Title:-

Longitudinal clonal tracking in humanized mice reveals sustained polyclonal repopulation of gene-modified human-HSPC despite 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:

Background

Current understanding of hematopoiesis is largely derived from mouse models that are physiologically distant from humans. Humanized mice provide the most physiologically relevant small animal model to study human diseases, most notably preclinical gene-therapy studies. However, the clonal repopulation dynamics of human hematopoietic stem and progenitor cells (HSPC) in these animal models is only partially understood. Using a new clonal tracking methodology designed for small sample volumes, we aim to reveal the underlying clonal dynamics of human cell repopulation in a mouse environment.

Methods

Humanized BLT (bone marrow-liver-thymus) mice were generated by transplanting lentiviral vector transduced human fetal liver HSPC (FL-HSPC) in NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ (NSG) mice implanted with a piece of human fetal thymus. We developed a methodology to track vector integration sites (VIS) in a mere 25µl of mouse blood for longitudinal and quantitative clonal analysis of human HSPC repopulation in mouse environment. We explored transcriptional and epigenetic features of human HSPC for possible VIS bias.

Results

897 HSPC clones were longitudinally tracked in BLT mice—providing a first-ever demonstration of clonal dynamics and competitive expansion of therapeutic and control vector-modified human cell populations simultaneously repopulating in the same humanized mice. The polyclonal repopulation stabilized at 19 weeks post-transplant and the contribution of the largest clone doubled within 4 weeks. Moreover, 550 (~60%) clones persisted over 6 weeks and were highly shared between different organs. The normal clonal profiles confirmed the safety of our gene therapy vectors. Multi-omics analysis of human FL-HSPC revealed that 54% of vector integrations in repopulating clones occurred within ±1kb of H3K36me3-enriched regions.

Conclusions

Human repopulation in mice is polyclonal and stabilizes more rapidly than that previously observed in humans. VIS preference for H3K36me3 has no apparent negative effects on HSPC repopulation. Our study provides a methodology to longitudinally track clonal repopulation in small animal models extensively used for stem cell and gene-therapy research and with lentiviral vectors designed for clinical applications. Results of this study provide a framework for understanding the clonal behavior of human HPSC repopulating in a mouse environment, critical for translating results from humanized mice models to the human settings.

List of abbreviations

75 HSPC: hematopoietic stem and progenitor cells; LoVIS-Seq: low volume vector integration site
76 sequencing; BLT: Bone marrow-liver-thymus; FL: fetal liver; MDA: Multiple displacement amplification.
77
78 **Background:**
79 Hematopoietic stem cells (HSC) are an ideal vehicle for introducing gene-modified cells to treat genetic
80 disorders, cancers, and viral infections. Humanized mouse models—immunodeficient mice transplanted
81 with human stem cells or tissues that generate a functioning human immune system—provide the most
82 practical in vivo system for human stem cell and disease research (reviewed in [1, 2]). In particular,
83 humanized bone marrow-liver-thymus mouse (hu-BLT mouse) models can support the development of
84 human T cells, B cells, monocytes, macrophages, and dendritic cells. Comprehensive analysis of human T-
85 cell development, diversity, and function in hu-BLT mice found intact TCRβ sequence diversity, thymic
86 development, and differentiation to memory and effector cells [3]. The hu-BLT mice demonstrate human
87 MHC-restricted T cell response to Epstein-Barr virus (EBV) infection and human dendritic cells-mediated
88 T cell response against toxic shock syndrome toxin 1 (TSST1) [4]. The hu-BLT mouse models have been
89 used in the field of cancer research to explore treatment strategy for non-Hodgkin's lymphoma [5]. As
90 models for sepsis studies, hu-BLT mice had high levels of human IL-6 in sera after cecal ligation and
91 puncture (CLP)-surgery and are more susceptible to CLP induced sepsis than C57BL/6 mice[6]. A triple
92 knockout (C57BL/6 Rag2$^{-/-}$γc$^{-/-}$CD47$^{-/-}$) -bone marrow, liver, thymus (TKO-BLT) humanized mouse
93 model was able to recapitulate the early in vivo human immune response against Ebola virus (EBOV) and
94 Marburg virus (MARV) [7]. The Human Cytomegalovirus (HCMV) latently-infected hu-BLT mice
95 developed human effector and central memory CD4+ and CD8+ T-cell responses and HCMV specific IgM
96 and IgG B-cell responses [8]. Thus hu-BLT mice provide a valuable animal model system to study virus
97 induced cellular and humoral immune response and to evaluate anti-viral treatments. To mimic anti-HIV
98 gene-therapy in HIV patients, a recent study used an HIV-1 pre-infected hu-BLT mouse model and
99 demonstrated that HIV-1 infection induces selective expansion of anti-HIV-1 dual shRNA gene-modified
100 (protected) CD4+ T cells over control vector-modified unprotected CD4+ T cells [9]. Another recent study
101 utilized HIV-infected hu-BLT mouse models to iteratively test CD28 and 4-1BB costimulation for
102 optimizing anti-HIV CAR T cell therapy [10]. These studies reaffirm that hu-BLT mouse model provides
103 an invaluable in vivo system for cancer, virus, and stem cell research and is particularly indispensable for
104 preclinical gene therapy studies. However, translating results from xenograft mouse models to human
105 setting is challenging as it remains unclear whether the human HSPC in mouse environment exhibit their
106 human traits or clonally behave like mouse HSPC.
107 Understanding the clonal behavior of human cells in mouse environment is essential for interpreting results
108 of humanized mice studies and effect of gene modification on the HSPC repopulation. In humans and

macaques, longitudinal clonal tracking of transplanted HSPC revealed a biphasic expansion: 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 [11, 12]. In mice transplanted with limited number of vector transduced mouse HSC, clonal tracking showed that clones start to stabilize around week 12 post-transplant and progressively fewer clones contribute to the overall repopulation [13]. Clonal tracking in mouse bone marrow also indicated similar time scale for clonal stabilization [14]. Another clonal tracking study in mice suggested transplantation dose-dependent change in HSC differentiation [15]. Clonal tracking in these mouse studies is done using genetic barcodes introduced in HSC using lentiviral vectors. Mice were transplanted with a limited number of mouse HSC that were transduced with lentiviral vector library consisting of individual vectors having a unique sequence tag or barcode. Integration of a barcoded vector in the DNA of host HSC introduces a unique genetic tag that can be used to identify and track individual HSC clones. This genetic barcoding assays enable quantitative clonal analysis and have been useful to study hematopoiesis and post-transplant clonal repopulation. However, generating a well curated barcode library for quantitative clonal tracking is both expensive and time consuming. Moreover, low DNA availability, lack of a universal barcode counting method, small barcode library size and other limitations [16] affect the accuracy and usability of vector barcoding techniques. Inability to identify genomic location of inserted barcode limits the use of genetic barcoding techniques to detect insertional mutagenesis induced clonal expansion.

