Title:-

LoVIS-Seq reveals clonal dynamics of human-HSPC in humanized mice and vector integration bias

Gajendra W. Suryawanshi[1,2]; Hubert Arokium[1,2]; Sanggu Kim[6,7,8]; Wannisa Khamaikawin[4,5]; Samantha Lin[4]; Saki Shimizu[4]; Koollawat Chupradit[4]; YooJin Lee[1,2]; Yiming Xie[1,2]; Xin Guan[1,2]; Vasantika Suryawanshi[9]; Angela P. Presson[10,11]; Dong-Sung An[2,4]; Irvin S. Y. Chen[1,2,3#]

[1]Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]UCLA AIDS Institute, Los Angeles, CA, 90095, USA
[3]Division of Hematology-Oncology, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA
[4]School of Nursing, University of California, Los Angeles, CA, 90095, USA.
[5]Present address: Faculty of Medicine, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
[6]Department of Veterinary Biosciences, College of Veterinary Medicine, [7]Center for Retrovirus Research, [8]Infectious Disease Institute, The Ohio State University, Columbus, OH 43210, USA
[9]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA, 90089, USA
[10]Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, 84108
[11]Department of Biostatistics, University of California, Los Angeles, 90095

#Corresponding authors:
Irvin S. Y. Chen;
615 Charles E. Young Dr. South
BSRB, Rm 173,
Los Angeles, CA 90095
syuchen@mednet.ucla.edu

Abstract:

Clonal repopulation of human hemopoietic stem and progenitor cells (HSPC) in humanized mouse models remains only partially understood due to the lack of a quantitative clonal tracking technique for low sample volumes. Here, we present a low-volume vector integration site sequencing (LoVIS-Seq) assay that requires a mere 25µl mouse blood for quantitative clonal tracking of HSPC. Using LoVIS-Seq, we longitudinally tracked 897 VIS clones—providing a first-ever demonstration of clonal dynamics of both therapeutic and control vector-modified human cell populations simultaneously repopulating in humanized mice. Polyclonal repopulation of human cells became stable at 19 weeks post-transplant indicating faster clonal repopulation than observed in humans. Multi-omics data of human fetal liver HSPC provides a definitive evidence of vector integration preference for H3K36me3-enriched regions. Despite this bias the repopulation remains normal, underscoring the safety of gene therapy vectors. LoVIS-Seq provides an efficient tool for exploring gene therapy and stem cell biology in small-animal models.

Introduction:

Hemopoietic stem cells (HSC) are an ideal vehicle for introducing gene-modified cells to treat genetic disorders, cancers, and viral infections. Humanized mouse models—immunodeficient mice transplanted with human stem cells or tissues that generate a functioning human immune system—provide the most practical in vivo system for stem cell and disease research (reviewed in[1]). In particular, humanized bone marrow-liver-thymus mouse (hu-BLT mouse) models can support the development of T cells, B cells, monocytes, macrophages, and dendritic cells. Moreover, these hu-BLT mice demonstrate human MHC-restricted T cell response to Epstein-Barr virus (EBV) infection and human Dendritic cells-mediated T cell response against toxic shock syndrome toxin 1 (TSST1)[2]. Capable of mounting both innate and adaptive immune response, the hu-BLT mouse model is well-suited for antiviral gene therapy research. A recent study used an HIV-1 pre-infected hu-BLT mouse model to demonstrate that HIV-1 infection induces selective expansion of anti-HIV-1 dual shRNA gene-modified (protected) CD4+ T cells over control vector-modified unprotected CD4+ T cells[3]. However, whether the human HSPC in xenograft mouse models exhibit their human traits or clonally behave like mouse cells remains unclear.

Longitudinal clonal tracking in humans and macaques revealed biphasic expansion of transplanted HSPC: an early phase of rapid and transient expansion of short-term HSC and a late phase (~1 year post-transplant) of sustained expansion of long-term HSC[4,5]. However, clonal tracking in mice autologously transplanted with a limited number of barcoded HSC (marked with a unique sequence tag using lentiviral vector) showed that clones start to stabilize around week 12 post-transplant and progressively fewer clones contribute to the overall repopulation[6]. Another barcode tracking study in mice suggested transplantation dose-dependent change in HSC differentiation[7]. However, generating a barcode library for every therapeutic test vector is

both cost-prohibitive and impractical; additionally, low DNA availability, lack of a universal barcode counting method, and small barcode library size limit the accuracy of barcoding techniques. Finally, these techniques lack the ability to identify genomic location of vector integration in host genomes.

In each transduced HSPC, the vector randomly integrates into the host genome, creating a unique vector-host DNA junction sequence or VIS clone. A high-throughput integration sites (IS) sequencing assay can simultaneously identify and track multiple VIS as well as detect probable mutagenic insertions. A quantitative high-throughput VIS assay revealed that of all HSPC transplanted in rhesus macaques, ~0.01% are long-term HSC and start contributing >1.5 years post-transplant[5]. Our long terminal repeat indexing–mediated integration site sequencing (LTRi-seq) method enables multiplexed and unbiased quantitative clonal analysis of cells gene-modified with anti-HIV or control vector and that of HIV-1 IS—all in the same hu-BLT mouse[8]. VIS analysis at endpoint showed HIV-1 infection induced selective clonal expansion in the anti-HIV (H1 LTR-index) gene-modified population without adverse impact on clonal expansion of the control (H5 LTR-index) vector-modified population. However, only 100µl blood (0.6µg of DNA assuming 1000 cell/µl) can be drawn biweekly from a typical humanized mouse, which is insufficient for longitudinal tracking with our VIS assay requiring ≥1µg DNA. Multiple displacement amplification (MDA), a whole genome amplification technique, used to increase the DNA amount with a very low error rate (1 in 10^6 to 10^7 nucleotides)[9] and high coverage[10]. MDA-amplified DNA has been used to detect retroviral IS[11] and to sequence full-length HIV-1 proviruses including the IS[12,13].

HIV-1 and HIV-1-based vectors are known to favor gene-dense regions in primary T cells and cell lines[14,15] with preference toward transcriptionally active genes, select H3 histone modifications[16], and open chromatin DNase I hypersensitive sites[17]. Implications of integration into active genes on cell fate are compounded by HIV-1 infection induced cytotoxicity. Self-inactivating lentiviral vectors are low risk yet, a strong promoter within the vector can upregulate the expression of endogenous genes where the vector integrated[18,19]. Remarkably, in human HSPC—more specifically human fetal HSPC (FL-HSPC)—integration preference and its impact on proliferation and differentiation these cells in a mouse environment is unknown.

In this study, we present LoVIS-Seq assay for quantitative clonal tracking in small-animal models using low-volume samples. Using this assay, we longitudinally track hundreds of clones from two different gene-modified cell populations simultaneously repopulating in hu-BLT mice. This polyclonal repopulation resembles typical after-transplant HSPC expansion in macaques and humans. In FL-HSPC, we found vector integration is biased for actively transcribed regions. Our method provides an efficient tool to study clonal repopulation in murine and humanized-mouse models used for stem cell and gene therapy research.

Results:

110 *Minimum 10,000 bone marrow cells or 25µl blood is sufficient for LoVIS-Seq:*
111 To test our new assay, we collected bone marrow (BM) cells from hu-BLT (bone marrow-liver-thymus)
112 mouse (m860). Fetal liver CD34+ cells transduced with sh1005(anti-CCR5 shRNA)-EGFP vector or control
113 mCherry vector were mixed in equal ratio and transplanted in the mouse (Figure 1a-b, details in Methods).
114 We estimated clonal composition of the EGFP-WT (WT LTR-index) and mCherry-H1 (H1 LTR-index)
115 populations using unamplified bulk DNA, in triplicate, as described previously[8]. A total of 300 ±42 SD (216
116 ±22 SD mCherry-H1 VIS and 84 ±20 SD EGFP-WT) VIS were detected. The polyclonal profile in mouse
117 bone marrow (Figure 1c) resembled that found in hu-BLT mice[8] and in autologously transplanted mice[6],
118 nonhuman primates[5,20], and humans[4,21]. Next, to test our LoVIS-Seq assay (Figure 1d) that combines MDA
119 with quantitative VIS assays, we first performed MDA directly on 81,000, 27,000, 9,000, 3,000, and 1,000
120 bone marrow cells, each in duplicate (Supplementary figure 1A, details in Methods section). Equal amounts
121 of MDA-amplified DNA and unamplified bulk DNA were used for the VIS assay (Supplementary table 1).
122 We found high reproducibility of clonal profiles in different MDA samples and within-MDA replicates of
123 81,000 to 9,000 cells (avg. Pearson's r value >0.91) (Figure 1e and Supplementary figure 1B-F); for less than
124 9,000 cells, the reproducibility dropped (avg. Pearson's r=0.87 for 3,000 and 0.73 for 1,000 cells).
125 Importantly, reduced cell numbers caused a modest reduction in VIS detection (Supplementary figure 1G).
126 These data validate that MDA-amplified DNA from >10,000 bone marrow cells is sufficient for LoVIS-Seq.
127 Next, to test accuracy of LoVIS-Seq with hu-BLT mouse blood, we collected 100µl blood at week 13, 15, 17
128 and ~1ml of whole blood at week 19 post-transplant. The hu-BLT mice were transplanted with an equal mix
129 of human CD34+ cells transduced with anti-HIV EGFP-WT vector and control mCherry-H5 vector (Figure
130 1f). Cells from 50µl blood were used for flow cytometry and the remaining cells were used for MDA
131 duplicates; each 25µl of blood (>10,000 human cells). High correlation (median Pearson's r =0.93) of
132 mCherry-H5 and EGFP-WT VIS clonal frequencies between unamplified and MDA-amplified DNA from
133 blood cells (Figure 1g) suggests that the clonal profile of entire mouse blood can be captured with 25µl of
134 blood. Importantly, the MDA replicates also showed high reproducibility (median Pearson's r>0.95,
135 Supplementary figure 3A). In conclusion, our LoVIS-Seq assay accurately captured the clonality of two
136 vector-modified cell populations in hu-BLT mouse blood using mere 25µl of blood or as few as 10,000 cells.
137 These results show that using very small sample volume, our assay can provide accurate and reproducible
138 clonal data.
139 *LoVIS-Seq for simultaneous clonal tracking of therapeutic vector-modified and control vector-modified*
140 *populations:*
141 Subsequently, using 25µl of hu-BLT mouse blood for LoVIS-Seq and LTR-indexes (Figure 1a, e), we
142 tracked change in the relative frequencies of 792 mCherry-H5 and 105 EGFP-WT VIS (897 total) clones
143 over 6 weeks (Figure 2a). High correlations between the total mCherry-H5 VIS clonal frequency and

