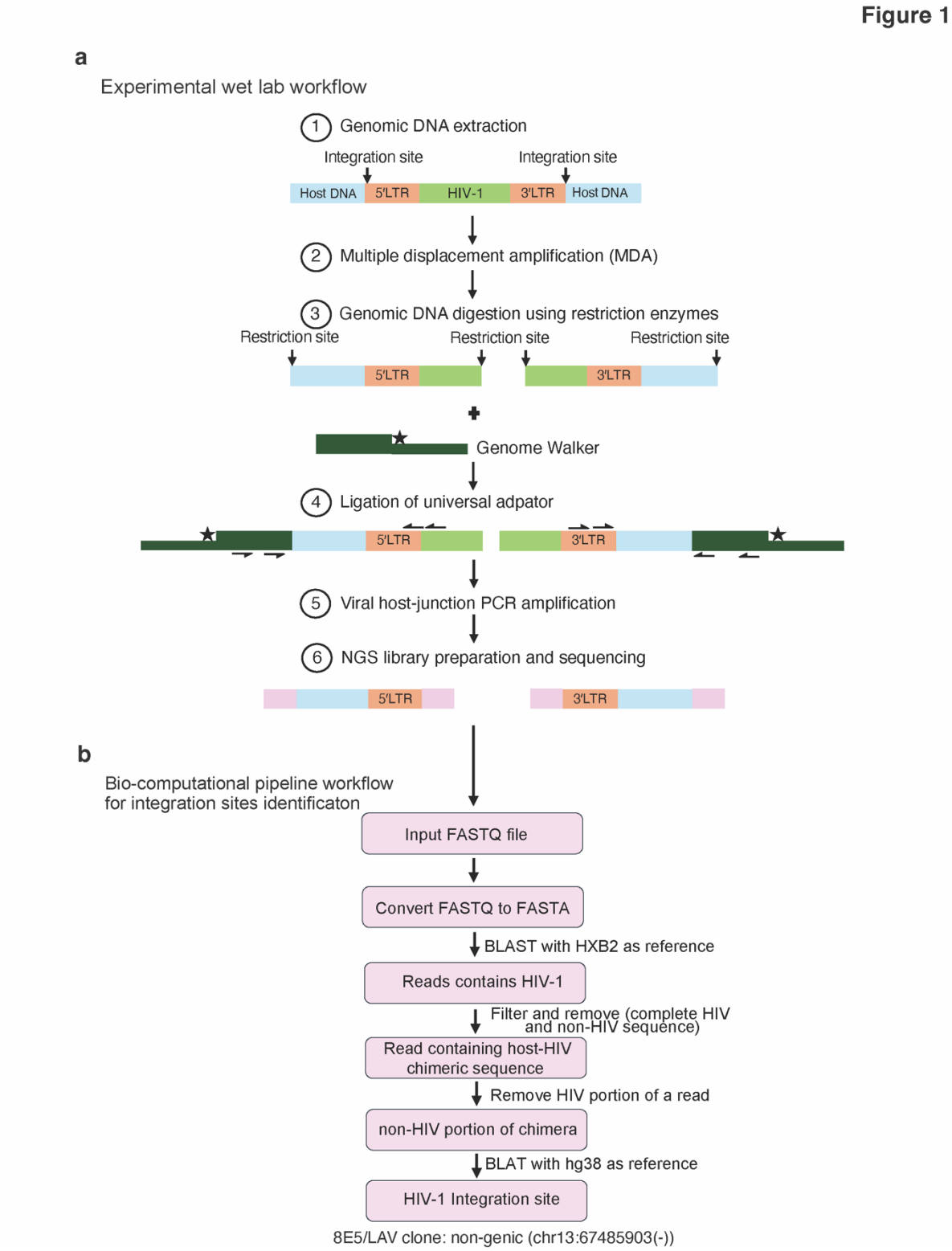


Figure 1. Experimental and bioinformatic workflow for proviral integration sites identification using PRISM-seq. (a) Experimental workflow of PRISM-seq. **(b)** Bioinformatics steps implemented in the BulkIntSiteR pipeline to process raw sequencing data and generate annotated proviral integration site outputs. *NGS, next-generation sequencing.*

Figure 2

a Pool 1 post-filter

Clone ID	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	CARD8 (chr19:48220395(-))	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	non-genic (chr2:1783182(+))	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	POLE (chr12:132654892(+))	1	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	RBM10 (chrX:47154026(+))	1	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
10	VPS54 (chr2:63945292(+))	1	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
10	WASF1 (chr6:110137385(-))	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

Clone ID	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	CARD8 (chr19:48220395(-))	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	non-genic (chr2:1783182(+))	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	POLE (chr12:132654892(+))	1	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
10	RBM10 (chrX:47154026(+))	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	VPS54 (chr2:63945292(+))	1	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
10	WASF1 (chr6:110137385(-))	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

c

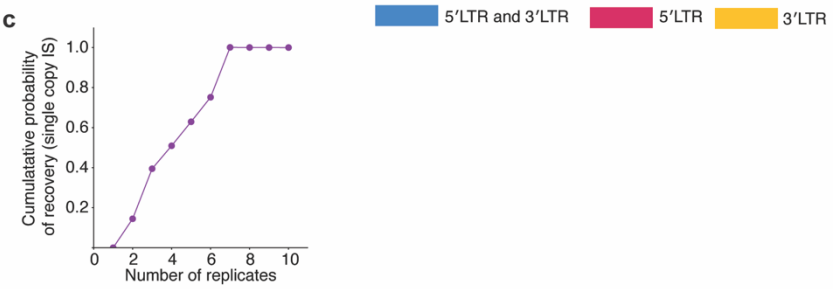


Figure 2. Detection efficiency of PRISM-seq in artificial mixes. (a and b) Two artificial pools were constructed to test PRISM-seq's ability to recover integration sites in samples with multiple unique integration sites at varying relative abundance. Pool 1 contained 11 of the 14 clones #1-14, whereas pool 2 contained all 14 clones #1-14. Each pool was divided into ten independent genomic DNA sampling replicates and independently subjected to PRIMS-seq. The "average copies per replicate" column shows the copy 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. True/false denotes if a high-confidence integration site was recovered by either proviral end. **(c)** Cumulative recovery of single-copy clone integration sites after adjusting for Poisson sampling expectation across 10 replicates for pool 2. Clones #6 and #10 were associated with more than one integration site and were excluded from this analysis. *IS, integration sites.*

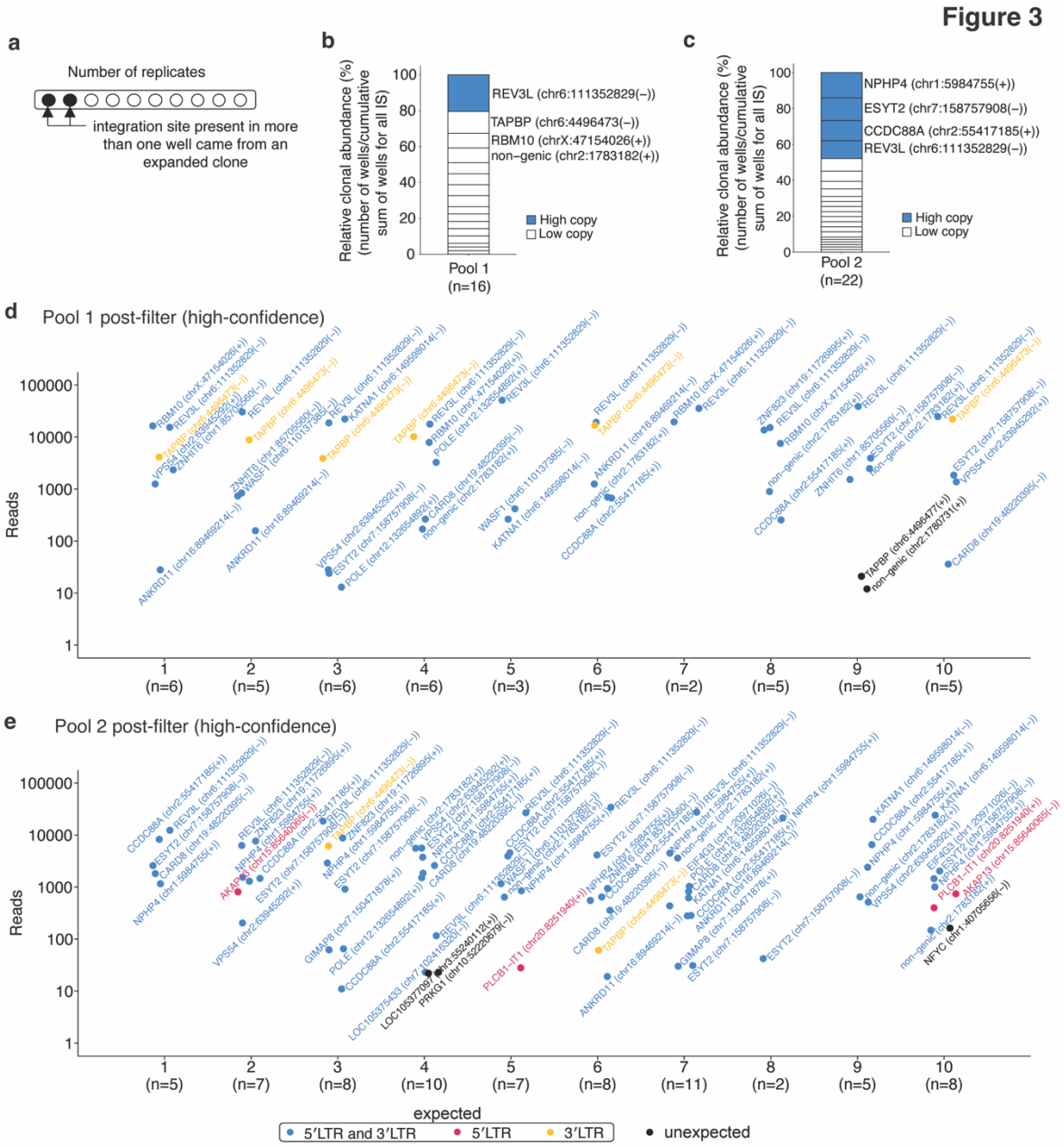


Figure 3. Post-quality filter recovery of high-confidence integration sites and estimation of clonal abundance. (a) Schematics showing the identification of expanded clones using PRISM-seq. We used the post-filtered data to evaluate integration site authenticity and to estimate clonal abundance in pool 1 and pool 2, containing multiple integration sites. **(b and c)** Post-filter estimation of the relative clonal abundance of individual clones found in pool 1 and pool 2. After applying steps1-4 of the quality-control filter, we recovered a total of 16 and 22 high-confidence integration sites in pool 1 and 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% noise) and 3 of 22 (13.6% noise), respectively. The top 4 abundant clones are labelled in the stack bar plot. **(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. Expected high-confidence integration sites are highlighted if recovered by both 5′LTR and 3′LTR (blue), or exclusively