Vector integration site (VIS) assays take advantage of a unique vector-host DNA junction sequence in each transduced HSPC due to the vector randomly integrating into host genome of these cells. A high-throughput integration sites (IS) sequencing assay can simultaneously identify multiple VIS and can quantitate relative contribution of each VIS clone to detect probable vector insertion induced clonal expansion. Our quantitative high-throughput VIS tracking study revealed that of all HSPC transplanted in rhesus macaques, ~0.01% are long-term HSC and start contributing >1.5 years post-transplant [12]. Recently, our VIS assay found polyclonal repopulation of anti-HIV CAR modified HSPC in peripheral blood of pigtail macaques [17] indicating normal clonal expansion of stem cell derived CAR cells. Our long terminal repeat indexing–mediated integration site sequencing (LTRi-seq) method now 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 [18]. Clonal analysis of different tissues/organs 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 [18]. While these results show the potential of an anti-HIV gene-therapy, the underlying dynamics of clonal selection remain unknown owing to lack of a reliable quantitative clonal tracking technique for small animal models.

143 Although self-inactivating lentiviral vectors are low risk, a strong promoter within the vector can
144 upregulate the expression of endogenous genes where the vector integrated [19, 20]. In rhesus macaques
145 use of lentiviral vector containing a strong murine stem cell virus (MSCV) constitutive promoter-enhancer
146 in the LTR induced aberrant clonal expansion [21]. Thus, promotor and transgene in vector can adversely
147 impact the clonal behavior of transduced cells despite no obvious genotoxicity. Longitudinal clonal
148 tracking of gene-modified cells in an appropriate small animal model. An in vitro study found lentiviral
149 vector integration preference for active genes [22] and specifically associated with H3K36me3 [23] in
150 activated human CD34+ HSC. In absence of appropriate in vivo longitudinal clonal tracking and vector
151 integration data, the association between VIS preference for genes and histone modifications and its impact
152 on in vivo proliferation, differentiation, and repopulation of gene-modified human HSPC remains partially
153 explored.
154 Developing a reliable quantitative VIS assay for longitudinal clonal tracking in small animal models is
155 challenging due to low sample volume. Only 100µl blood/biweekly can be safely drawn from a typical
156 humanized mouse (~ 0.6µg of DNA assuming 1000 cell/µl), whereas VIS assays require more than 1µg
157 DNA. Multiple displacement amplification (MDA), a whole genome amplification technique, can increase
158 the DNA amount from few nano grams to micro grams with a very low error rate (1 in 10^6 to 10^7
159 nucleotides) [24] and high coverage [25]. MDA amplified DNA has been used for various sequence
160 sensitive applications such as single nucleotide polymorphism (SNP), next generation sequencing studies
161 [26], detection of retroviral IS, and sequencing full-length HIV-1 proviruses including the IS [27, 28]. We
162 hypothesize that MDA amplified DNA from low volume samples can be used with VIS assay for
163 quantitative longitudinal clonal tracking in hu-BLT mouse. Thereby, providing an essential tool to study
164 the clonal dynamics of HSPC repopulating in a mouse environment, to assess the safety of gene-therapy
165 vectors, and to explore the association between in vivo clonal behavior of gene-modified HSPC and
166 location of vector integration in host genome.
167 In this study, we developed a quantitative low volume vector integration site sequencing (LoVIS-Seq)
168 assay, a combined MDA and VIS assay for low-volume samples, for longitudinal clonal tracking in small
169 animal models. We longitudinally tracked hundreds of clones in two different gene-modified cell
170 populations simultaneously repopulating in hu-BLT mice. Human cell repopulation in a mouse
171 environment is polyclonal and resembles typical after-transplant HSPC expansion in other small and large
172 animal models. Analysis of multi-omics data of FL-HSPC and VIS in repopulating clones shows vector
173 integration bias for actively transcribed regions of the genome. Our assay provides an efficient tool to study
174 clonal repopulation in murine and humanized-mouse models used for stem cell and gene therapy research.
175 Results of clonal and genomic analysis revealed valuable insight into the clonal dynamics of human HSPC
176 repopulation in a mouse environment.

177

178 **Methods:**

179 *Human fetal thymus and isolation of FL-CD34+ cells from fetal tissue*

180 Human fetal thymus and livers were obtained from Advanced Bioscience Resources (ABR) and the UCLA
181 CFAR Gene and Cellular Therapy Core. Human fetal liver CD34+ HSPC and thymus pieces were
182 processed as previously described [29]. Briefly, a single cell suspension of fetal liver cells was strained
183 through 70µm mesh and layered onto density gradient separation media (Ficol Paque PLUS, GE
184 Healthcare). After 20 minutes of centrifugation, the mononuclear cells layer was collected. Anti-CD34+
185 microbeads (Miltenyi Biotech, San Diego, CA) were used for magnetic isolation of CD34+ cells from
186 mononuclear cells. Calvanese et al. [30] also obtained fetal liver from the UCLA CFAR Gene and Cellular
187 Therapy Core and followed identical CD34+ magnetic sorting to isolate uncultured FL-HSPC for RNA-
188 seq, ATAC-seq, and ChIP-seq assay.