144 mCherry+ cell percentage by flow cytometry (Pearson's r = 0.8) as well as between total EGFP-WT VIS
145 clonal frequency and EGFP+ cells percentage (Pearson's r = 0.9, Supplementary figure 3B) are consistent
146 with our previous study[8]. These results show LoVIS-Seq can quantitatively capture the flow cytometry data.
147 Notably, the expansion in EGFP-WT clones coincided with reduction of mCherry-H5 contribution and vice-
148 versa (Figure 2b; solid lines); these changes closely match the change in $EGFP^+$ and $mCherry^+$ cell
149 percentages measured by flow cytometry (Figure 2b; dashed lines). Interestingly, repopulation in both
150 EGFP-WT and mCherry-H5 populations is largely driven by expansion of a few HSPC clones—a
151 characteristic feature of after-transplant repopulation. These results present a first-ever demonstration of
152 clonal expansion of two populations, therapeutic vector-modified and control vector-modified,
153 simultaneously repopulating in hu-BLT mice.
154 *Sustained polyclonal repopulation with rapid clonal expansion and stabilization:*
155 After identifying polyclonal repopulation of human cells in mice, we investigated the properties of its clonal
156 dynamics. The maximum number of VIS were detected at week 13 and on average, only 13% (±5%) fewer
157 VIS were detected at week 19 compared to week 13. Also, the number of total VIS contributing to
158 repopulation at each timepoint decreased with time (Figure 2c). On average, 61% (±12.8%) of persistent VIS
159 clones (m599: 235 clones, m598: 150 clones, and m591: 165 clones; total 550 clones) consistently
160 contributed for 6 weeks and provided stable polyclonal repopulation (Figure 2a area plots). While the
161 number of VIS clones steadily dropped over time in both the mCherry-H5 and EGFP-WT populations, their
162 clonal profiles became increasingly similar (Supplementary figure 4A-B), comparable to polyclonal
163 repopulation patterns that have been reported in nonhuman primates[5]. Moreover, correlation between time
164 points indicates clonal distribution at week 13 differs from week 19, with clonal expansion stabilizing around
165 week 17 (Supplementary figure 4B). We also examined Rényi's diversity[22] profiles for each animal at each
166 time point (details in Methods section). The clonal diversity at week 13 was highest and subsequently
167 decreased with time (Figure 2d). Diversities were similar between weeks 17 and 19 as indicated by their
168 overlapping diversity profiles. The Shannon[23] and Simpson[24] indices dropped between weeks 13 and 17,
169 indicating expansion of fewer clones (Supplementary figure 4C). The similarity of the indices between
170 weeks 17 and 19 suggests stabilization of clones. Contribution by the most frequent clone rose ~2.2 times,
171 from 0.078 (±0.005, n=3) at week 13 to 0.174 (±0.030, n=3) at week 17 (Figure 2e), signaling rapid
172 expansion of a few clones. Overall, the clonal repopulation remained normal despite faster expansion and
173 earlier stabilization of human HSPC clones in hu-BLT mice compared to humans, wherein stable clones
174 appear >1 year post-transplant[4]. Using the quantitative data for 897 VIS clones generated by LoVIS-Seq we
175 were able to investigate the polyclonality and stabilization of human HSPC clones in mouse environment.
176 *Clonal sharing between organs reveals normal repopulation and unique tissue distribution pattern:*

177 Our VIS data shows normal repopulation in blood of hu-BLT mice; however, in nonhuman primates, the
178 early post-transplant clonal expansion patterns differ between blood and organs[25]. We performed VIS
179 analysis on bulk cells from bone marrow (BM) and spleens of our hu-BLT mice to investigate whether the
180 tissue/organ clonal expansion pattern differed from blood. We found a very similar clonal expansion pattern
181 (avg. Pearson's r =0.94) between blood and spleen (Figure 3); however, clonal expansion in bone marrow
182 differed from blood and spleen. Interestingly, we observed that in all three tissue compartments, persistent
183 clones contributed the most to repopulation, showing significant clonal sharing among the three tissue
184 compartments. These results show that LoVIS-Seq enables tracing of hundreds of persistent clones and their
185 contribution in blood relative to other tissue compartments.
186 *Influence of genomic location and proximal genes on clonal growth:*
187 For each VIS, our assay provided both relative frequency and genomic location of vector integration. Similar
188 to the HIV-1 integration pattern[14], our VIS data showed preference for high gene density chromosomes
189 (Supplementary figure 5A). Additionally, similar genomic distribution of low, medium, and high frequency
190 VIS clones (Figure 4a) suggests clonal expansion is unrelated to genomic location of integration. We found
191 that ~80% of the VIS occurred within ±1Kb of protein-coding genes, significantly higher than 46% of 1000
192 random IS ($p < 0.001$, Supplementary figure 6A). About 8% of VIS were within ±1Kb of long non-coding
193 RNA (lncRNA) and 10% were outside ±1Kb of any genes (distal VIS, Figure 4b inside pie chart). Persistent
194 and top 10 VIS clones (top 10 most frequent VIS clones from each timepoint) showed similar preference for
195 gene biotypes (Figure 4B inside pie chart). Out of all VIS, only 66 (including 1 out of 42 top 10 VIS) were
196 proximal to known cancer consensus genes (Figure 4a Circos plot). These results indicate clonal stability or
197 growth is unrelated to the biological function of proximal genes.
198 *Integration bias for transcriptionally active genes in human fetal liver HSPC*:
199 Our data shows VIS preference for genic regions; intriguingly, the transcriptional state of these VIS-
200 proximal genes in uncultured human FL-HSPC is unknown. To address this, we analyzed transcriptomic
201 (RNA-seq) and functional genomic (ATAC-seq and ChIP-seq) data from uncultured human FL-HSPC[26].
202 Processing of these cells is identical to the FL-HSPC used in this study (see Methods). The RNA-seq data
203 shows that >77% of all VIS, including the persistent and top 10 VIS, are within ±1Kb of transcriptionally
204 active genes (FPKM>1) (Figure 4b outer donut chart). This is significantly higher than the ~27% of random
205 IS proximal to active genes ($p < 0.001$, Supplementary figure 6A). The level of transcriptional activity of
206 VIS-proximal genes (based on the FPKM value) was slightly higher than the median expression level of all
207 active genes (Supplementary figure 5B). Moreover, the median gene activity level (FPKM) varied based on
208 gene biotype, with protein coding proximal genes of all, persistent, and top 10 VIS clones having higher
209 activity than lncRNA or pseudogenes of proximal genes. Overall, our analysis shows that VIS prefer active
210 genes but not highly active genes.