743 by either 5′LTR (red), or 3′LTR (yellow) viral-host junction. Unexpected integration sites are
744 colored in black. *IS, integration sites.*

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

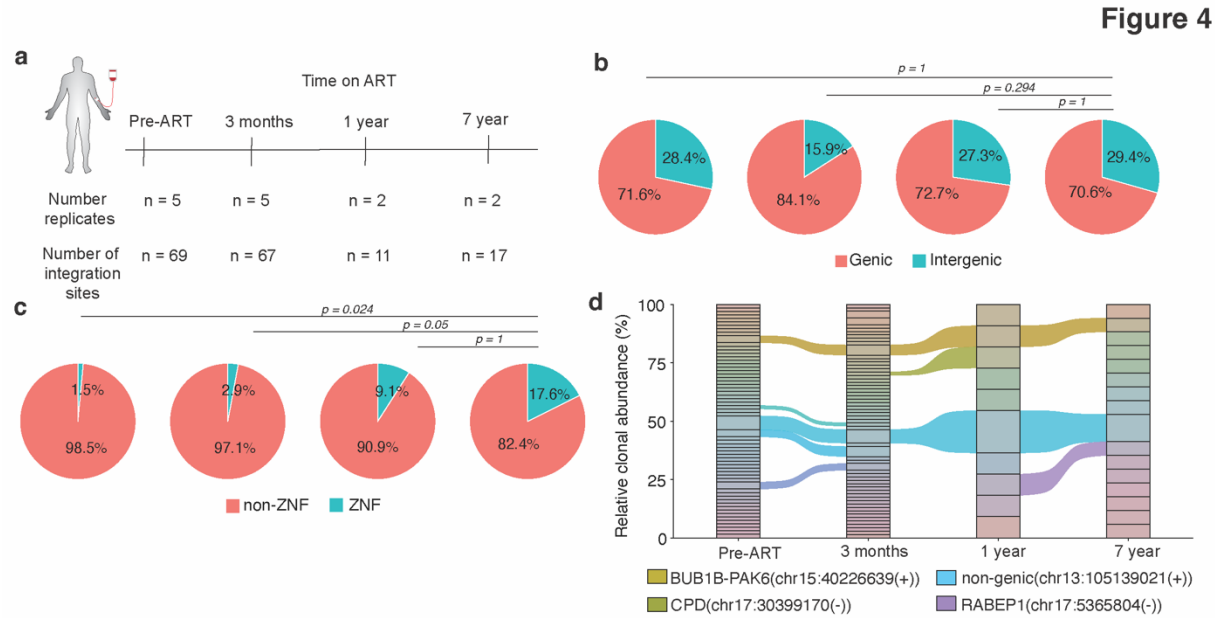


Figure 4. Longitudinal clonal tracking and identification of proviral integration sites in a person living with HIV-1. (a) Number of unique integration sites identified at each time point using PRISM-seq from a person living with HIV-1 under long-term antiretroviral therapy (ART) treatment. Across the number of replicates performed, we sampled in total 250, 250, 100, and 100 HIV-containing templates at each indicated time point. **(b)** Precent genic and intergenic distribution of proviral integration sites. **(c)** 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. Shared clones identified at all the time points are labelled with their respective proviral integration sites.

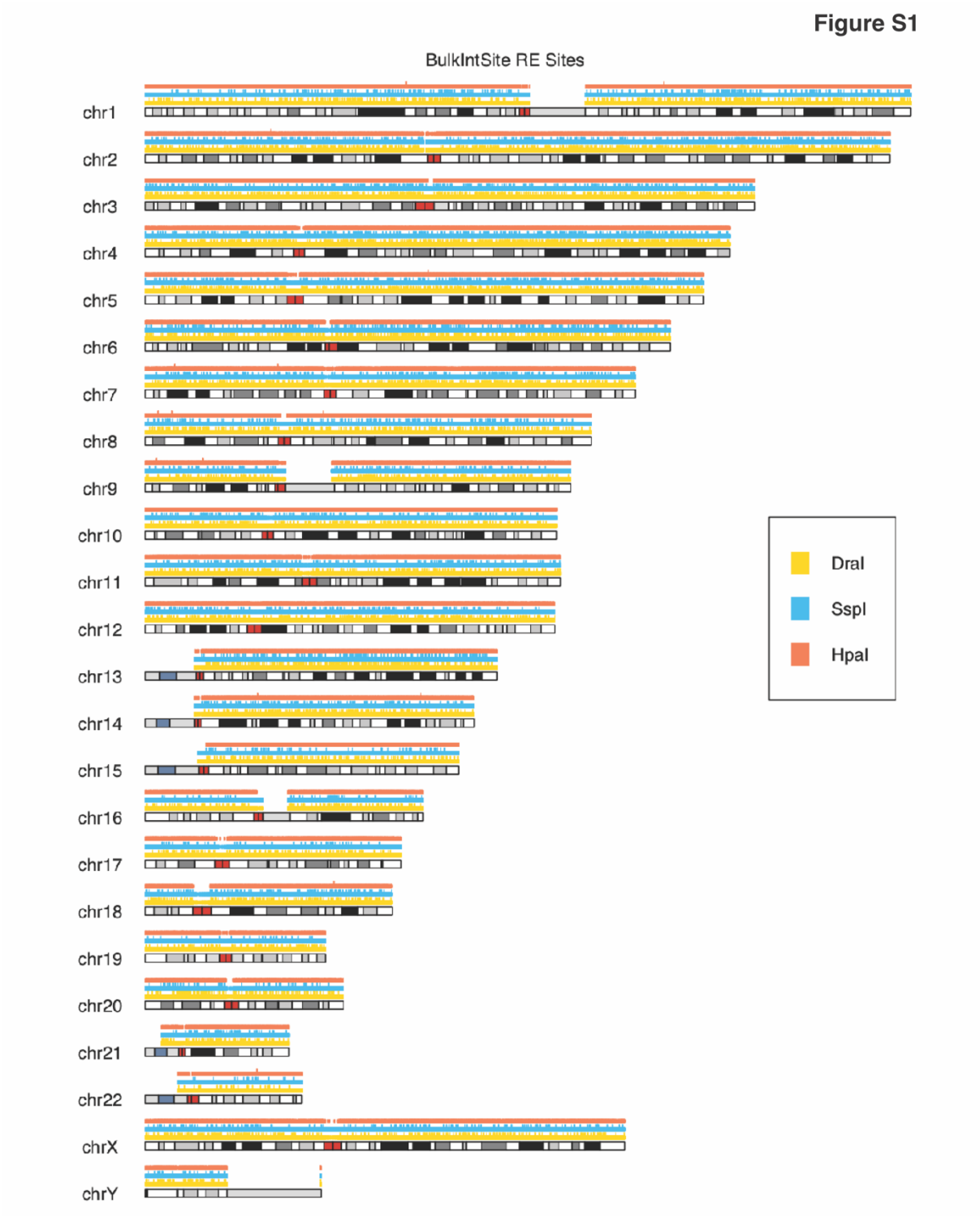


Figure S1. Genome-wide distribution of the restriction enzyme recognition site. The karyoplot shows the DraI (yellow), SspI (cyan), and HpaI (orange) restriction sites distribution genome-wide on the human reference genome assembly hg38. *RE, restriction enzymes.*