189 *Vector transduction of FL-CD34+ cells*

190 The isolated FL-CD34+ cells were transduced overnight using either mCherry or EGFP vector at MOI
191 (multiplicity of infection) of 1 and 3, respectively. We used higher MOI for EGFP vector to achieve gene
192 marking level comparable to mCherry vector. Fraction of vector transduced cells were cultured for 4 days
193 and percentage of EGFP or mCherry expressing cells were measured by flow cytometry (Supplementary
194 figure 2A). Our data showed of the FL-CD34 cells transduced with mCherry vector, 68.4% cells were
195 mCherry+. For EGFP vector of the FL-CD34 cells transduced with EGFP vector 88.5% cells were EGFP+.

196 *Humanized BLT mouse and sample collection*

197 NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ (NSG) mice, 6-8-week-old, were myeloablated 1 day before transplant by
198 intraperitoneal (i.p.) injection with 10 mg/kg of 6-thioguanine (6TG) (Sigma-Aldrich, Saint Louis, MO) or
199 35mg/kg of Busulfan for mouse m860. Myelo-preconditioned mice were transplanted with human fetal
200 liver CD34+ HSPCs transduced with Anti-HIV vector (EGFP+) (0.5×10^6 cells/mouse) and mixed with
201 HSPCs transduced with the control (mCherry+) vector (0.5×10^6 cells/mouse). The mice were transplanted
202 with a two-step procedure: half the mixture of EGFP+ and mCherry+ transduced cells was solidified by
203 matrigel (BD Bioscience, San Jose, CA), mixed with CD34- cells as feeder cells (4.5×10^6 cells), and
204 implanted with a piece of human thymus under the mouse kidney capsule. Then, mice were injected with
205 the other half of the mixed EGFP+ and mCherry+ transduced cells via retro-orbital vein plexus using a 27-
206 gauge needle on the same day. Bone marrow cells for MDA were harvested from mouse m860 at week 25
207 post-transplant. Low human reconstitutions at early timepoints and graft versus host-disease related
208 illnesses at very late timepoints post-transplant have potential to impact interpretation of the clonal data. In
209 order to avoid these potential confounding effects, we longitudinally tracked clones between week 13 to 19
210 post-transplant. Every two weeks from weeks 13-19 post-transplant, 100µl of mouse blood was collected

211 from the retro-orbital vein for longitudinal clonal tracking and for monitoring reconstitution of human
212 leukocyte and level of the EGFP+ and mCherry+ marked cells. Plasma was removed and peripheral blood
213 cells were stained with monoclonal antibodies for 30 minutes (Supplementary figure 9). Red blood cells
214 were lysed with red blood cell lysis buffer (4.15 g of NH4Cl, 0.5 g of KHCO3, and 0.019 g of
215 ethylenediaminetetraacetic acid in 500 mL of H2O) for 10 minutes and washed with FACS buffer (2% fetal
216 calf serum in phosphate-buffered saline [PBS]). Stained cells were resuspended in 16µl PBS, of which 8µl
217 was split equally into two tubes for MDA replicates. The remaining 8µl was mixed with 300µl of 1%
218 formaldehyde in PBS and examined with Fortessa (BD Biosciences) flow cytometers. Flow cytometry data
219 was utilized to monitor human reconstitution (Supplementary table 2) and count human cells, mCherry+
220 cells, and EGFP+ cells as well as human T and B cells in blood (Supplementary figure 2B). The following
221 monoclonal antibodies with fluorochromes were used: human CD45-eFluor 450 (HI30, eBioscience), CD3-
222 APC-H7 (SK7: BD Pharmingen), and CD19-BV605 (HIB19: BioLegend). Data were analyzed on FlowJo
223 (TreeStar, Ashland, OR) software.
224 The UCLA Institutional Review Board has determined that fetal tissues from diseased fetuses obtained
225 without patient identification information are not human subjects. Written informed consent was obtained
226 from patients for use of these tissues for research purposes. All mice were maintained at the UCLA Center
227 for AIDS Research (CFAR) Humanized Mouse Core Laboratory in accordance with a protocol approved
228 by the UCLA Animal Research Committee.
229
230 *LoVIS-Seq workflow with whole genome amplification and quantitative VIS assay*
231 Multiple displacement amplification for whole genome amplification: To estimate the minimum number of
232 cells required for LoVIS-Seq, we collected 81,000, 27,000, 18,000, 9,000, 3,000, and 1,000 bone marrow
233 cells from mouse m860 by serial dilution and stored in 4µl of PBS at -20ºC. MDA was done directly on
234 cells (Supplementary figure 9) using the REPLI-g Single Cell Kit from Qiagen (Cat #150343) following
235 kit-specific protocol. For longitudinal clonal tracking in the blood compartment, 100 µl blood was drawn at
236 weeks 13, 15 & 17. At end point (week 19), max blood (≈ 1ml) was collected, out of which 100µl was used
237 for flow cytometric analysis along with MDA; the remainder was used to isolate unamplified whole blood
238 DNA. Cells for MDA were isolated as described above and stored in 4µl of PBS at -20ºC. MDA-amplified
239 DNA was then used for quantitative VIS assay. A Qiagen DNeasy Blood & Tissue Kit was used to extract
240 unamplified DNA from max blood cells, splenocytes, and bone marrow cells.
241
242 Quantitative VIS assay and data analysis workflow: For VIS sequencing, we followed the procedures
243 described in our previous publication [12, 18, 31, 32] and focused on analyzing only the right LTR
244 junctions using CviQI and RsaI restriction enzymes. For our VIS assay, we used one microgram MDA-