211 *Vector integration favors H3K36me3 enriched regions:*
212 We speculate that the chromatin structure of active genes strongly influences vector integration. To
213 investigate this, we analyzed functional genomic data for 10 different chromatin features in uncultured FL-
214 HSPC[26] (listed in Figure 5). These features include chromatin accessibility (ATAC-seq), active RNA
215 polymerase II, and 8 histone modifications (ChIP-seq). For VIS-proximal active genes (FPKM>1), we found
216 active transcription markers such as open chromatin region (ATAC-seq peaks) and histone marks (e.g.
217 H3K4me3) near TSS as well as H3K36me3 and H3K79me2-enriched regions within the gene body.
218 However, these marks were less prominent in inactive (FPKM<1) VIS-proximal genes (Figure 5a profile
219 plots and heatmaps). Random IS-proximal genes showed similar enrichment profiles in active genes and
220 their lack in inactive genes (Supplementary figure 6B). Further analysis showed that 420 VIS (46% of total
221 VIS) were within actively transcribed regions marked by histone-modification marks H3K36me3 and/or
222 H3K79me2. This is significantly higher compared to only 10% random IS within actively transcribed regions
223 ($p<0.001$, Supplementary figure 6C). Principal component analysis (PCA) on normalized enrichment levels
224 (RPKM values) over a ±1Kb region of VIS (Figure 5b data) clearly separated H3K36me3 and H3K79me2
225 from all other features and showed no bias for random IS (Supplementary figure 7). Importantly, VIS
226 avoided open DNA and transcription-regulator histone marks H3K4me3 and RNApolII as well as repression
227 marker H3K27me3 (Figure 5b and 5c top panel). The top 10 most frequent VIS clones also displayed similar
228 preference for all chromatin features (Figure 5c middle panel). Comparatively, random IS were evenly
229 distributed across all 10 histone marks (Figure 5c bottom panel). Although our analysis reveals a significant
230 vector integration bias for H3K36me3 and/or H3K79me2 enriched regions, the clonal expansion of human
231 HSPC in hu-BLT mice remained normal. Overall, LoVIS-Seq assay provided both the genomic location of
232 each VIS clone and its contribution to the repopulation which revealed determinants of VIS selection and
233 that clonal behavior was unrelated to genomic location of vector integration.
234
235 **Discussion:**
236 In this study we presented LoVIS-Seq, a new longitudinal clonal tracking method requiring a mere 25µl
237 blood or less to monitor clonal behavior of gene-modified cells in small-animal models. LoVIS-Seq
238 quantitatively captured the clonality of both control (mCherry-H1/H5) and anti-HIV (EGFP-WT)
239 populations in whole blood. We provide the first-ever demonstration of simultaneous polyclonal
240 repopulation of therapeutic vector-modified and control vector-modified cell populations in hu-BLT mouse
241 blood. The polyclonal expansion resembles normal post-transplant HSPC clonal repopulation in mice[6],
242 nonhuman primates[5,20] and humans[4,21]. Notably, the clonal frequency data recapitulated the flow cytometric
243 measurements. Persistent clones are major contributors in blood, BM, and spleen. In uncultured FL-HSPC,
244 VIS bias toward H3K36me3 and/or H3K79me2 enriched regions. Significantly, this VIS bias appears

245 inconsequential with respect to clonal repopulation, as gene-modified HSPC differentiated normally in vivo;
246 this confirms the safety of our therapeutic and control lentiviral vectors.
247 We recently showed that in hu-BLT mice, monitoring clonal expansion of gene-modified and control
248 populations within the same animal gives an unbiased analysis[8]. In the current study, we longitudinally
249 tracked hundreds of HPSC clones in both anti-HIV gene-modified and control vector-modified populations
250 in the same hu-BLT mouse. Interestingly, we found a competitive growth pattern between two populations,
251 with few clones from each population leading the expansion. Previous studies propose that in myeloablated
252 mice, hematopoiesis tends to stabilize around 22 weeks post-transplant[7,27] while a recent study suggested 16
253 to 24 weeks[28]; our data indicates clonal stabilization between 17 to 19 weeks post-transplant. Overall, the
254 clonal repopulation of human HSPC in hu-BLT mice resembles that of mouse HSPC after autologous
255 transplant. Cord blood HSPC transplanted in NGS mice also showed similar clonal behavior, with clonal
256 stabilization starting near week 18 to 20 post-transplant[29,30]. Although the timescales compare well with
257 other studies, caution is due considering high incidence of graft versus host-disease-related illnesses in
258 xenograft mouse models.
259 VIS preference for transcriptionally active genes is also present in FL-HSPC. However, RNA-Seq data
260 indicates bias against highly expressed genes due to either obstruction of integration by transcriptional
261 machinery or detrimental effect of integration on survival of the cell. Investigation of 10 chromatin features
262 showed strong VIS bias toward actively transcribed regions marked by histone modifications H3K36me3
263 and H3K79me2. This bias could be attributed to LEDGF/p75, a chromatin binding protein essential for
264 efficient HIV-1 integration[31], that binds to integrases of HIV-1[32,33] and protects the pre-integration complex
265 from degradation[34]. Importantly, the N-terminal PWWP domain of LEDGF is known to preferentially
266 interact with H3K36me3[35,36]. H3K79me2 and H3K36me3 mark the gene body[37] and H3K36me3 marks
267 exons and is positioned near the 5` end of the exon and is correlated with alternative splicing[38,39]. Thus, VIS
268 in proximity of H3K36me3 are likely to influence co-transcriptional splicing of the proximal gene as well as
269 expression of the vector itself. However, tracking of hundreds of single HSPC clones suggests that such
270 integration has miniscule to no impact on the stemness of repopulating vector-modified HSPC.
271 A recent study demonstrated use of CRISPR/Cas to introduce barcodes in the long-term HSPC and
272 longitudinally tracked a very limited number of HSPC clones[30]. Comparatively, using LoVIS-Seq we have
273 tracked ~10 times more HSPC clones per animal with high accuracy and reproducibility. It is pertinent to
274 note that to enable insertion of barcoded donor DNA into the host genome, HSPC need to undergo in vitro
275 preconditioning and incubation before transplantation. The double stranded breaks introduced by
276 CRISPR/Cas activate DNA damage responses causing significant delays in HSPC proliferation and affects
277 their in vivo repopulation[40]. Additionally, off-target gene-editing by CRISPR/Cas remains a major concern.
278 In contrast, LoVIS-Seq does not require preconditioning of HSPC and provides a ready to use high-

279 throughput clonal tracking assay for small-animal models. Furthermore, LoVIS-Seq has wider applicability
280 as it can be tailored to work with many different lentiviral vectors commonly used to insert transgenes into
281 the genome of host cells.
282
283 LoVIS-Seq with whole genome amplification allows for quantitative assessment of clonal repopulation in
284 small-animal models. However, the accuracy and reproducibility of our assay depends on the initial number
285 of cells used for MDA (Figure 1d). To minimize sampling errors, it is important to have a sufficient number
286 of gene-marked human cells in each 25µl of blood or 10,000 cells to represent each clone in similar
287 proportions as in the bulk population. Higher human reconstitution and gene marking are often desirable and
288 necessary conditions wherein our assay provides optimal results.
289 Overall, using a mere 25µl blood and LTR-indexed vectors, we explored polyclonal expansion in both
290 control and therapeutic vector-modified populations in the same hu-BLT mice. LoVIS-Seq revealed the in
291 vivo dynamics of clonal expansion, emergence of stable stem cell clones, and consequences of vector
292 integration bias on repopulation. Thus, LoVIS-Seq provides an efficient tool for multifaceted analysis of
293 clonal dynamics in murine and humanized-mouse models that are used extensively in HIV, cancer, gene-
294 therapy, and stem cell research.
295
296 **Methods:**
297 *Human fetal thymus and isolation of FL-CD34+ cells from fetal tissue*
298 Human fetal thymus and livers were obtained from Advanced Bioscience Resources (ABR) and the UCLA
299 CFAR Gene and Cellular Therapy Core. Human fetal liver CD34+ HSPC and thymus pieces were processed
300 as previously described[41]. Briefly, a single cell suspension of fetal liver cells was strained through 70 µm
301 mesh and layered onto density gradient separation media (Ficol Paque PLUS, GE Healthcare). After 20
302 minutes of centrifugation, the mononuclear cells layer was collected. Anti-CD34+ microbeads (Miltenyi
303 Biotech) were used for magnetic isolation of CD34+ cells from mononuclear cells. Calvanese et al.[26] also
304 obtained fetal liver from the UCLA CFAR Gene and Cellular Therapy Core and followed identical CD34+
305 magnetic sorting to isolate uncultured FL-HSPC for RNA-seq, ATAC-seq, and ChIP-seq assay.
306 *Humanized BLT mouse and sample collection*
307 NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ (NSG) mice, 6-8-week-old, were myeloablated 1 day before transplant by
308 intraperitoneal (i.p.) injection with 10 mg/kg of 6-thioguanine (6TG) (Sigma-Aldrich, Saint Louis, MO) or
309 35mg/kg of Busulfan for mouse m860. Myelo-preconditioned mice were transplanted with human fetal liver
310 CD34+ HSPCs transduced with Anti-HIV vector (EGFP+) (0.5×10^6 cells/mouse) and mixed with HSPCs
311 transduced with the control (mCherry+) vector (0.5×10^6 cells/mouse). The mice were transplanted with a
312 two-step procedure: half the mixture of EGFP+ and mCherry+ transduced cells was solidified by matrigel