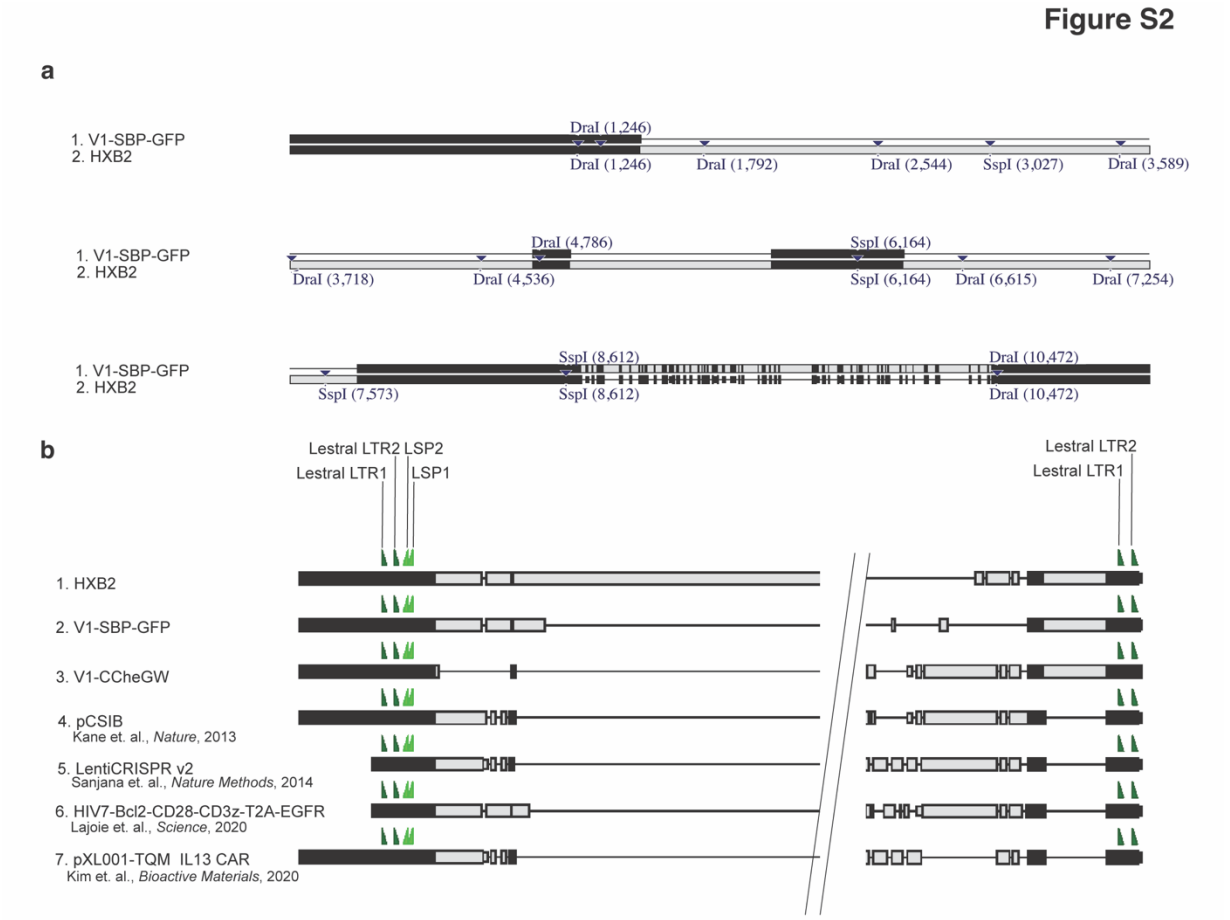


Figure S2. Distribution of restriction enzyme sites within the provirus genome and PRISM-seq primer binding sites compatibility across HIV and lentiviral vector delivery systems. (a) Location of DraI and SspI restriction sites in the genome of the full-length HIV-1 reference sequence (HXB2) and replication-deficient HIV-1 reporter (V1-SBP-GFP) from 5′ to 3′ direction. Restriction sites coordinates in the proviral genome are mentioned within brackets next to each restriction enzyme. 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. Primer binding sites for 5′LTR (LSP1 and LSP2) and 3′LTR (LestralLTR1 and LestralLTR2) viral-host junction amplification are conserved across these sequences.

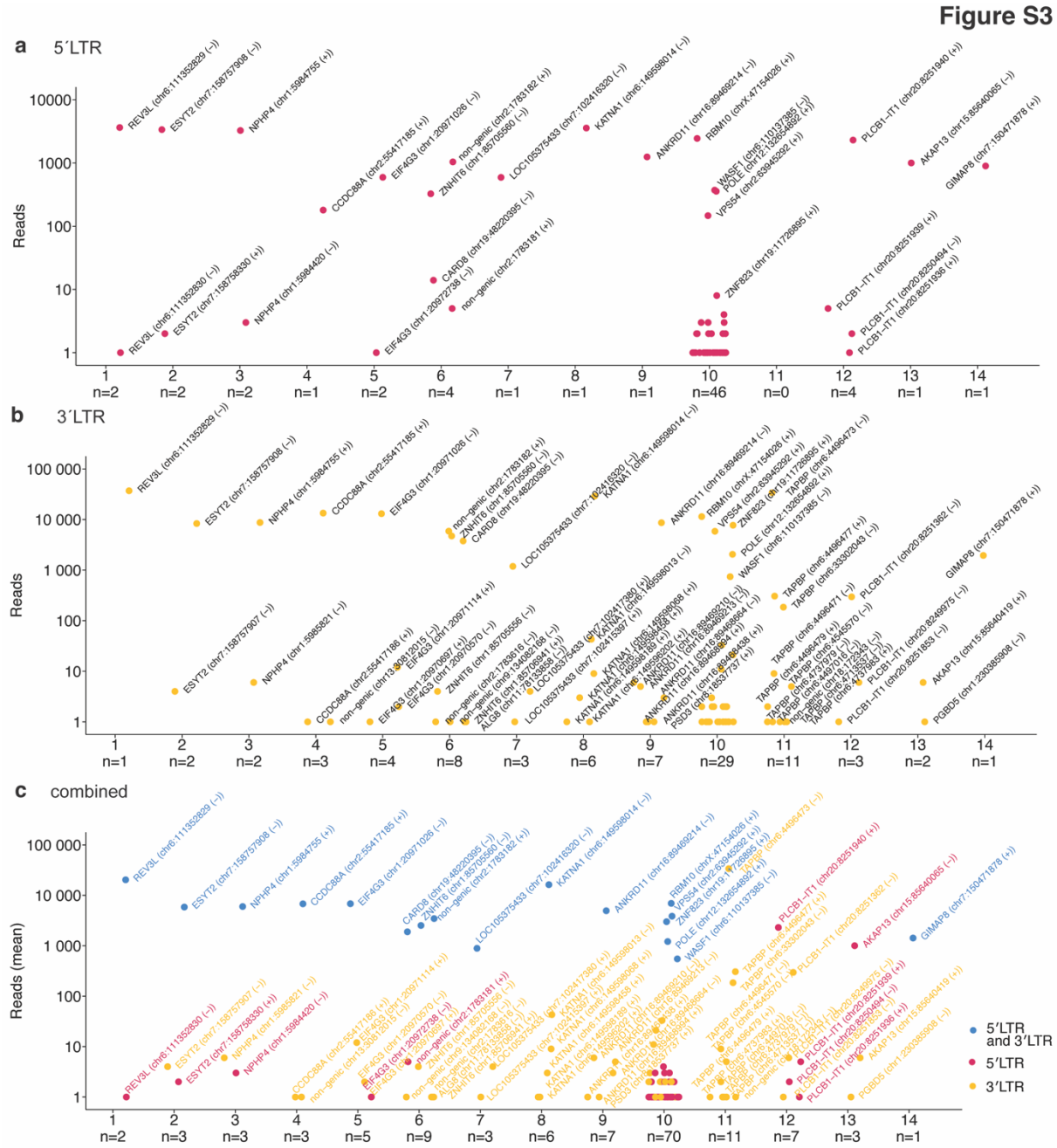


Figure S3. Identification of proviral integration sites using PRISM-seq in putative single-cell clones #1-14, raw pre-filter data. (a-c) Proviral integration sites retrieved from individual putative clones, either from the 5′LTR (a) or 3′LTR (b) or combined (c) viral host junction. Blue, red, and yellow 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.