245 amplified or unamplified genomic DNA for animal m860 samples and two micrograms MDA-amplified or
246 unamplified genomic DNA for different time point samples, with a few exceptions (see Supplementary
247 Tables 1 and 2). DNA samples were subject to extension PCR using LTR specific biotinylated primers
248 /5BiotinTEG/CTGGCTAACTAGGGAACCCACT 3' and
249 /5BiotinTEG/CAGATCTGAGCCTGGGAGCTC 3'. The extension PCR product was then digested using
250 CviQI and RsaI restriction enzymes and biotin primer bound DNAs isolated using streptavidin-agarose
251 Dynabeads using magnetic separator as per manufactures instructions. The vector-host junctions capture on
252 streptavidin beads were processed for linker-mediated PCR (LM-PCR) methods as described
253 previously[31, 32]. The linker ligated vector-host junction DNA was subjected to two step PCR. First step
254 amplification was done using primer 5' CTGGCTAACTAGGGAACCCACT 3' and first linker primer
255 GTGTCACACCTGGAGATAT. We removed the internal vector sequence by restriction enzyme (SfoI)
256 digestion. The digested product of first PCR was then amplified using primer
257 5'ACTCTGGTAACTAGAGATCC 3' and second linker primer 5' GGAGATATGATGCGGGATC 3'.
258 Since the LTR index sequence is included in the vector-host junction, we obtain unbiased amplification all
259 the H1, H5 and/or WT VIS sequences. Lentiviral vectors used in this study as derived from FG12-mCherry
260 lentiviral vector[29] and all the primers are designed accordingly. A detailed protocol for VIS assay is
261 provided in supplementary text. The amplicon libraries prepared using custom made Illumina (Illumina,
262 San Diego, CA) sequencing primers for Illumina MiSeq (m860 samples) or iSeq100 (m599, m599, and
263 m591 samples) sequencer. Sequences with a virus-host junction with the 3' end LTR, including both the 3'-
264 end U5 LTR DNA and ≥ 25 base host DNA (with ≥ 95% homology to the human genome), were
265 considered true VIS read-outs. The sequence mapping and counting method was performed as described
266 previously[18]. In brief, sequences that matched the 3'end LTR sequence joined to genomic DNA as well
267 as LTR-indexes (H1-TGGAAAATCTCCAACA, H5-TGGAAAATATCCAACA or WT-
268 TGGAAAATCTCTAGCA) were identified using a modified version of SSW library in C++ [33]. Reads
269 were classified as H1, H5, or WT VIS based on the LTR barcodes used in the experiment. VIS sequences
270 were mapped onto the human genome (Version hg38 downloaded from https://genome.ucsc.edu/) using
271 Burrows-Wheeler Aligner (BWA) software. Mapped genomic regions were then used as reference and VIS
272 reads were remapped using BLAST to further remove poorly mapped reads to get an accurate estimate of
273 sequence count. Final VIS counting was done after correcting for VIS collision events and signal crossover
274 as described previously. VIS with a final sequence count less than the total number of samples analyzed per
275 animal were removed. VIS clones with maximum frequency values below 1st quartile were classified as
276 "low frequency", clones with maximum frequency value above 3rd quartile were classified as "high
277 frequency", and clones with maximum frequency between the 1st and 3rd quartiles were designated
278 "medium frequency". VIS clones that were detected with frequency >0 at every week from 13-19 are

279 termed "persistent clones". The 10 high frequency VIS clones at each timepoint were selected as top 10
280 VIS. All the VIS data and list of VIS-proximal genes is provided in supplementary file.
281 *Random integration sites*
282 Random integration sites were generated in silico using a custom python script. To mimic our VIS assay,
283 we randomly selected 1000 integration sites that were within ±1500bp of the nearest CviQI/RsaI (GTAC)
284 site in the human genome (hg38).
285 *Clonal diversity analysis*
286 For diversity analysis, we used Rényi's diversity/entropy [34] of order α defined as follows

287 $$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum p_i^\alpha\right),$$

288 where p_i is the proportional abundance of the ith VIS clone for $i = 1, \ldots, n$. At each timepoint, an average
289 Rényi's diversity profile was obtained by calculating average values of $\boldsymbol{H_\alpha}$ for $\alpha \geq 0$. The α is considered
290 as a weighting parameter such that increasing α leads to increased influence of high frequency VIS clones.
291 The proportional abundance is calculated as $p_i = s_i/S$, where s_i is the sequence count of the ith VIS clone
292 and S is the sum of sequence counts from all VIS clones. The Rényi's diversity $\boldsymbol{H_\alpha}$ values are averaged
293 over two replicates and plotted as a function of α. If all VIS clones contributed equally, i.e. $p_i = \frac{1}{n}$ for all
294 $i = 1, \ldots, n$, then $\boldsymbol{H_\alpha}$ for all values of α would be equal and the profile (line) would be horizontal. VIS
295 clones expanding at different rates would show decreasing $\boldsymbol{H_\alpha}$ values as α increases, generating a
296 downward-sloped diversity profile that is steeper with more non-uniform clonal expansion. $\boldsymbol{H_\alpha}$ indicates
297 clonal diversity of the repopulating cells, such that consistently higher values of $\boldsymbol{H_\alpha}$ indicate a more diverse
298 clonal population. If the profiles for two populations/samples cross, then their relative diversities are
299 similar. For $\alpha = 0$, $\boldsymbol{H_0} = \log(n)$ and the antilogarithm of this value equates to the richness or number of
300 unique IS. $\boldsymbol{H_\alpha}$ at $\alpha = 1$ and $\alpha = 2$ are the Shannon and 1/Simpson indexes, respectively. We calculated
301 Renyi's diversity using the R package BiodiversityR (*https://cran.r-*
302 *project.org/web/packages/BiodiversityR/index.html*). For the above analysis, we used raw sequence counts
303 from two replicates without distinguishing between mCherry-H5 VIS and EGFP-WT VIS.
304 *RNA-seq data analysis*
305 Raw sequence data of uncultured FL-HSPC (in triplicate) was pre-processed for quality using Fastqc.
306 Trimmomatic was used to remove adaptors and for quality trimming. After this, reads were aligned onto
307 human genome hg38 using RNA STAR aligner [35]. SAMtools was used to remove reads with low
308 mapping scores (< 20) and to generate BAM files. Cufflinks [36] was used to calculate FPKM values for all
309 genes. The human cancer consensus gene list is from Catalogue of Somatic Mutations In Cancer
310 (https://cancer.sanger.ac.uk/census).
311 *ATAC-seq analysis*