313 (BD Bioscience, San Jose, CA), mixed with CD34- cells as feeder cells ($4.5x10^6$ cells), and implanted with a
314 piece of human thymus under the mouse kidney capsule. Then, mice were injected with the other half of the
315 mixed EGFP+ and mCherry+ transduced cells via retro-orbital vein plexus using a 27-gauge needle on the
316 same day. Bone marrow cells for MDA were harvested from mouse m860 at week 25 post-transplant. For
317 longitudinal clonal tracking and to monitor human leukocyte reconstitution and percentage of the EGFP+
318 and mCherry+ marked cells, 100µl of mouse blood was collected from the retro-orbital vein every two
319 weeks from weeks 13-19 post-transplant. Plasma was removed and peripheral blood cells were stained with
320 monoclonal antibodies for 30 minutes. Red blood cells were lysed with red blood cell lysis buffer (4.15 g of
321 NH4Cl, 0.5 g of KHCO3, and 0.019 g of ethylenediaminetetraacetic acid in 500 mL of H2O) for 10 minutes
322 and washed with FACS buffer (2% fetal calf serum in phosphate-buffered saline [PBS]). Stained cells were
323 resuspended in 16µl PBS, of which 8µl was split equally into two tubes for MDA replicates. The remaining
324 8µl was mixed with 300µl of 1% formaldehyde in PBS and examined with Fortessa (BD Biosciences) flow
325 cytometers. Flow cytometry data was utilized to monitor human reconstitution (Supplementary table 2) and
326 count human cells, mCherry+ cells, and EGFP+ cells as well as human T and B cells in blood
327 (Supplementary figure 2). The following monoclonal antibodies with fluorochromes were used: human
328 CD45-eFluor 450 (HI30, eBioscience), CD3-APC-H7 (SK7: BD Pharmingen), and CD19-BV605 (HIB19:
329 BioLegend). Data were analyzed on FlowJo (TreeStar, Ashland, OR) software.
330 The UCLA Institutional Review Board has determined that fetal tissues from diseased fetuses obtained
331 without patient identification information are not human subjects. Written informed consent was obtained
332 from patients for use of these tissues for research purposes. All mice were maintained at the UCLA Center
333 for AIDS Research (CFAR) Humanized Mouse Core Laboratory in accordance with a protocol approved by
334 the UCLA Animal Research Committee.
335
336 *LoVIS-Seq workflow with whole genome amplification and quantitative VIS assay*
337 Multiple displacement amplification for whole genome amplification: To estimate the minimum number of
338 cells required for LoVIS-Seq, we collected 81,000, 27,000, 18,000, 9,000, 3,000, and 1,000 bone marrow
339 cells from mouse m860 by serial dilution and stored in 4µl of PBS at -20ºC. MDA was done directly on cells
340 using the REPLI-g Single Cell Kit from Qiagen (Cat #150343) following kit-specific protocol. For
341 longitudinal clonal tracking in the blood compartment, 100 µl blood was drawn at weeks 13, 15 & 17. At end
342 point (week 19), max blood (≈ 1ml) was collected, out of which 100µl was used for flow cytometric analysis
343 along with MDA; the remainder was used to isolate unamplified whole blood DNA. Cells for MDA were
344 isolated as described above and stored in 4µl of PBS at -20ºC. MDA-amplified DNA was then used for
345 quantitative VIS assay. A Qiagen DNeasy Blood & Tissue Kit was used to extract unamplified DNA from
346 max blood cells, splenocytes, and bone marrow cells.

347

348 Quantitative VIS assay and data analysis workflow: For VIS sequencing, we followed the procedures
349 described in our previous publication[5,8,42,43] and focused on analyzing only the right LTR junctions using
350 CviQI and RsaI restriction enzymes. For our VIS assay we used one microgram MDA-amplified or
351 unamplified genomic DNA for animal m860 samples and two micrograms MDA-amplified or unamplified
352 genomic DNA for different time point samples, with a few exceptions (see Supplementary Tables 1 and 2).
353 DNA samples were subject to extension PCR using LTR specific biotinylated primers
354 /5BiotinTEG/CTGGCTAACTAGGGAACCCACT 3' and /5BiotinTEG/CAGATCTGAGCCTGGGAGCTC
355 3'. The extension PCR product was then digested using CviQI and RsaI restriction enzymes and biotin
356 primer bound DNAs isolated using streptavidin-agarose Dynabeads using magnetic separator as per
357 manufactures instructions. The vector-host junctions capture on streptavidin beads were processed for linker-
358 mediated PCR (LM-PCR) methods as described previously[42,43]. The linker ligated vector-host junction DNA
359 was subjected to two step PCR. First step amplification was done using primer 5'
360 CTGGCTAACTAGGGAACCCACT 3' and first linker primer GTGTCACACCTGGAGATAT. We
361 removed the internal vector sequence by restriction enzyme (SfoI) digestion. The digested product of first
362 PCR was then amplified using primer 5'ACTCTGGTAACTAGAGATCC 3' and second linker primer 5'
363 GGAGATATGATGCGGGATC 3'. Since the LTR index sequence is included in the vector-host junction
364 the we obtain unbiased amplification all the H1, H5 and/or WT VIS sequences. Lentiviral vectors used in
365 this study as derived from FG12-mCherry lentiviral vector[41] and all the primers are designed accordingly. A
366 detailed protocol for VIS assay is provided in supplementary text. The amplicon libraries prepared using
367 custom made Illumina sequencing primers for Illumina MiSeq (m860 samples) or iSeq100 (m599, m599,
368 and m591 samples) sequencer. Sequences with a virus-host junction with the 3' end LTR, including both the
369 3'-end U5 LTR DNA and ≥ 25 base host DNA (with ≥ 95% homology to the human genome), were
370 considered true VIS read-outs. The sequence mapping and counting method was performed as described
371 previously[8]. In brief, sequences that matched the 3'end LTR sequence joined to genomic DNA as well as
372 LTR-indexes (H1, H5 or WT) were identified using a modified version of SSW library in C++[44]. Reads
373 were classified as H1, H5, or WT VIS based on the LTR barcodes used in the experiment. VIS sequences
374 were mapped onto the human genome (Version hg38 downloaded from https://genome.ucsc.edu/) using
375 Burrows-Wheeler Aligner (BWA) software. Mapped genomic regions were then used as reference and VIS
376 reads were remapped using BLAST to further remove poorly mapped reads to get an accurate estimate of
377 sequence count. Final VIS counting was done after correcting for VIS collision events and signal crossover
378 as described previously. VIS with a final sequence count less than the total number of samples analyzed per
379 animal were removed. VIS clones with maximum frequency values below 1st quartile were classified as "low
380 frequency", clones with maximum frequency value above 3rd quartile were classified as "high frequency",

381 and clones with maximum frequency between the 1st and 3rd quartiles were designated "medium frequency".
382 VIS clones that were detected with frequency >0 at every week from 13-19 are termed "persistent clones".
383 The 10 high frequency VIS clones at each timepoint were selected as top 10 VIS. All the VIS data and list of
384 VIS-proximal genes is provided in supplementary file.

385 *Random integration sites*

386 Random integration sites were generated in silico using a custom python script. To mimic our VIS assay, we
387 randomly selected 1000 integration sites that were within ±1500bp of the nearest CviQI/RsaI (GTAC) site in
388 the human genome (hg38).

389 *Clonal diversity analysis*

390 For diversity analysis, we used Rényi's diversity/entropy[22] of order α defined as follows

391 $$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum p_i^\alpha\right),$$

392 where p_i is the proportional abundance of the ith VIS clone for $i = 1, \ldots, n$. At each timepoint, an average
393 Rényi's diversity profile was obtained by calculating average values of $\boldsymbol{H_\alpha}$ for $\alpha \geq 0$. The α is considered as
394 a weighting parameter such that increasing α leads to increased influence of high frequency VIS clones. The
395 proportional abundance is calculated as $p_i = s_i/S$, where s_i is the sequence count of the ith VIS clone and S
396 is the sum of sequence counts from all VIS clones. The Rényi's diversity $\boldsymbol{H_\alpha}$ values are averaged over two
397 replicates and plotted as a function of α. If all VIS clones contributed equally, i.e. $p_i = \frac{1}{n}$ for all $i = 1, \ldots, n$,
398 then $\boldsymbol{H_\alpha}$ for all values of α would be equal and the profile (line) would be horizontal. VIS clones expanding
399 at different rates would show decreasing $\boldsymbol{H_\alpha}$ values as α increases, generating a downward-sloped diversity
400 profile that is steeper with more non-uniform clonal expansion. $\boldsymbol{H_\alpha}$ indicates clonal diversity of the
401 repopulating cells, such that consistently higher values of $\boldsymbol{H_\alpha}$ indicate a more diverse clonal population. If
402 the profiles for two populations/samples cross, then their relative diversities are similar. For $\alpha = 0$, $\boldsymbol{H_0} =$
403 $\log(n)$ and the antilogarithm of this value equates to the richness or number of unique IS. $\boldsymbol{H_\alpha}$ at $\alpha = 1$ and
404 $\alpha = 2$ are the Shannon and 1/Simpson indexes, respectively. We calculated Renyi's diversity using the R
405 package BiodiversityR (*https://cran.r-project.org/web/packages/BiodiversityR/index.html*). For the above
406 analysis, we used raw sequence counts from two replicates without distinguishing between mCherry-H5 VIS
407 and EGFP-WT VIS.