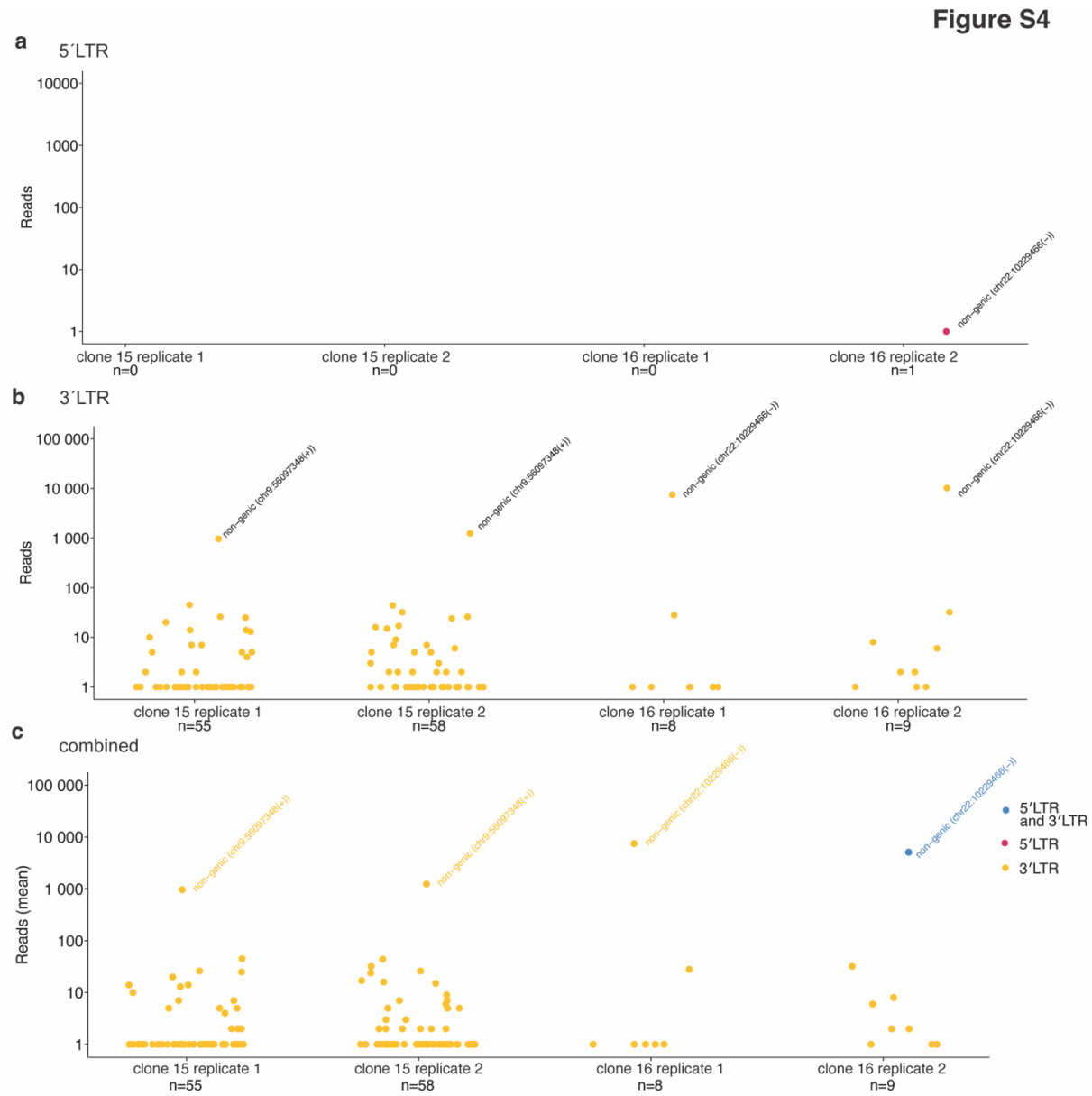


Figure S4. Identification of proviral integration sites using PRISM-seq in clones with known centromeric integrations #15 and #16, raw pre-filter data. (a-c) Proviral integration sites were identified in clones #15 and #16, either from the 5′LTR (a) or 3′LTR (b) or combined (c) viral host junction using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

Figure S5

a

Types of staggers	examples
Integration sites 1 base pair apart (1 bp stagger)	REV3L (chr6:111352829 in negative strand) and REV3L (chr6:111352830 in negative strand)
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)
Integration sites 10-100 bp pair apart (10-100 bp stagger)	EIF4G3 (chr1:20971026 in negative strand and EIF4G3 (chr1:20971114 in positive strand)
Integration sites 100- 3000 bp apart (100-3000 bp stagger)	KATNA1(chr6: 149598014 in negative strand and KATNA1 (chr6: 149596202 in positive strand)

b

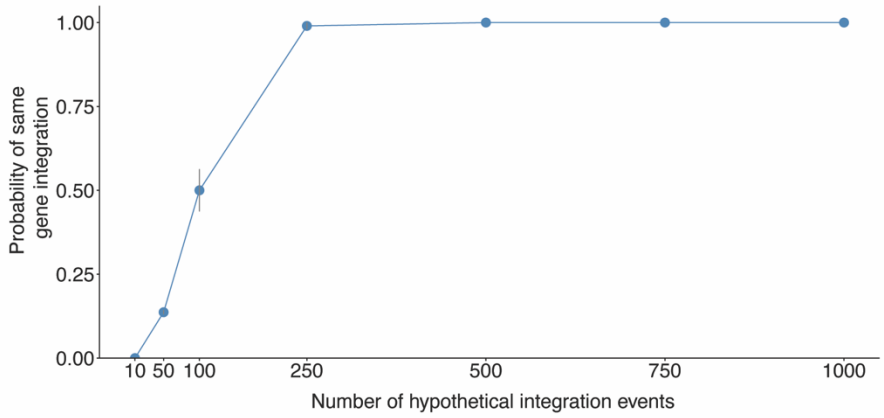
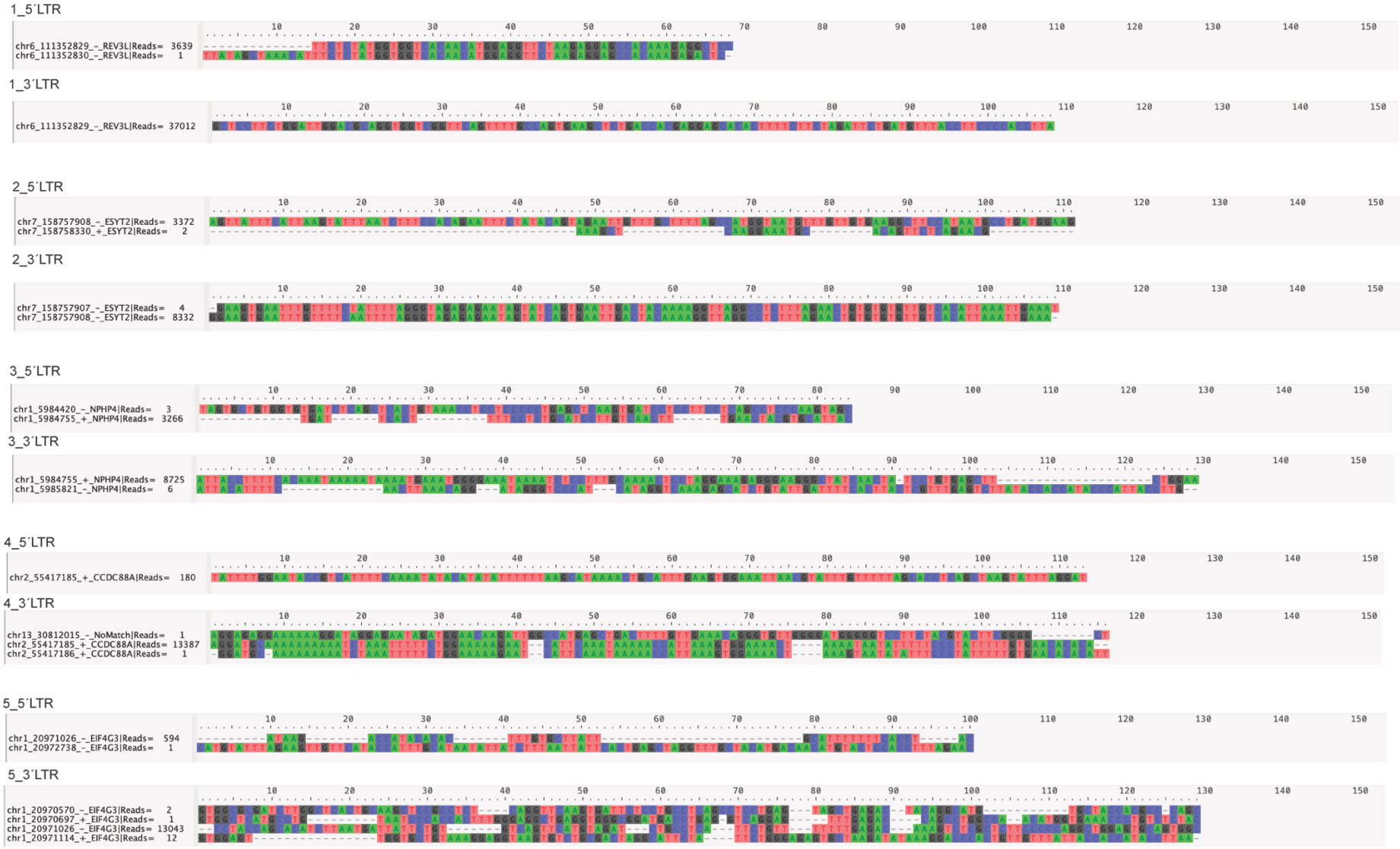


Figure S5. Identification of look-alike integration site artifact cases and the relationship between integration events and the probability of having true multiple integrations into the same gene. (a) Characterization of artifacts generated during the integration sites assay. Each staggered case is distinguished 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. **(b)** Regarding Step 3 in our proposed quality filter, which involves collapsing integration sites that map to the same gene when found within a single reaction: We recognize that when PRISM-seq is applied to non-clonal clinical samples, there is a genuine possibility that proviruses have integrated into distinct locations within the same gene. The likelihood of detecting genuine same 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 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. *Bp, base pairs.*

Figure S6


1_5´LTR
chr6_111352829_-_REV3L|Reads= 3639
chr6_111352830_-_REV3L|Reads= 1
1_3´LTR
chr6_111352829_-_REV3L|Reads= 37012
2_5´LTR
chr7_158757908_-_ESYT2|Reads= 3372
chr7_158758330_+_ESYT2|Reads= 2
2_3´LTR
chr7_158757907_-_ESYT2|Reads= 4
chr7_158757908_-_ESYT2|Reads= 8332
3_5´LTR
chr1_5984420_-_NPHP4|Reads= 3
chr1_5984755_+_NPHP4|Reads= 3266
3_3´LTR
chr1_5984755_+_NPHP4|Reads= 8725
chr1_5985821_-_NPHP4|Reads= 6
4_5´LTR
chr2_55417185_+_CCDC88A|Reads= 180
4_3´LTR
chr13_30812015_-_NoMatch|Reads= 1
chr2_55417185_+_CCDC88A|Reads= 13387
chr2_55417186_+_CCDC88A|Reads= 1
5_5´LTR
chr1_20971026_-_EIF4G3|Reads= 594
chr1_20972738_-_EIF4G3|Reads= 1
5_3´LTR
chr1_20970570_-_EIF4G3|Reads= 2
chr1_20970697_+_EIF4G3|Reads= 1
chr1_20971026_-_EIF4G3|Reads= 13043
chr1_20971114_+_EIF4G3|Reads= 12

835 **Figure S6. Comparison of host sequences for each unique integration site identified within**
836 **the putative clones**. Low-confidence integration sites shared the same junction sequence as
837 the corresponding high-confidence integration site in clones #1, #2, and #4.