312 Raw sequence data of uncultured FL-HSPC (in triplicate) was pre-processed for quality using Fastqc.
313 Adaptor removal and quality trimming was done using Trimmomatic. After this, reads were mapped onto
314 human genome hg38 using bowtie2 with parameter --very-sensitive -X 2000 -k 1. SAMtools was used to
315 remove reads with low mapping (<20) scores, blacklisted regions, and to generate BAM files. Picard tool
316 kit was used to remove duplicate reads. We used Genrich, a paired end peak caller, to identify ATAC
317 peaks. Software deepTools[37] was used to generate coverage (.bw) files and for visualization of open
318 DNA in genes and VIS-proximal regions.
319 *ChIP-seq data analysis*
320 Raw sequence data of uncultured FL-HSPC for histones, RNAploII, and input were pre-processed for
321 quality using Fastqc. Trimmomatic was used to remove adaptors and for quality trimming. After this, reads
322 were mapped onto human genome hg38 using bowtie2 with parameter --local. SAMtools was used to
323 remove reads with low mapping (<20) scores, blacklisted regions, and to convert SAM to BAM format.
324 Picard tool kit was used to remove duplicates. MACS2 tool was used to call peaks for all histone marks and
325 RNApolII using input sample as control. Software deepTools [37] was used to generate coverage .bw files
326 and for visualization of histone/RNApolII in genes and VIS proximal regions. Proximity of VIS to a peak
327 was calculated by measuring the distance between VIS and the coordinates of the nearest peak; for VIS
328 within the peak distance is 0bp.
329 *Statistical analysis*
330 Relative clonal frequencies are summarized as means ± standard deviations (SDs). Pearson correlation (r) is
331 used to assess similarity in terms of a linear relationship among pairs of samples. The interclass correlation
332 coefficient, ICC(3,1) two-way mixed effects, absolute agreement is used to compare the reproducibility
333 (ICC) of clonal profiles among replicates [38]. ICC, Pearson's r and p values are calculated using statistical
334 software R (version 3.6, https://www.r-project.org/). To determine if VIS preference for genomic and
335 epigenetic features differs significantly from random IS, we used Pearson's chi-squared test with Yate's
336 continuity correction (function chisq.test() in software R). We used Principal Component Analysis (PCA)
337 to reduce the complexity of read coverage data of multiple chromatin feature in proximity to VIS. The
338 dimensionality reduction by PCA method is similar to clustering and allows detection of patterns in the
339 data. In this study, PCA was done using software deepTools [37].
340
341 **Results:**
342 *Minimum 10,000 bone marrow cells or 25µl blood is sufficient for LoVIS-Seq:*
343 To estimate the minimum number of cells sufficient for LoVIS-seq assay to provide a quantitative clonal
344 analysis, we collected bone marrow (BM) cells from hu-BLT (bone marrow-liver-thymus) mouse (m860).
345 Fetal liver CD34+ cells transduced with sh1005(anti-CCR5 shRNA)-EGFP vector or control mCherry

346 vector were mixed in equal ratio and transplanted in the mouse (Figure 1a-b, details in Methods). We first
347 estimated clonal composition of the EGFP-WT (WT LTR-index) and mCherry-H1 (H1 LTR-index)
348 populations using unamplified bulk DNA, in triplicate, as described previously [18]. A total of 300 ±42 SD
349 (216 ±22 SD mCherry-H1 VIS and 84 ±20 SD EGFP-WT) VIS were detected indicating polyclonal
350 expansion. The clonal size distribution in the mouse bone marrow (Figure 1c) resembled that of recently
351 observed hu-BLT mice [18] and pigtail macaques [17] by us as well as that of observed by others in mice
352 [13], autologously transplanted nonhuman primates [12, 39], and humans [11, 40] overall, indicating
353 normal clonal repopulations.
354 For low volume sample with DNA yield of <1µg, we propose MDA to increasing the quantity of and use
355 the amplified DNA for VIS assays. For our LoVIS-seq assay, we first perform MDA directly on the cells
356 and use the MDA amplified DNA for VIS assays (Supplementary figure 9). We hypothesized that LoVIS-
357 seq can be used for clonal analysis of low volume samples without loss of quantitative accuracy. To test our
358 hypothesis, we first performed MDA directly on 81,000, 27,000, 9,000, 3,000, and 1,000 bone marrow
359 cells, each in duplicate (Supplementary figure 1A, details in Methods section). We used equal amounts of
360 MDA-amplified DNA and unamplified bulk DNA for the VIS assay (Supplementary table 1). We assessed
361 reproducibility of clonal profiles in MDA amplified DNA of 81,000, 27,000, 9,000, 3,000, and 1,000 bone
362 marrow cells against that of unamplified DNA sample from bulk BM cells. We performed routinely used
363 Pearson correlation analysis and also did Intraclass Correlation Coefficient (ICC) analysis to test the
364 reproducibility of clonal abundances (details in Method section). We found high reproducibility of clonal
365 profiles in different MDA samples and within-MDA replicates of 81,000 to 9,000 cells (avg. Pearson's r
366 value >0.91) (Figure 1d and Supplementary figure 1B-F). To determine the minimum number of BM cells
367 that are required for obtaining high quantitative accuracy, we compared the reproducibility of MDA
368 amplified 81,000, 27,000, 9,000, 3,000, and 1,000 bone marrow cells and found that for less than 9,000
369 cells, the reproducibility dropped (avg. Pearson's r=0.87 for 3,000 and 0.73 for 1,000 cells). Importantly,
370 reduced cell numbers caused a modest reduction in VIS detection (Supplementary figure 1G) showing
371 minimal impact on the sensitivity of our assays. In order to further validate the reproducibility of our assay,
372 we used ICC analysis. ICC can be used to simultaneously compare replicates of unamplified DNA against
373 replicates of MDA amplified samples. The ICC values (Supplementary figure 1H) were similar to
374 Pearson's r, further indicating loss of reproducibility when fewer cells were used as source material for
375 MDA. Overall, both the Pearson and ICC analysis, validate that MDA-amplified DNA from >10,000 bone
376 marrow cells is sufficient for LoVIS-Seq.
377 After establishing the accuracy and reproducibility using BM cells, we tested LoVIS-seq assays with hu-
378 BLT mouse blood. We collected 100µl blood at week 13, 15, 17 and ~1ml of whole blood at week 19 post-
379 transplant from hu-BLT mice. These mice were transplanted with an equal mix of human CD34+ cells