408 *RNA-seq data analysis*

409 Raw sequence data of uncultured FL-HSPC (in triplicate) was pre-processed for quality using Fastqc.
410 Trimmomatic was used to remove adaptors and for quality trimming. After this, reads were aligned onto
411 human genome hg38 using RNA STAR aligner[45]. SAMtools was used to remove reads with low mapping
412 scores (< 20) and to generate BAM files. Cufflinks[46] was used to calculate FPKM values for all genes. The

413 human cancer consensus gene list is from Catalogue of Somatic Mutations In Cancer
414 (https://cancer.sanger.ac.uk/census).
415 *ATAC-seq analysis*
416 Raw sequence data of uncultured FL-HSPC (in triplicate) was pre-processed for quality using Fastqc.
417 Adaptor removal and quality trimming was done using Trimmomatic. After this, reads were mapped onto
418 human genome hg38 using bowtie2 with parameter --very-sensitive -X 2000 -k 1. SAMtools was used to
419 remove reads with low mapping (<20) scores, blacklisted regions[47], and to generate BAM files. Picard tool
420 kit was used to remove duplicate reads. We used Genrich, a paired end peak caller, to identify ATAC peaks.
421 Software deepTools[48] was used to generate coverage (.bw) files and for visualization of open DNA in genes
422 and VIS-proximal regions.
423 *ChIP-seq data analysis*
424 Raw sequence data of uncultured FL-HSPC for histones, RNAploII, and input were pre-processed for quality
425 using Fastqc. Trimmomatic was used to remove adaptors and for quality trimming. After this, reads were
426 mapped onto human genome hg38 using bowtie2 with parameter --local. SAMtools was used to remove
427 reads with low mapping (<20) scores, blacklisted regions[47], and to convert SAM to BAM format. Picard tool
428 kit was used to remove duplicates. MACS2 tool was used to call peaks for all histone marks and RNApolII
429 using input sample as control. Software deepTools[48] was used to generate coverage .bw files and for
430 visualization of histone/RNApolII in genes and VIS proximal regions.
431 Raw RNA-seq, ATAC-seq, ChIP-seq data of uncultured FL-HSPC from published reference is available in
432 Gene Expression Omnibus (GEO) with the accession code GSE111484[26].
433
434 *Statistical analysis*
435 Clonal frequencies are summarized as means ± standard deviations (SDs). Pearson correlations are used to
436 compare the similarity and reproducibility of clonal profiles between two samples and replicates,
437 respectively. Pearson's r values and p values are calculated using statistical software R (version 3.6,
438 https://www.r-project.org/). To determine if VIS preference for genomic and epigenetic features differs
439 significantly from random IS, we used Pearson's chi-squared test with Yate's continuity correction (function
440 chisq.test() in software R). We use Principle component analysis (PCA) to reduce the complexity of read
441 coverage data of multiple chromatin feature in proximity to VIS. The dimensionality reduction by PCA
442 method is similar to clustering and allows detection of patterns in the data. In this study, PCA was done
443 using software deepTools[48].
444
445 **Acknowledgements:** We thank the technical support from the UCLA Center for AIDS research (CFAR)
446 research cores including the CFAR Gene and Cellular Therapy Core and the CFAR Humanized Mouse Core.

We would like also to thank Anna Sahakyan and Ruth Cortado for their help with making lentiviral vectors, transducing human CD34+ HSPC, and performing humanized mouse experiments. **Funding:** This work was supported by grants from the NIH (5U19AI117941, AI110297, AI145038, HL125030, HL126544), CIRM (DR1-01431, TRX-01431-1), the James B. Pendleton Charitable Trust, and the McCarthy Family Foundation (I.S.Y.C).

Competing interests: Dr. Irvin S.Y. Chen has a financial interest in CSL Behring and Calimmune Inc. No funding was provided by these companies to support this work; Dr. Dong Sung An has a financial interest in Calimmune Inc and CSL Behring that the University of California Regents have licensed intellectual property invented by Dong Sung An, that is being used in the research, to Calimmune Inc. No funding was provided by these companies to support this work. All other authors declare no competing interests.

Reference:

1 Shultz, L. D., Ishikawa, F. & Greiner, D. L. Humanized mice in translational biomedical research. *Nat Rev Immunol* **7**, 118-130, doi:10.1038/nri2017 (2007).

2 Melkus, M. W. *et al.* Humanized mice mount specific adaptive and innate immune responses to EBV and TSST-1. *Nat Med* **12**, 1316-1322, doi:10.1038/nm1431 (2006).

3 Khamaikawin, W. *et al.* Modeling Anti-HIV-1 HSPC-Based Gene Therapy in Humanized Mice Previously Infected with HIV-1. *Mol Ther Methods Clin Dev* **9**, 23-32, doi:10.1016/j.omtm.2017.11.008 (2018).

4 Biasco, L. *et al.* In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. *Cell Stem Cell* **19**, 107-119, doi:10.1016/j.stem.2016.04.016 (2016).

5 Kim, S. *et al.* Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. *Cell Stem Cell* **14**, 473-485, doi:10.1016/j.stem.2013.12.012 (2014).

6 Verovskaya, E. *et al.* Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood* **122**, 523-532, doi:10.1182/blood-2013-01-481135 (2013).

7 Brewer, C., Chu, E., Chin, M. & Lu, R. Transplantation Dose Alters the Differentiation Program of Hematopoietic Stem Cells. *Cell Rep* **15**, 1848-1857, doi:10.1016/j.celrep.2016.04.061 (2016).

8 Suryawanshi, G. W. *et al.* The clonal repopulation of HSPC gene modified with anti–HIV-1 RNAi is not affected by preexisting HIV-1 infection. *Science Advances* **6**, eaay9206, doi:10.1126/sciadv.aay9206 (2020).

9 Esteban, J. A., Salas, M. & Blanco, L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* **268**, 2719-2726 (1993).

10 Paez, J. G. *et al.* Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* **32**, e71, doi:10.1093/nar/gnh069 (2004).

11 Bleier, S. *et al.* Multiple displacement amplification enables large-scale clonal analysis following retroviral gene therapy. *J Virol* **82**, 2448-2455, doi:10.1128/JVI.00584-07 (2008).

12 Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest* **129**, 988-998, doi:10.1172/Jci124291 (2019).

13 Patro, S. C. *et al.* Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *P Natl Acad Sci USA* **116**, 25891-25899, doi:10.1073/pnas.1910334116 (2019).

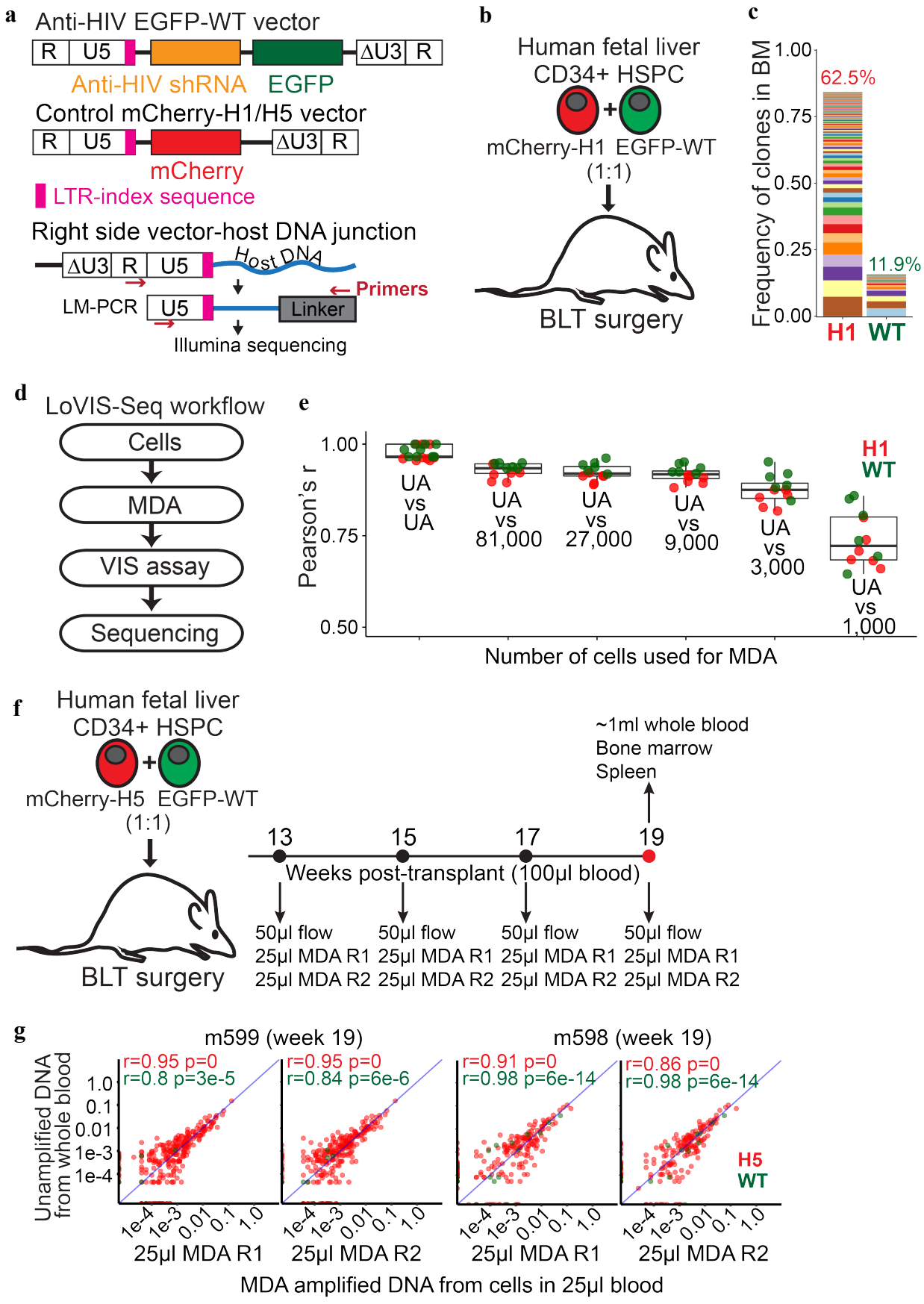
14 Schroder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-529, doi:10.1016/s0092-8674(02)00864-4 (2002).
15 Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**, E234, doi:10.1371/journal.pbio.0020234 (2004).
16 Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* **17**, 1186-1194, doi:10.1101/gr.6286907 (2007).
17 Lelek, M. *et al.* Chromatin organization at the nuclear pore favours HIV replication. *Nat Commun* **6**, doi:ARTN 648310.1038/ncomms7483 (2015).
18 Modlich, U. *et al.* Insertional Transformation of Hematopoietic Cells by Self-inactivating Lentiviral and Gammaretroviral Vectors. *Molecular Therapy* **17**, 1919-1928, doi:10.1038/mt.2009.179 (2009).
19 Montini, E. *et al.* The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. *J Clin Invest* **119**, 964-975, doi:10.1172/Jci37630 (2009).
20 Wu, C. *et al.* Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* **14**, 486-499, doi:10.1016/j.stem.2014.01.020 (2014).
21 Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **341**, 1233151, doi:10.1126/science.1233151 (2013).
22 Rényi, A. On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (1961).
23 Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst Tech J* **27**, 379-423, doi:DOI 10.1002/j.1538-7305.1948.tb01338.x (1948).
24 Simpson, E. H. Measurement of Diversity. *Nature* **163**, 688-688, doi:DOI 10.1038/163688a0 (1949).
25 Wu, C. *et al.* Geographic clonal tracking in macaques provides insights into HSPC migration and differentiation. *J Exp Med* **215**, 217-232, doi:10.1084/jem.20171341 (2018).
26 Calvanese, V. *et al.* MLLT3 governs human haematopoietic stem-cell self-renewal and engraftment. *Nature* **576**, 281-286, doi:10.1038/s41586-019-1790-2 (2019).
27 Lu, R., Czechowicz, A., Seita, J., Jiang, D. & Weissman, I. L. Clonal-level lineage commitment pathways of hematopoietic stem cells in vivo. *Proc Natl Acad Sci U S A* **116**, 1447-1456, doi:10.1073/pnas.1801480116 (2019).
28 Rodriguez-Fraticelli, A. E. *et al.* Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* **583**, 585-589, doi:10.1038/s41586-020-2503-6 (2020).
29 Belderbos, M. E. *et al.* Donor-to-Donor Heterogeneity in the Clonal Dynamics of Transplanted HumanCord Blood Stem Cellsin Murine Xenografts. *Biol Blood Marrow Transplant* **26**, 16-25, doi:10.1016/j.bbmt.2019.08.026 (2020).
30 Ferrari, S. *et al.* Efficient gene editing of human long-term hematopoietic stem cells validated by clonal tracking. *Nat Biotechnol*, doi:10.1038/s41587-020-0551-y (2020).
31 Llano, M. *et al.* An essential role for LEDGF/p75 in HIV integration. *Science* **314**, 461-464, doi:10.1126/science.1132319 (2006).
32 Vanegas, M. *et al.* Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin tethering. *J Cell Sci* **118**, 1733-1743, doi:10.1242/jcs.02299 (2005).
33 Cherepanov, P. *et al.* Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. *Nat Struct Mol Biol* **12**, 526-532, doi:10.1038/nsmb937 (2005).
34 Llano, M., Delgado, S., Vanegas, M. & Poeschla, E. M. Lens epithelium-derived growth factor/p75 prevents proteasomal degradation of HIV-1 integrase. *J Biol Chem* **279**, 55570-55577, doi:10.1074/jbc.M408508200 (2004).
35 Eidahl, J. O. *et al.* Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. *Nucleic Acids Res* **41**, 3924-3936, doi:10.1093/nar/gkt074 (2013).