838

839

840

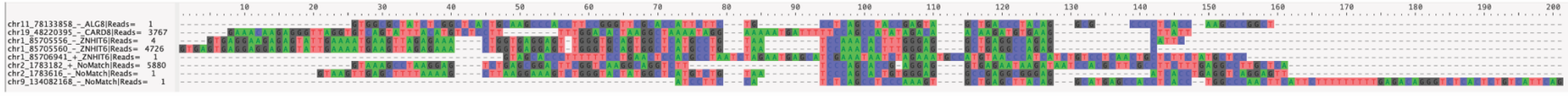
841

842

Figure S7

6_5´LTR

6_3´LTR

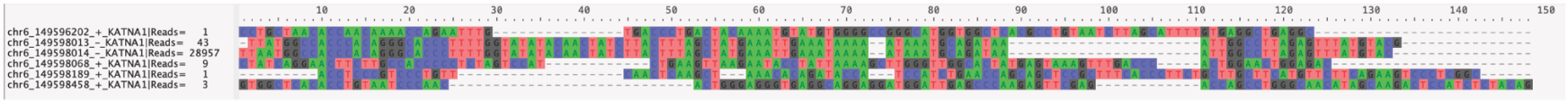


7_5´LTR

7_3´LTR

8_5´LTR

8_3´LTR



9_5´LTR

9_3´LTR

843 **Figure S7. Comparison of host sequences for each unique integration site identified within**
844 **the putative clones**. Low-confidence integration sites shared the same junction sequence as
845 the corresponding high-confidence integration site in clones #6, #8, and #9.

846

847

848

849

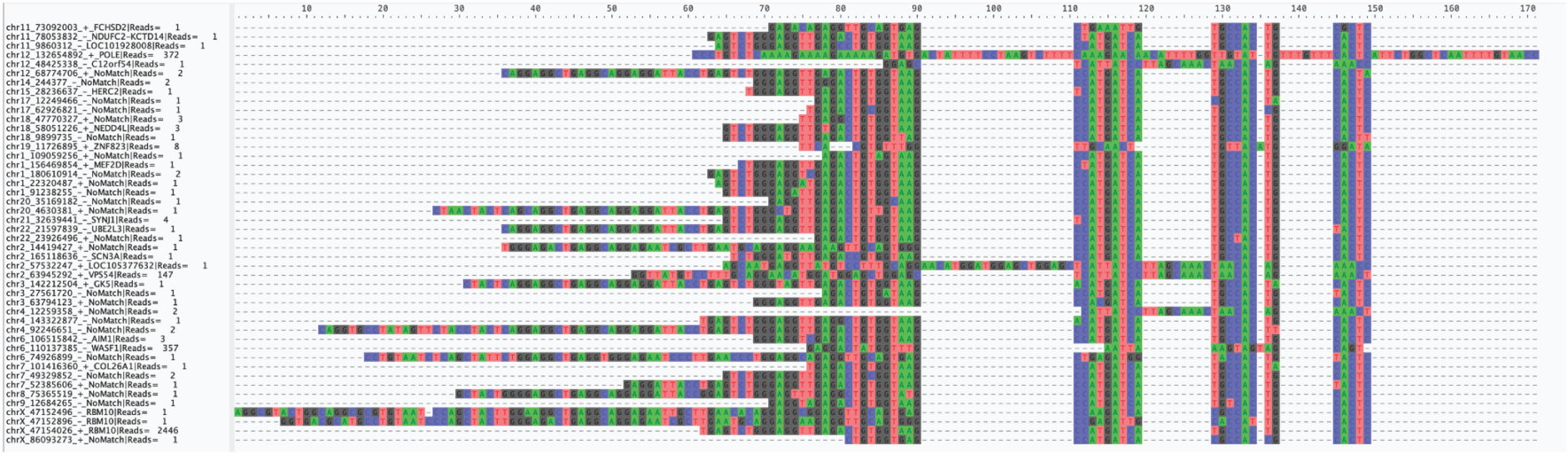
850

851

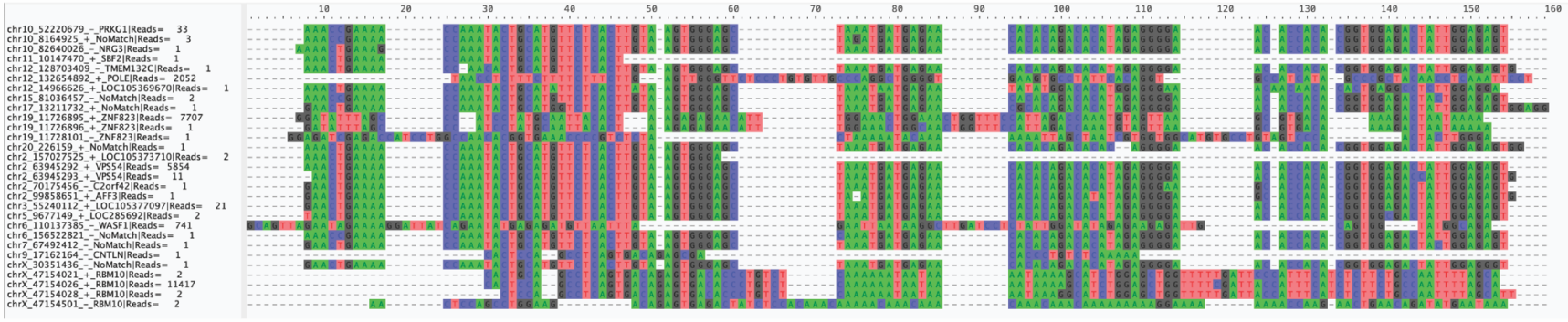
852

Figure S8

10_5´LTR



10_3´LTR



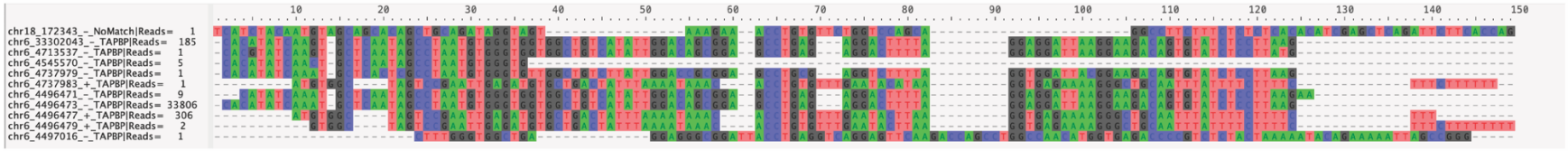
854 **Figure S8. Comparison of host sequences for each unique integration site identified within**
855 **the putative clones**. Low-confidence integration sites shared the same junction sequence as
856 the corresponding high-confidence integration site in clone #10.