380 transduced with anti-HIV EGFP-WT vector and control mCherry-H5 vector (Figure 1e). Cells from 50µl
381 blood were used for flow cytometry and the remaining cells were used for MDA duplicates; each 25µl of
382 blood (~400 and ~600 human cells/µl blood for mouse m599 and m598, respectively). High correlation
383 (median Pearson's r =0.93) of mCherry-H5 and EGFP-WT VIS clonal frequencies between unamplified
384 and MDA-amplified DNA from blood cells (Figure 1f) indicates clonality of entire mouse blood can be
385 captured with 25µl of blood. Importantly, the MDA replicates also showed high reproducibility (median
386 Pearson's r>0.95, Supplementary figure 3A). To further validate the reproducibility, we simultaneously
387 compared two MDA replicates and unamplified DNA from whole blood. The ICC values of >0.95 further
388 confirmed reproducibility of our assay with mere 25µl blood. In conclusion, our LoVIS-Seq assay
389 accurately captured the clonality of two vector-modified cell populations in hu-BLT mouse blood using
390 mere 25µl of blood or as few as 10,000 cells. These results indicate that LoVIS-Seq assay can be used to
391 study clonal repopulation of gene-modified cells in mouse blood using only a fraction of (approximately
392 2.5%) of whole blood.
393 *Simultaneous clonal tracking of therapeutic vector-modified and control vector-modified populations:*
394 After demonstrating the high accuracy and reproducibility of LoVIS-Seq assay, we aim to examine clonal
395 dynamics of anti-HIV EGFP-WT vector and control mCherry-H5 vector modified two human cell
396 populations simultaneously repopulating the mouse blood. We collected 25µl of hu-BLT mouse blood
397 samples for each MDA replicate on week 13, 15, 17 and 19 post-transplant. Despite the variation in cell
398 count over time each replicate/sample had >10,000 (avg. 21,189; ±7,364) cells (Supplementary table 4 and
399 Supplementary figure 2B). After performing LoVIS-Seq assay on these cell samples, we identified 329
400 VIS in m599, 222 VIS in m598, and 346 VIS in m591 (total 897 VIS clones) and longitudinally tracked
401 change in the relative contribution of these clones to the repopulation from week 13 to 19 (Figure 2a).
402 From total 897 VIS clones, we classified mCherry-H5 (792) and EGFP-WT (105) VIS clones using LTR-
403 indexes in the two vectors. Consistent with our previous study [18], we observed high correlations between
404 the total mCherry-H5 VIS clonal frequency and mCherry^{+} cell percentage by flow cytometry (Pearson's r =
405 0.8) as well as between total EGFP-WT VIS clonal frequency and EGFP^{+} cells percentage (Pearson's r =
406 0.9, Supplementary figure 3B). The high correlation between flow cytometry and clonal frequency data
407 indicates quantitative accuracy of LoVIS-Seq assay and ability to simultaneously track clonal expansion of
408 two human cell populations repopulating in the same animal using a mere 25µl of blood. In addition,
409 longitudinal clonal tracking further shows expansion in EGFP-WT clones coincided with reduction of
410 mCherry-H5 contribution and vice-versa (Figure 2b; solid lines); these changes closely match the change in
411 EGFP^{+} and mCherry^{+} cell percentages measured by flow cytometry (Figure 2b; dashed lines). Furthermore,
412 repopulation in both EGFP-WT and mCherry-H5 populations is largely driven by expansion of a few
413 HSPC clones—a characteristic feature of after-transplant repopulation. These results present a first-ever

414 demonstration of clonal expansion of two populations, therapeutic vector-modified and control vector-
415 modified, competing to repopulate in hu-BLT mice.
416 *Sustained polyclonal repopulation with rapid clonal expansion and stabilization:*
417 The clonal analysis revealed that at each timepoint hundreds of VIS clones contributed to the repopulation
418 and contributions of EGFP-WT and mCherry-H5 clones changed over time. To investigate the properties of
419 the clonal dynamics of human cell repopulation in mouse environment, we analyzed the quantitative clonal
420 tracking data from different timepoints. We found that the maximum number of VIS were detected at week
421 13 and on average, only 13% (±5%) fewer VIS were detected at week 19 compared to week 13. Also, the
422 number of total VIS contributing to repopulation at each timepoint decreased with time (Figure 2c). On
423 average, 61% (±12.8%) of persistent VIS clones (m599: 235 clones, m598: 150 clones, and m591: 165
424 clones; total 550 clones) consistently contributed for 6 weeks and provided stable polyclonal repopulation
425 (Figure 2a area plots). While the number of VIS clones steadily dropped over time in both the mCherry-H5
426 and EGFP-WT populations, their clonal profiles became increasingly similar (Supplementary figure 4A-B),
427 comparable to polyclonal repopulation patterns that have been reported in nonhuman primates [12].
428 Moreover, correlation between time points indicates clonal distribution at week 13 differs from week 19,
429 with clonal expansion stabilizing around week 17 (Supplementary figure 4B). We also examined Rényi's
430 diversity profiles for each animal at each time point (details in Methods section). The clonal diversity at
431 week 13 was highest and subsequently decreased with time (Figure 2d). Diversities were similar between
432 weeks 17 and 19 as indicated by their overlapping diversity profiles. The two commonly used diversity
433 indexes Shannon Index [41] ($\alpha = 1$) and ln(1/Simpson index) ($\alpha = 2$) referred here on as Simpson index
434 [42] to capture the overall change in the diversity using single numerical values. Shannon index is more
435 sensitive to number VIS clones and Simpson index to the relative frequency of each VIS clones. The
436 Shannon[41] and Simpson[42] indices dropped between weeks 13 and 17, indicating expansion of fewer
437 clones (Supplementary figure 4C). The similarity of the indices between weeks 17 and 19 suggests
438 stabilization of clones. Contribution by the most frequent clone rose ~2.2 times, from 0.078 (±0.005, n=3)
439 at week 13 to 0.174 (±0.030, n=3) at week 17 (Figure 2e), signaling rapid expansion of a few clones.
440 Compared to humans, wherein stable clones appear >1 year post-transplant [11], in hu-BLT mice the clonal
441 repopulation remained normal despite faster expansion and earlier stabilization of human HSPC clones.
442 Overall, these results indicate notable difference in the time scale of clonal repopulation of human cells in
443 mouse from that in human environment and these finding have larger implication while interpreting and
444 translating results from hu-BLT mouse models to human settings.
445 *Clonal sharing between organs reveals normal repopulation and unique tissue distribution pattern:*
446 We showed first-ever demonstration of clonal repopulation of two human cell population in mouse
447 environment with sustained polyclonal repopulation and competitive expansion of two human cell