541 36 Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R. & Bickmore, W. A. Psip1/Ledgf p52
542 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative
543 splicing. *PLoS Genet* **8**, e1002717, doi:10.1371/journal.pgen.1002717 (2012).
544 37 Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone
545 modification regions in humans. *Nat Struct Mol Biol* **17**, 1495-1499, doi:10.1038/nsmb.1924 (2010).
546 38 Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the
547 human genome. *PLoS Comput Biol* **5**, e1000566, doi:10.1371/journal.pcbi.1000566 (2009).
548 39 Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000,
549 doi:10.1126/science.1184208 (2010).
550 40 Schiroli, G. *et al.* Precise Gene Editing Preserves Hematopoietic Stem Cell Function following
551 Transient p53-Mediated DNA Damage Response. *Cell Stem Cell* **24**, 551-565 e558,
552 doi:10.1016/j.stem.2019.02.019 (2019).
553 41 Shimizu, S. *et al.* A highly efficient short hairpin RNA potently down-regulates CCR5 expression in
554 systemic lymphoid organs in the hu-BLT mouse model. *Blood* **115**, 1534-1544, doi:10.1182/blood-
555 2009-04-215855 (2010).
556 42 Kim, S. *et al.* High-throughput, sensitive quantification of repopulating hematopoietic stem cell
557 clones. *J Virol* **84**, 11771-11780, doi:10.1128/JVI.01355-10 (2010).
558 43 Suryawanshi, G. W. *et al.* Bidirectional Retroviral Integration Site PCR Methodology and
559 Quantitative Data Analysis Workflow. *J Vis Exp*, doi:10.3791/55812 (2017).
560 44 Zhao, M., Lee, W. P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith-Waterman C/C++
561 library for use in genomic applications. *PLoS One* **8**, e82138, doi:10.1371/journal.pone.0082138
562 (2013).
563 45 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21,
564 doi:10.1093/bioinformatics/bts635 (2013).
565 46 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with
566 TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).
567 47 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic
568 Regions of the Genome. *Sci Rep* **9**, 9354, doi:10.1038/s41598-019-45839-z (2019).
569 48 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis.
570 *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