857

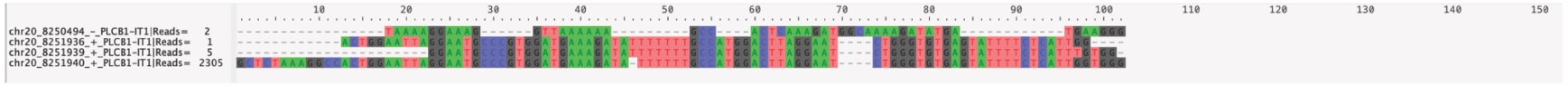
858

Figure S9

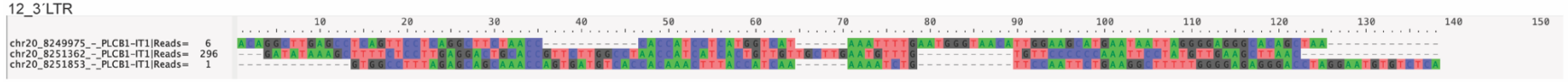
11_3´LTR



12_5´LTR



12_3´LTR

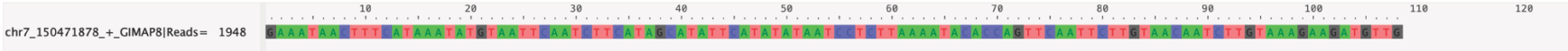


13_5´LTR

13_3´LTR

14_5´LTR

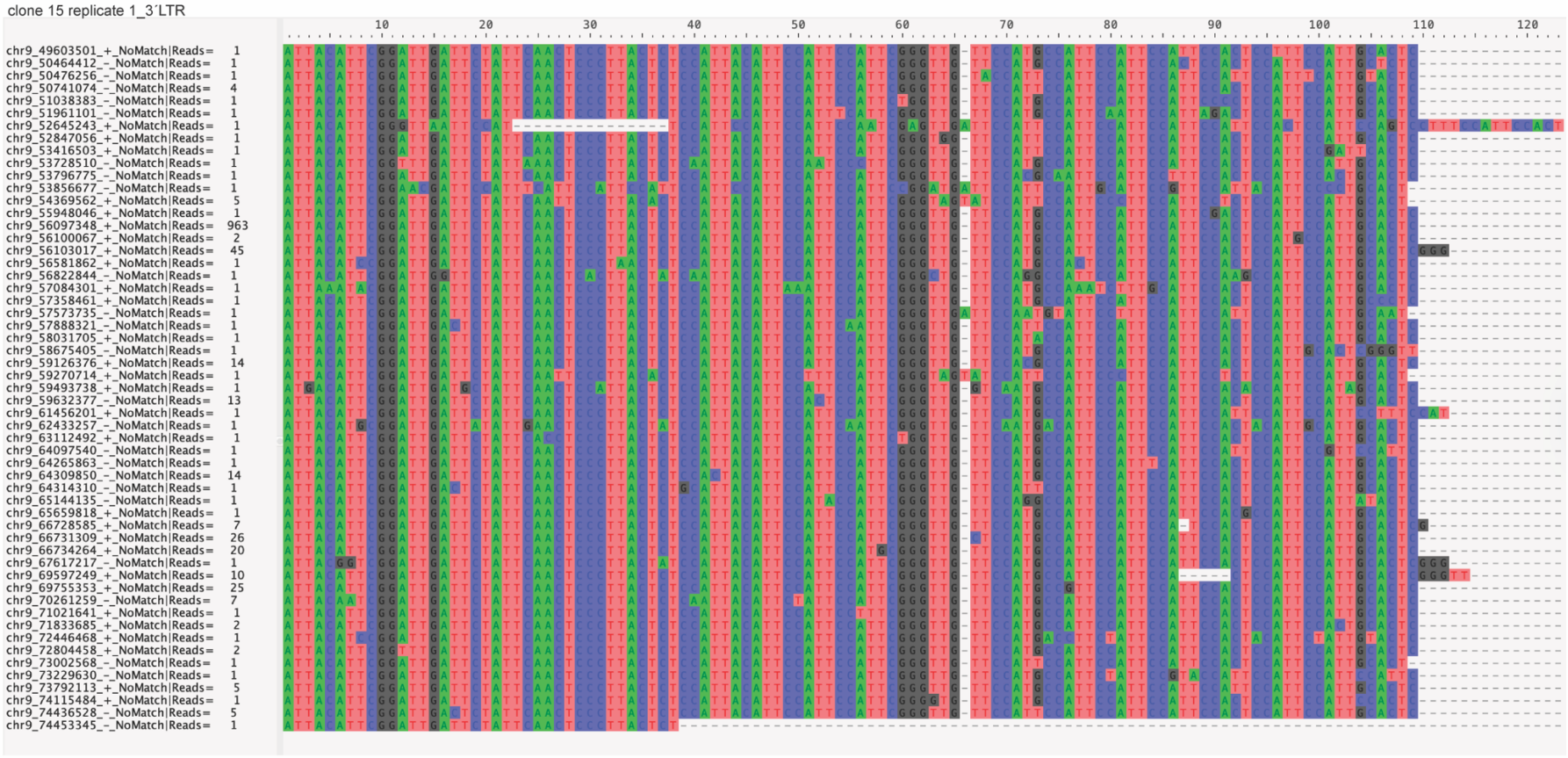
14_3´LTR



860 **Figure S9. Comparison of host sequences for each unique integration site identified within**
861 **the putative clones**. Low-confidence integration sites shared the same junction sequence as
862 the corresponding high-confidence integration site in clones #11 and #12.

863

Figure S10

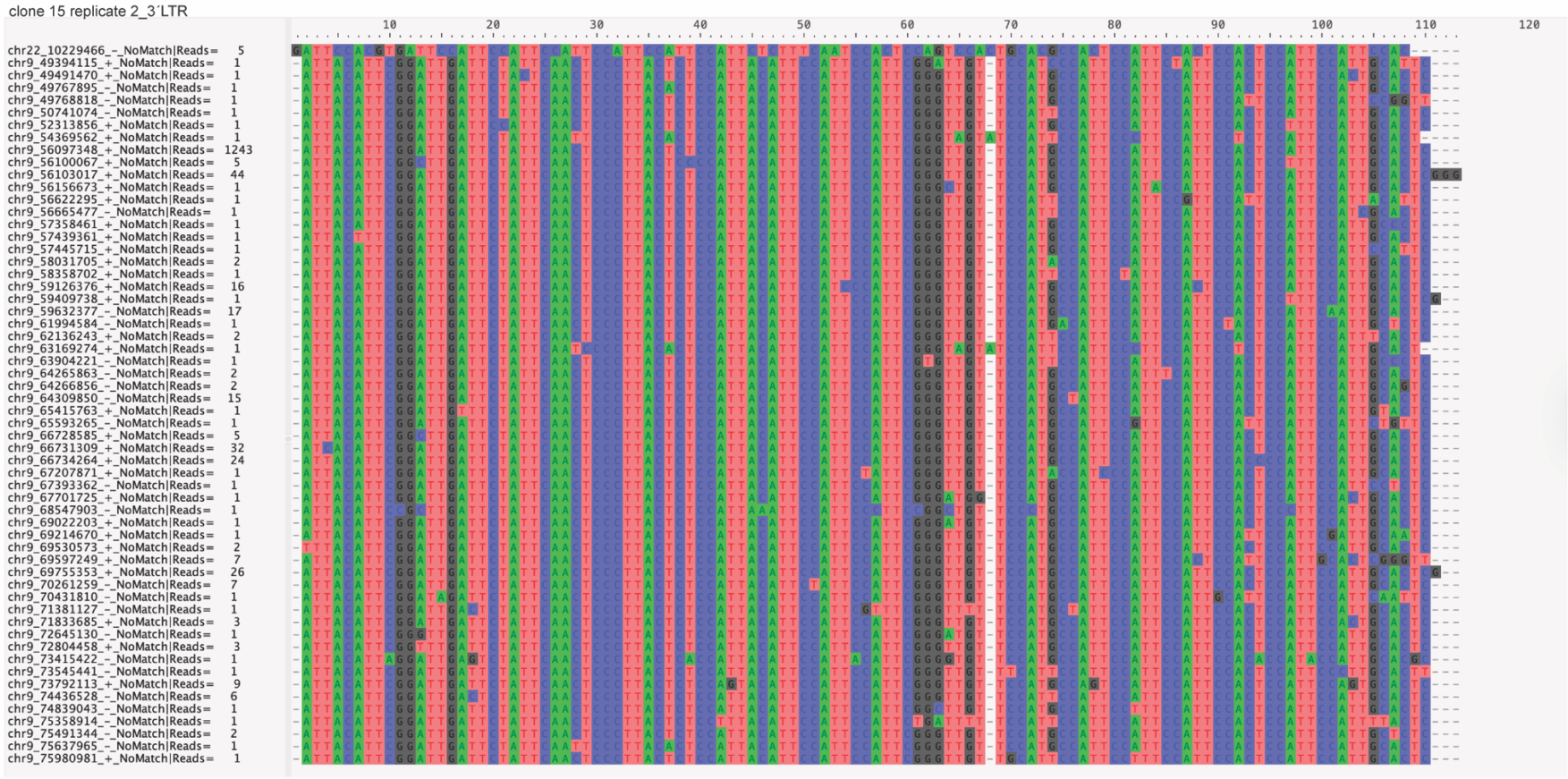

clone 15 replicate 1_3´LTR
10
20
30
40
50
60
70
80
90
100
110
120
chr9_49603501_+_NoMatch|Reads= 1
chr9_50464412_-_NoMatch|Reads= 1
chr9_50476256_-_NoMatch|Reads= 1
chr9_50741074_-_NoMatch|Reads= 4
chr9_51038383_-_NoMatch|Reads= 1
chr9_51961101_-_NoMatch|Reads= 1
chr9_52645243_+_NoMatch|Reads= 1
chr9_52847056_+_NoMatch|Reads= 1
chr9_53416503_+_NoMatch|Reads= 1
chr9_53728510_-_NoMatch|Reads= 1
chr9_53796775_-_NoMatch|Reads= 1
chr9_53856677_-_NoMatch|Reads= 1
chr9_54369562_+_NoMatch|Reads= 5
chr9_55948046_+_NoMatch|Reads= 1
chr9_56097348_+_NoMatch|Reads= 963
chr9_56100067_+_NoMatch|Reads= 2
chr9_56103017_+_NoMatch|Reads= 45
chr9_56581862_+_NoMatch|Reads= 1
chr9_56822844_-_NoMatch|Reads= 1
chr9_57084301_+_NoMatch|Reads= 1
chr9_57358461_+_NoMatch|Reads= 1
chr9_57573735_-_NoMatch|Reads= 1
chr9_57888321_-_NoMatch|Reads= 1
chr9_58031705_+_NoMatch|Reads= 1
chr9_58675405_-_NoMatch|Reads= 1
chr9_59126376_+_NoMatch|Reads= 14
chr9_59270714_+_NoMatch|Reads= 1
chr9_59493738_+_NoMatch|Reads= 1
chr9_59632377_-_NoMatch|Reads= 13
chr9_61456201_+_NoMatch|Reads= 1
chr9_62433257_-_NoMatch|Reads= 1
chr9_63112492_+_NoMatch|Reads= 1
chr9_64097540_-_NoMatch|Reads= 1
chr9_64265863_-_NoMatch|Reads= 1
chr9_64309850_-_NoMatch|Reads= 14
chr9_64314310_-_NoMatch|Reads= 1
chr9_65144135_-_NoMatch|Reads= 1
chr9_65659818_+_NoMatch|Reads= 1
chr9_66728585_+_NoMatch|Reads= 7
chr9_66731309_+_NoMatch|Reads= 26
chr9_66734264_+_NoMatch|Reads= 20
chr9_67617217_-_NoMatch|Reads= 1
chr9_69597249_+_NoMatch|Reads= 10
chr9_69755353_+_NoMatch|Reads= 25
chr9_70261259_-_NoMatch|Reads= 7
chr9_71021641_+_NoMatch|Reads= 1
chr9_71833685_+_NoMatch|Reads= 2
chr9_72446468_+_NoMatch|Reads= 1
chr9_72804458_+_NoMatch|Reads= 2
chr9_73002568_-_NoMatch|Reads= 1
chr9_73229630_-_NoMatch|Reads= 1
chr9_73792113_+_NoMatch|Reads= 5
chr9_74115484_+_NoMatch|Reads= 1
chr9_74436528_-_NoMatch|Reads= 5
chr9_74453345_-_NoMatch|Reads= 1

865 **Figure S10. Comparison of host sequences for each unique integration site identified in a**
866 **clone associated with centromeric integration.** Low-confidence integration sites shared the
867 same junction sequence as the corresponding high-confidence integration site in clone #15,
868 replicate 1.