448 population in mouse blood. However, clonal studies in nonhuman primates revealed that clonal expansion
449 patterns in blood during early phase of repopulation differ from other organs [43]. In hu-BLT mice whether
450 clonality of blood differs from the tissue/organ is unknown. We determined the clone profiles is bone
451 marrow (BM) and spleens compartments of these mice using bulk cells for our VIS assay. Similar to blood,
452 BM and spleen compartments also showed normal polyclonal repopulation. Between organ/tissue
453 comparison of clonal size distributions showed a highly correlated clonal expansion pattern (avg. Pearson's
454 r =0.96) between blood and spleen (Figure 3); however, clonal expansion in bone marrow was less
455 correlated to blood or spleen. Importantly, we observed that in all three tissue compartments, persistent
456 clones contributed the most to repopulation. These results show substantial clonal sharing among yet,
457 differential clonal contributions. Our data also indicate that clonal tracking using BM samples would reflect
458 the clonality only of the BM compartment. Compared to serial bone marrow extractions, clonal tracking
459 using blood is more practical and closely reflect clonal composition of spleen.
460 *Influence of genomic location and proximal genes on clonal growth:*
461 We found normal and polyclonal expansion in human cells in mouse blood and tissue compartment yet
462 mutagenic insertions causing abnormal growth remains a major concern for gene-therapy vectors. For each
463 VIS, our assay provided both relative frequency and genomic location of integration allowing us to identify
464 possible of insertional mutagenesis induced clonal growth. Our VIS data from in vivo repopulating clones
465 showed preference for high gene density chromosomes (Supplementary figure 5A) also observed with the
466 HIV-1 integration pattern [44]. Additionally, similar genomic distribution of low, medium, and high
467 frequency in vivo repopulating VIS clones (Figure 4a) suggests clonal expansion is unrelated to genomic
468 location of integration. We found that in ~80% of the total detected clones, VIS occurred within ±1Kb of
469 protein-coding genes, significantly higher than 54% of 1000 random IS ($p < 0.001$, Supplementary figure
470 6A). About 8% of VIS were within ±1Kb of long non-coding RNA (lncRNA) and 10% were outside ±1Kb
471 of any genes (distal VIS, Figure 4b inside pie chart). Persistent and top 10 VIS clones (top 10 high
472 frequency VIS clones collected from each timepoint) showed similar preference for gene biotypes (Figure
473 4B inside pie chart). Out of all VIS, only 66 (including 1 out of 42 top 10 VIS) were proximal to known
474 cancer consensus genes (Figure 4a Circos plot). Gene ontology analysis of proximal genes and their mouse
475 orthologs showed significant enrichment ($P < 0.01$) in various biological pathways such as cell-cell
476 interaction, viral process and transcription regulation (Supplementary figure 5B). In vitro data of VIS-
477 proximal gene in vector transduced human CD34+ HSC showed enrichment in similar biological
478 processes.[22] A recent study claims that biological function of genes near HIV integrations sites influence
479 in vivo clonal expansion of HIV infected cells.[45] However, we found no strong evidence of link between
480 in vivo differentiation potential of VIS clones and genomic location of its VIS or biological function of the
481 VIS-proximal genes.

482 *Vector integration bias for transcriptionally active genes in human fetal liver HSPC*:
483 For in vivo repopulating clones we found VIS preference for genic regions in FL-HSPC however, the
484 transcriptional state and expression level of the VIS-proximal genes prior to vector integration remains
485 unknown. To address this, we analyzed transcriptomic (RNA-seq) and functional genomic (ATAC-seq and
486 ChIP-seq) data from uncultured human FL-HSPC[30] isolated and processed using protocol identical to
487 one used in our study (see Methods). Owing to the direct biological relevance of human FL-HSPC to our
488 humanized BLT mice models, the multi-omics data is well suited to investigate the impact of vector
489 integration on stemness and in vivo repopulation of vector-modified HSPC.
490 The gene expression (RNA-seq) data shows that of all in vivo detected clones, including the persistent
491 clones and top 10 VIS clones, 694 out of 897 VIS (>77%) are within ±1Kb of transcriptionally active genes
492 (FPKM>1) (Figure 4b outer donut chart). These VIS-proximal active genes include 668 protein coding
493 genes, 17 lncRNA (non-coding genes) and 9 other genes (other non-coding genes and pseudogenes). This is
494 significantly higher than the ~27% of random IS proximal to active genes ($p < 0.001$, Supplementary figure
495 6A). The level of transcriptional activity of VIS-proximal genes (based on the FPKM value) was slightly
496 higher than the median expression level of all active genes (Supplementary figure 5C). Moreover, the
497 median gene activity level (FPKM) varied based on gene biotype, with protein coding proximal genes of
498 all, persistent, and top 10 VIS clones having higher activity than proximal lncRNA or pseudogenes. Similar
499 to in vitro observations using cell line and primary cells have reported vector integration bias for active
500 genes with low to moderate expression [22, 46]. Our in vivo clonal tracking data show that for repopulating
501 clones VIS bias for moderately active genes and against highly active genes. These findings suggest that
502 similar to in vitro, in vivo viability and repopulation of gene-modified HSPC is likely linked to expression
503 level of VIS proximal gene.
504 *Epigenetics of FL-HSPC reveals VIS bias for actively transcribed genes:*
505 We speculate that similar to expression activity of genes the chromatin structure of active genes strongly
506 influences vector integration. Previous in vitro studies in cell lines suggested vector integration preference
507 for genomic features such as select histone modification and DNase I hypersensitivity sites. However,
508 neither such analysis is available for human FL-HSPC nor the influence of genomic location of VIS on in
509 vivo engraftment of these HSPC has been explored. To investigate this, we analyzed functional genomic
510 data of uncultured human FL-HSPC [30] with direct biological relevance to our humanized BLT mice
511 models which are constructed by transplanting FL-HSPC in to NSG mice. To investigate whether vector
512 preferred open chromatin or chromatin associated proteins, we analyzed chromatin accessibility (ATAC-
513 seq) ChIP-seq data of 8 histone modifications H3K4me3, H3K9ac, H3K9cr, H3K27ac, H3K4me1,
514 H3K79me2, H3K36me3, and H3K27me3 and RNApolII (chromatin associated RNA polymerase II), all
515 from uncultured human FL-HSPC [30].