571
572
573
574
575
576
577
578
579
580
581
582
583
584

585 **Figures-**

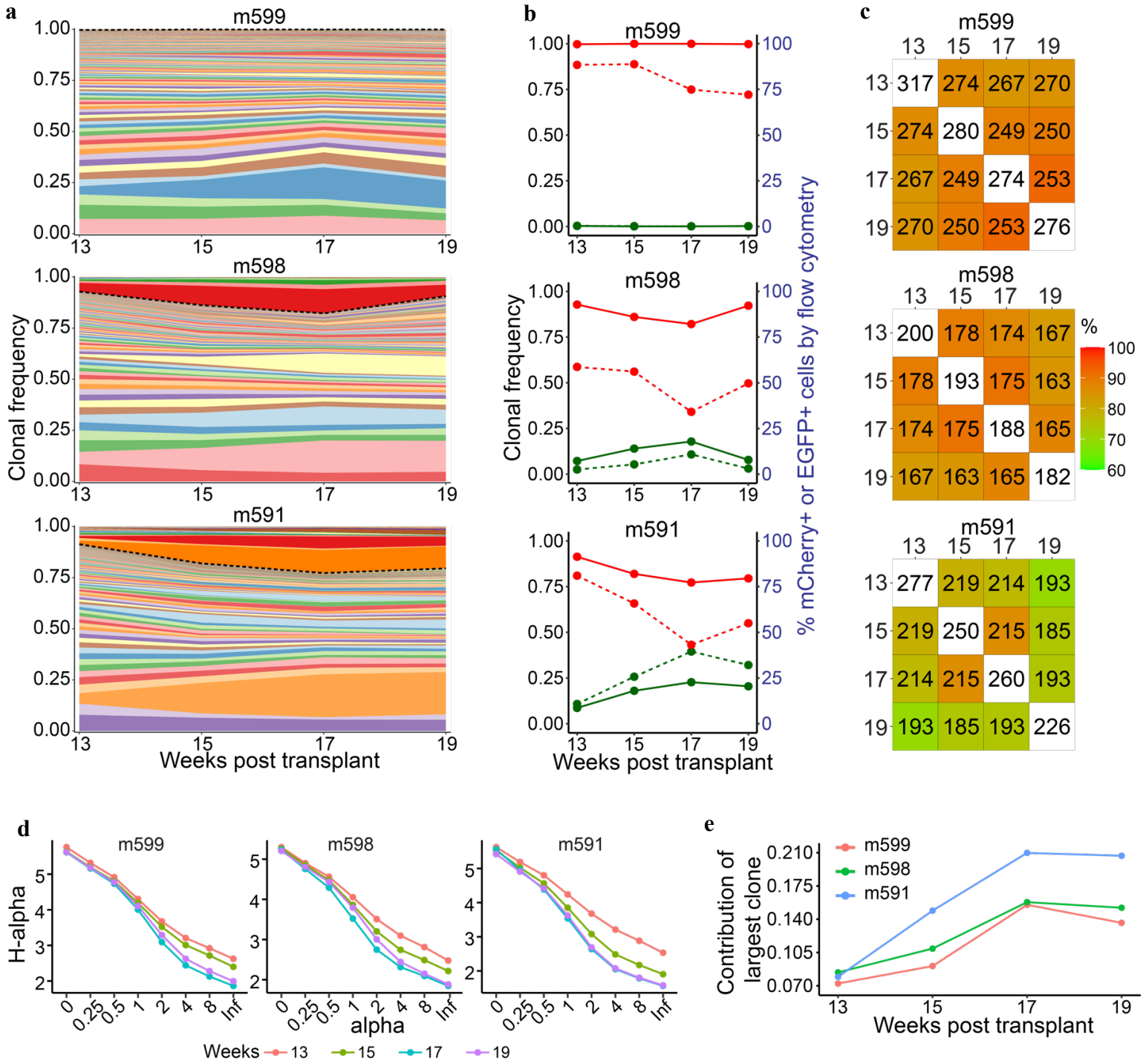
586 Figure 1



587

588 **Figure 1: LoVIS-Seq reproduces clonal distribution of entire mouse blood using 25µl blood: a**)
589 Diagram showing Anti-HIV-EGFP-WT and control mCherry-H1/H5 vectors having WT, H1, or H5 LTR-
590 index and strategy for VIS assay with LTRi-seq. **b**) Hu-BLT mouse model: Fetal liver CD34+ cells were
591 separately transduced with either anti-HIV or control vectors and transduced cells were mixed 1:1 for
592 transplant. The mix of transduced cells was transplanted in myeloablated NSG mice with a fetal thymus
593 tissue implant. **c**) Stacked bar plot showing clonal frequencies of VIS in BM of hu-BLT mouse. Clones from
594 mCherry-H1 and EGFP-WT cells were identified by corresponding LTR barcodes. In the stacked bar plot,
595 each band represents a unique VIS (HSPC clone) and band thickness represents clonal frequency or
596 abundance of that HSPC clone. Percentage of mCherry+ or EGFP+ cells within human cell (hCD45+)
597 population are shown on top of the corresponding stacked-bar. **d**) LoVIS-Seq workflow. **e**) Plot showing
598 Pearson's r for correlations of mCherry-H1 (red dots) and EGFP-WT (green dots) VIS clonal profiles
599 between unamplified DNA replicates and replicates of MDA-amplified DNA samples for different cell
600 numbers. **f**) Experimental protocol for longitudinal clonal tracking in humanized BLT mice. **g**) Scatter plot
601 showing VIS clonal frequencies between unamplified whole blood DNA and two replicates of MDA-
602 amplified DNA from 25µl blood at week 19 (r= Pearson's r, diagonal line is r=1) for m599 and m598. Clonal
603 frequency of mCherry-H5 (red dots) and EGFP-WT (green dots) VIS clones in unamplified DNA samples
604 (y-axis) and MDA replicates (x-axis).
605
606
607

608 Figure 2



609

610 **Figure 2: Longitudinal clonal tracking in hu-BLT mice: a**) Area plots show clonal repopulation in whole
611 blood over time from week 13-19. Each colored band is a unique VIS clone and thickness of the band
612 corresponds to frequency of the VIS clone. Dashed black line separates mCherry-H5 VIS clones (below) and
613 EGFP-WT VIS clones (above). **b**) Line plots show changes over time in the total frequency of mCherry-H1
614 VIS clones (solid red line) and total frequency of EGFP-WT VIS clones (solid green line) as well as
615 percentages of mCherry+ cells (dashed red line) and EGFP+ cells (dashed green line). **c**) Heatmaps showing

percentage change in shared clones between two timepoints. Digits inside white tiles on the diagonal show number of VIS detected at each time point. Colors of each heatmap tile correspond to percentage of clones shared and color key is provided on the right. Digits in each tile show the number of VIS shared between two timepoints. **d**) Renyi's diversity profiles evaluated using raw count data from two replicates at each timepoint and by varying value of alpha. Renyi's diversity profiles are arranged with highest diversity at the top to lowest at the bottom. Topmost curve with no overlap or intersection with any other curve has the highest overall diversity. Diversity of curves that overlap or intersect is undefined. **e**) Line plot showing contribution of highest contributing clone at different timepoints. Values reported are exp(H(alpha)) at alpha=∞.

650 Figure 3

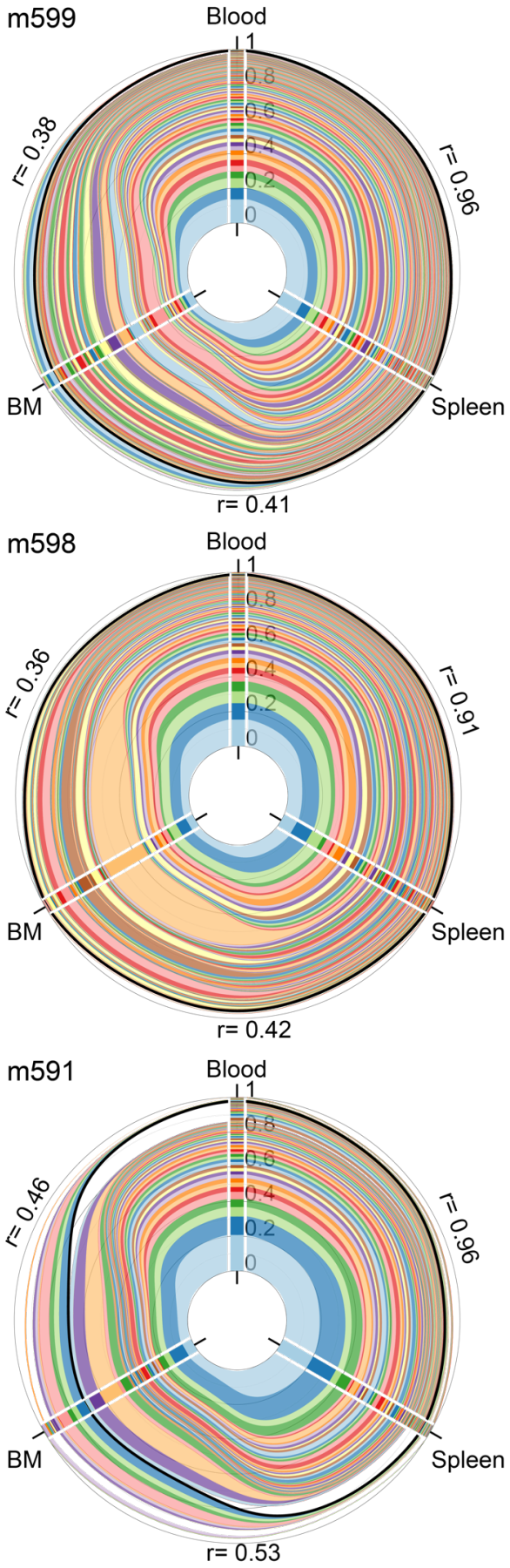
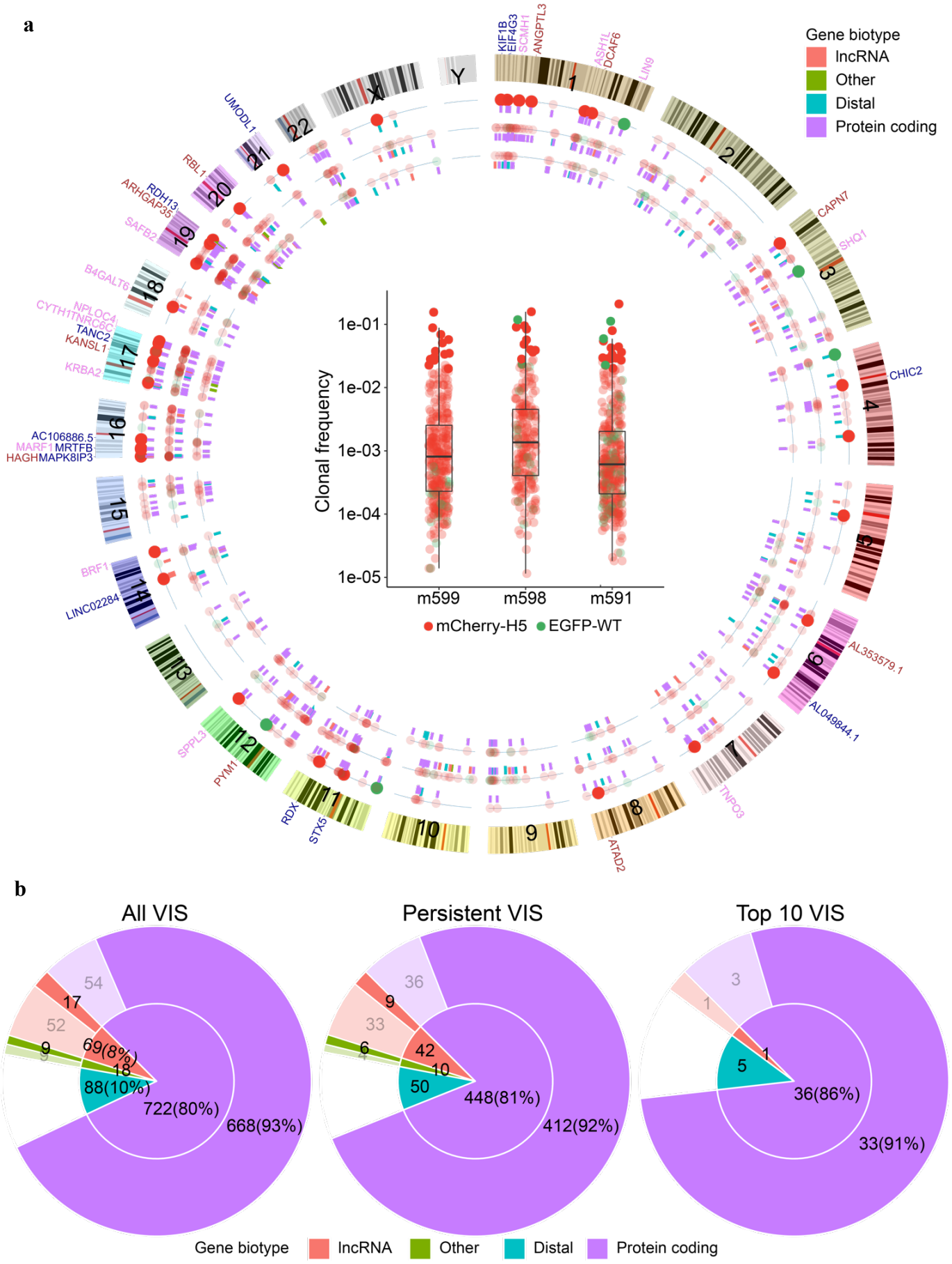

m599
Blood
1
0.8
0.6
0.4
0.2
0
r= 0.38
r= 0.96
BM
Spleen
r= 0.41
m598
Blood
1
0.8
0.6
0.4
0.2
0
r= 0.36
r= 0.91
BM
Spleen
r= 0.42
m591
Blood
1
0.8
0.6
0.4
0.2
0
r= 0.46
r= 0.96
BM
Spleen
r= 0.53