869

870

Figure S11

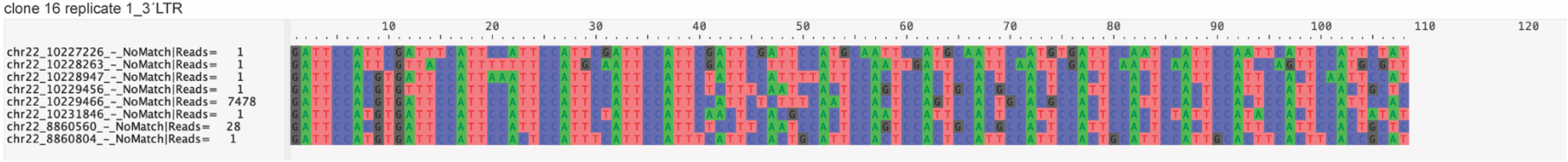

clone 15 replicate 2_3´LTR
10
20
30
40
50
60
70
80
90
100
110
120
chr22_10229466_-_NoMatch|Reads= 5
chr9_49394115_+_NoMatch|Reads= 1
chr9_49491470_+_NoMatch|Reads= 1
chr9_49767895_-_NoMatch|Reads= 1
chr9_49768818_-_NoMatch|Reads= 1
chr9_50741074_-_NoMatch|Reads= 1
chr9_52313856_+_NoMatch|Reads= 1
chr9_54369562_+_NoMatch|Reads= 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

872 **Figure S11. Comparison of host sequences for each unique integration site identified in a**
873 **clone associated with centromeric integration.** Low-confidence integration sites shared the
874 same junction sequence as the corresponding high-confidence integration site in clone #15,
875 replicate 2.

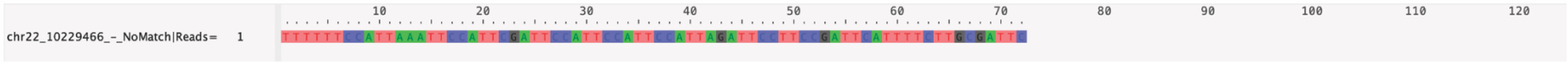
876

Figure S12

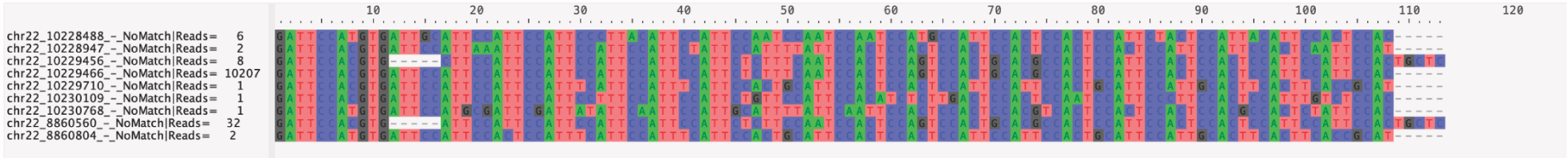
clone 16 replicate 1_3´LTR



clone 16 replicate 2_5´LTR



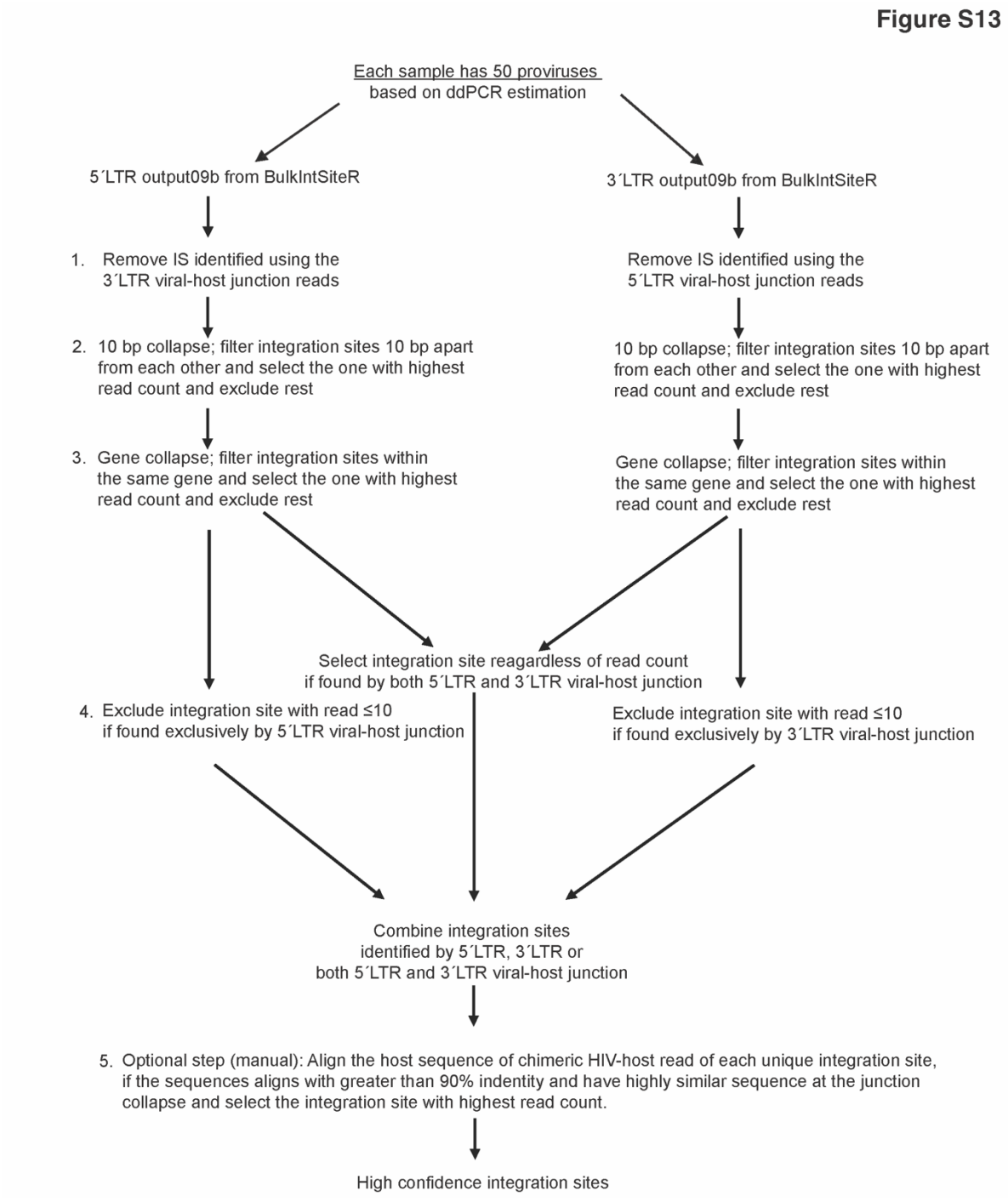
clone 16 replicate 2_3´LTR



878 **Figure S12. Comparison of host sequences for each unique integration site identified in a**
879 **clone associated with centromeric integration.** Low-confidence integration sites shared the
880 same junction sequence as the corresponding high-confidence integration site in clone #16.

881

882



883

884 **Figure S13. Data-guided noise removal approach.** Leveraging the clonal nature of Clones
885 #1-16, we devised a quality control filter to remove assay noise. This set of quality filters is
886 applied to the BulkIntSiteR output; (1) filter and remove the spillover noise of integration site
887 identified by 3′ LTR in the 5′ LTR library and vice versa. (2) integration sites within ≤10 bp of

each other 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. Steps 1-4 are fully automated, and the accompanying script is included with the BulkIntSiteR package; Step 5 requires user input and manual curation.

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
c
clone 15 replicate 2_3´LTR
d
clone 16 replicate 1_3´LTR
e
clone 16 replicate 2_5´LTR
clone 16 replicate 2_3´LTR

Figure S14. Post-filter recovery of proviral integration sites found within centromeric clones #15 and #16. (a) Integration sites identified by the BulkIntSiteR software were processed using the QC filtering pipeline (with steps 1-4 applied), 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.