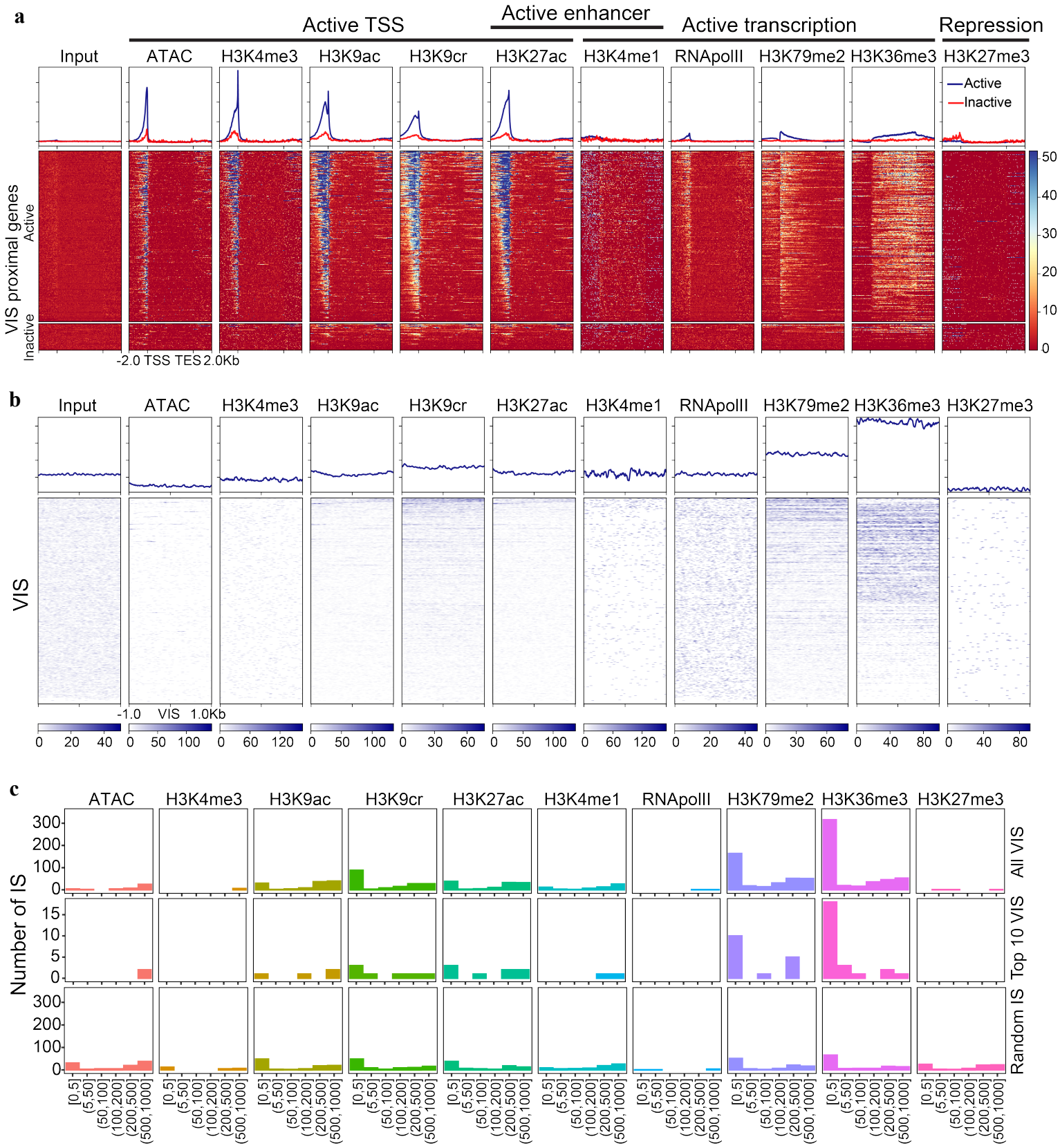
Figure 3: Unique clonal sharing pattern between different tissues: Polar area plots of clonal expansion and sharing in peripheral blood, spleen, and bone marrow (BM). There are three axes, one for each tissue. Stacked bar plot on each axis shows size distribution of clones in the tissue. Each colored stack represents a VIS clone and its thickness shows abundance of the clone. Clones shared between tissues are connected using ribbons with colors matching the clone's stack color in the bar plot. Black line encompasses total size distribution of persistent clones. Pearson's r values are shown in black.

Figure 4


a
Gene biotype
lncRNA
Other
Distal
Protein coding
KIF1B
EIF4G3
SCMH1
ANGPTL3
ASH1L
DCAF6
LIN9
CAPN7
SHQ1
CHIC2
AL353579.1
AL049844.1
TNPO3
ATAD2
RDX
STX5
PYM1
SPPL3
LINC02284
BRF1
AC106886.5
MARF1
MRTFB
HAGH
MAPK8IP3
KRBA2
KANSL1
TANC2
CYTH1
TNRC6C
NPLOC4
B4GALT6
SAFB2
ARHGAP35
RDH13
RBL1
UMODL1
Clonal frequency
1e-01
1e-02
1e-03
1e-04
1e-05
m599
m598
m591
mCherry-H5
EGFP-WT
b
All VIS
Persistent VIS
Top 10 VIS
54
17
52
9
9
69(8%)
18
88(10%)
722(80%)
668(93%)
36
9
33
6
4
42
10
50
448(81%)
412(92%)
3
1
1
5
36(86%)
33(91%)
Gene biotype
lncRNA
Other
Distal
Protein coding

687 **Figure 4: Chromosomal distribution of VIS and its bias for transcriptionally active genes: a**) Circos
688 plot shows genomic location of all 792 mCherry-H5 (red dots) and 105 EGFP-WT (green dots) VIS from
689 mice m599, m598, and m591. Box plots in the center show maximum frequency of mCherry-H5 (red dots)
690 and EGFP-WT (green dots) VIS clones in mice m599, m598, and m591 over 6 weeks. Genomic location of
691 mCherry-H5 (red dots) and EGFP-WT (green dots) VIS clones are plotted on three concentric circles
692 depending on their maximum frequency over 6 weeks: Low frequency clones with maximum frequency
693 below the 1st quartile value (innermost circle), High frequency clones above 3rd quartile value (outermost
694 circle), and Medium frequency clones between 1st and 3rd quartile (middle circle). Top 10 high frequency
695 VIS clones are shown in darker colors. Functional classification of VIS-proximal genes is shown by short
696 line segments, color coded as in the legend. Gene symbols above ideograms represent genes proximal to the
697 top 10 VIS clones from mice m599 (blue), m598 (brown), and m591 (light pink). **b**) Classification of all,
698 persistent, and Top 10 VIS clones based on biotype of proximal gene. Inner pie chart shows clones classified
699 based on gene biotype of the most proximal gene. Outer Donut plots show number of VIS and numbers in
700 bracket show % of VIS proximal to active (dark color) or inactive (faded colors) genes. Active proximal
701 genes have FPKM >1.
702

703 Figure 5



704

705 **Figure 5: Epigenetic determinants of vector integration: a**) Profile plots and heatmap for 10 chromatin

706 features and input sample in active and inactive VIS-proximal genes in uncultured FL-HSPC. Profile plots

707 show mean score for active (blue line) and inactive (red line) proximal genes. Score is calculated from
708 normalized read count (RPKM) for each sample. Each row in heatmap shows expression level of 10
709 chromatin features in proximal genes from TSS to TES with 2Kb flanks upstream and downstream. Color
710 scale key shows range of normalized expression. **b**) Profile plots and heatmaps for 10 chromatin features in
711 region flanking ±1Kb of each VIS. Profile plots showing mean scores over ±1Kb region flanking VIS. Each
712 row in the heatmap shows the expression level of 10 chromatin features in regions flanking ±1Kb of VIS.
713 Individual color scale key shows the range of normalized expression for corresponding features. **c**) Bar plots
714 show number of VIS within ±1Kb of enriched region (peak) of different chromatin features. VIS clones and
715 random IS are binned by absolute distance in base pairs (bp) between the enriched region and IS. Bars show
716 number of VIS in each bin. Top panel shows binning for all VIS, middle panel shows top 10 VIS clones, and
717 bottom panel shows random IS falling within ±1Kb of enriched region.
718