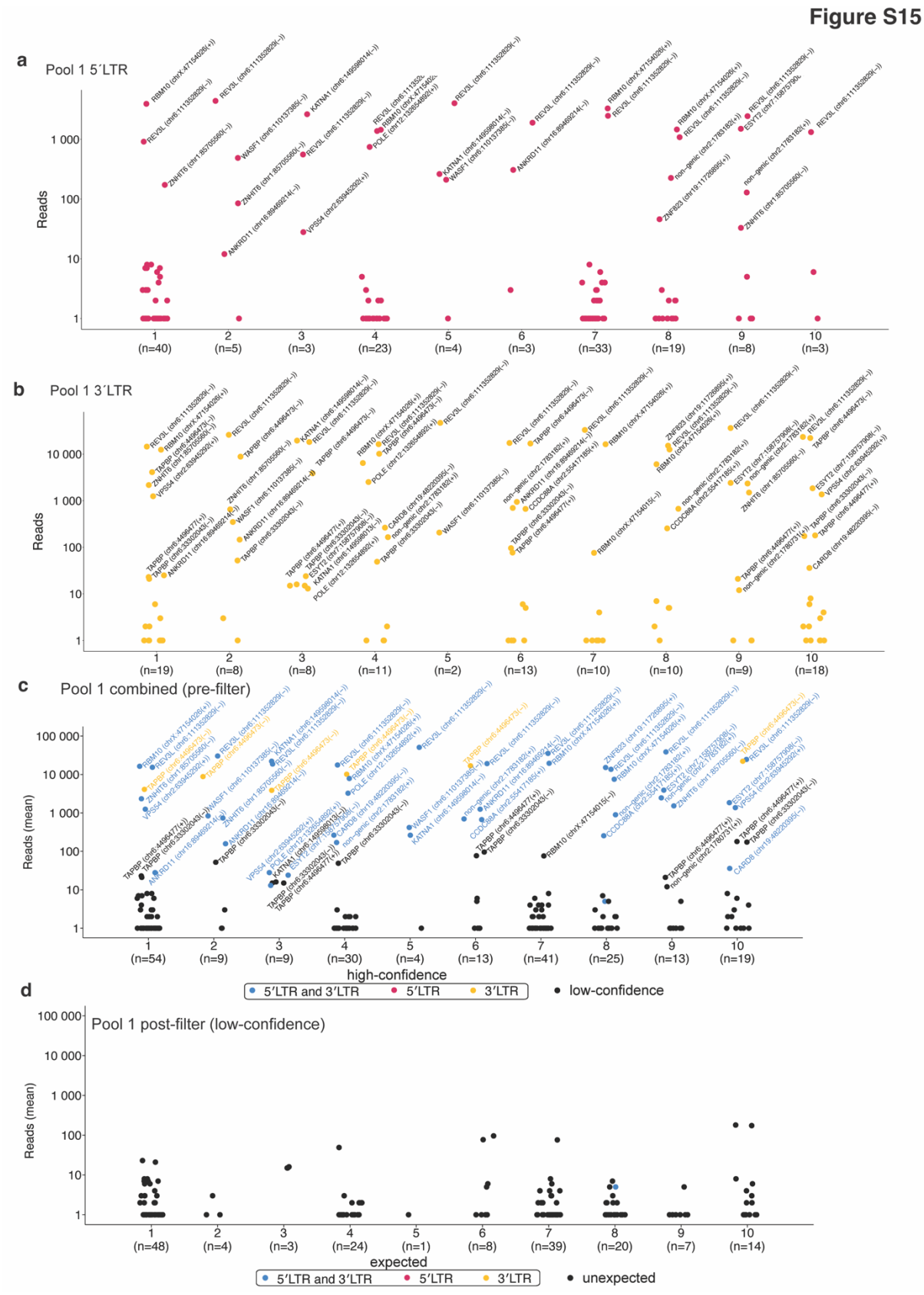


Figure S15. Proviral integration sites identification in pool 1. (a-c) Dot plots representing the total number of unique integration sites identified in each replicate of pool 1 by either the 5′LTR (a) or 3′LTR (b) or combined (c) viral-host junction. The total number of unique

913 integration sites retrieved in an individual replicate is represented as an n value. Blue, red, and
914 yellow dots represent integration sites identified by both 5′LTR and 3′LTR, exclusively by
915 5′LTR or exclusively by 3′LTR PRISM-seq reactions, respectively. (d) Post-filter identification
916 of low-confidence expected and unexpected integration sites.

917

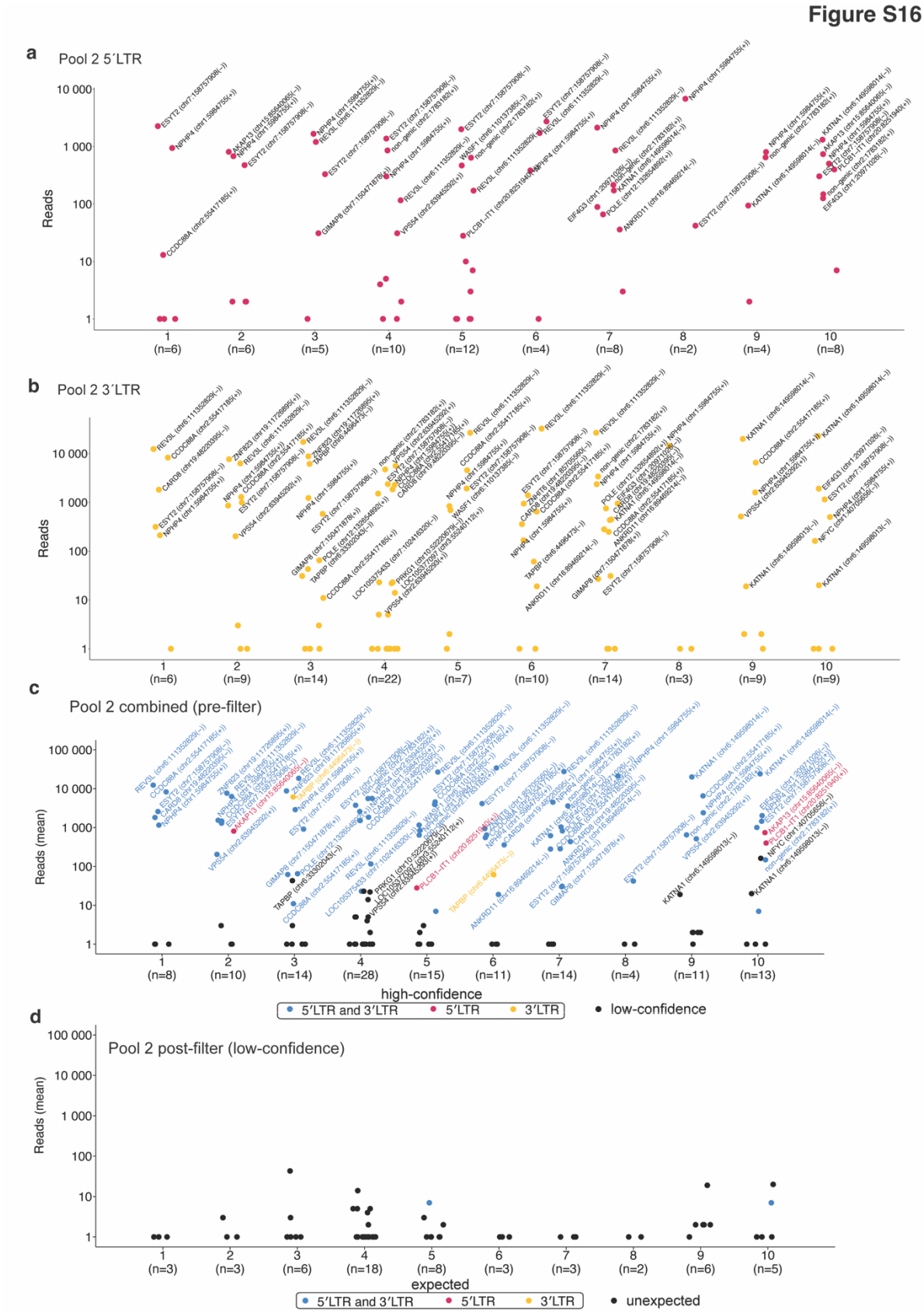


Figure S16. Proviral integration sites identification in pool 2. (a-c) Dot plots representing the total number of unique integration sites identified in each replicate of pool 2 by either the 5′LTR (a) or 3′LTR (b) or combined (c) viral-host junction. The total number of unique

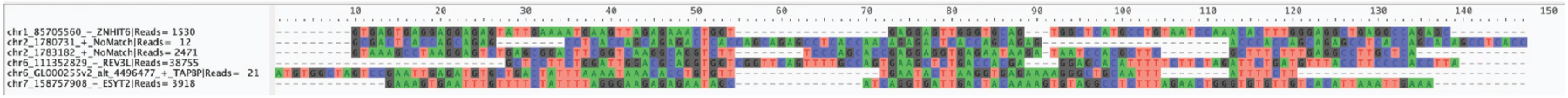
integration sites retrieved in an individual replicate is represented as an n value. Blue, red, and yellow dots represent the integration sites identified by both 5′LTR and 3′LTR, exclusively by 5′LTR or exclusively by 3′LTR PRISM-seq reactions, respectively. (d) Post-filter identification of low-confidence expected and unexpected integration sites.

Figure S17

a

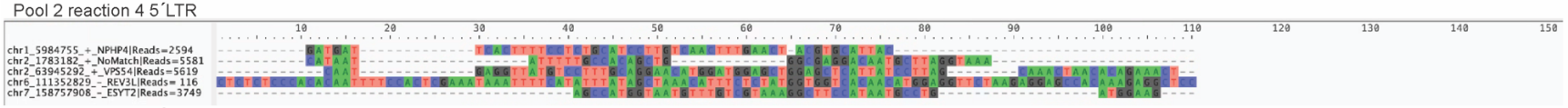
Pool 1 reaction 9 5´LTR

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). Host sequence derived from unexpected integration sites found in pool 1 replicate 9 (a) and pool 2 replicate 4 (b) by both 5′LTR and 3′LTR PRISM-seq reactions. Unexpected integration sites sharing the same junction sequence as the high-confidence site and/or are ±10 kb away from the expected high-confidence integration site are labelled with an asterisk (*).