516 We explored the epigenetic landscape of all the VIS-proximal genes in FL-HSPC using the multi-omics
517 data (Figure 5a profile plots and heatmaps). For active VIS-proximal genes (FPKM>1), our analysis
518 showed active transcription markers such as open chromatin region (ATAC-seq peaks) and H3K4me3,
519 H3K9ac, H3K9cr, and H3K27ac enriched regions near TSS as well as active enhancer markers (H3K27ac
520 and H3K4me1 enriched regions). H3K36me3 and H3K79me2-enriched regions within the gene body of
521 VIS-proximal active genes indicated genes being actively transcribed. We also detected the enrichment of
522 RNApolII-bound chromatin (RNApolII) in active gene that coincided with TSS regions. For inactive
523 (FPKM<1) VIS-proximal genes, these active gene marks were less prominent. Random IS-proximal genes
524 showed similar enrichment profiles in active genes and their lack in inactive genes (Supplementary figure
525 6B). Overall, the epigenetic markers of the VIS-proximal genes show that the vector integration is biased
526 for actively transcribed genes. However, further analysis was needed to determine presence of VIS bias for
527 particular epigenetic features. We investigated possible enrichment of epigenetic features within ±1Kb
528 region of VIS (Figure 5b). We found that vector integration did not occur near open chromatin (ATAC-seq)
529 regions. Similarly, for histones H3K4me3, H3K9ac, H3K9cr, and H3K27ac high enrichment was not
530 detected near VIS (Figure 5b lineplots), indicating that vector integration likely avoids transcription
531 regulatory regions such as TSS and enhancer. The enrichment of RNApolII-bound chromatin (RNApolII)
532 was also not visible near VIS further confirming that transcription regulatory regions are avoided. In
533 contrast, enrichment of H3K36me3 and H3K79me2, near VIS, showing clear preference for actively
534 transcribed regions within gene body. A Principal component analysis (PCA) on normalized enrichment
535 levels (RPKM values) over a ±1Kb region of VIS (Figure 5b data) clearly separated H3K36me3 and
536 H3K79me2 from all other features and showed no bias for random IS (Supplementary figure 7). These
537 results further confirmed the VIS bias for actively transcribed regions.
538 *Vector integrations in repopulating clones were outside gene regulatory regions:*
539 In order to further investigate the proximity of VIS to epigenetic features at finer resolution, we called
540 peaks for all the epigenetic markers (described in Methods) and measured the distance between the VIS and
541 the center of the peak. Our analysis showed that 659 (73.5% of total) VIS were near (±1Kb) one of the 10
542 epigenetic features (Supplementary figure 8). Previous study using epigenetic data from different cell lines
543 found a strong positive association of HIV integration to histones H3K4me1, H3K4me2, H3K4me3, and
544 H3K9ac [47]. However, our analysis showed that only 6 out of 897 VIS were near H3K4me3, a marker
545 for gene TSS, indicating disfavoring of transcription regulatory regions. Occurrence of only 63 VIS in
546 H3K4me1 enriched regions indicating disfavoring of both H3K4 mono- and tri-methylation. Interestingly,
547 compared to H3K4m1 and H4K4me3 more VIS associated with H3K9ac, H3K9cr, and H3K27ac
548 (Supplementary figure 8A). Consistent with previous studies, we also observed that very few integration
549 events (3 VIS) were near H3K27me3 enriched regions indicating that in repopulating clones vector

550 integration happened outside of transcriptionally repressed regions (Figure 5b and 5c top panel,
551 Supplementary figure 8). Analysis of open chromatin regions revealed that only 42 out of 897 VIS were in
552 proximity of ATAC-seq peaks. Overall, fewer vector integration events happened in proximity of open
553 chromatin (ATAC peak) or histones modification H3K4me3 and H3K4me1. These finding show that
554 despite preference for active genes, in majority of repopulating clones, vector integration occurred outside
555 the transcription regulatory regions (TSS and enhancers) of the genes. These finding also suggest
556 integration in regulatory regions may adversely impacts survival of vector transduced FL-HSPC.
557 *Vector integrations predominantly occurred in H3K36me3 enriched regions:*
558 In repopulating clones, we observed bias against transcription regulatory regions of VIS-proximal genes
559 despite significant VIS bias for active genes. While random IS were evenly distributed across all 10 histone
560 marks (Figure 5c bottom panel, Supplementary figure 8C), in repopulating clones VIS distribution skewed
561 towards H3K36me3 and/or H3K79me2 (Figure 5c top panel; Supplementary figure 8A-B). Out of 897
562 repopulating clones, 486 VIS (54% of total VIS) were within ±1Kb of histone marks H3K36me3,
563 significantly higher compared to only 11.6% random IS being associated with the histone ($p<0.001$,
564 Supplementary figure 6C). For the top 10 most frequent VIS clones, preference for all chromatin features
565 (Figure 5c middle panel) was also similar to that for all VIS (Figure 5c top panel). A previous in vitro study
566 using HIV based vector found that in human CD34+ cells, only 14.5% of vector integrations occurred
567 within ±1Kb of H3K36me3 [23]. Compared to in vitro, we found 3.7 times more VIS to be within ±1Kb of
568 H3K36me3 [23] for in vivo repopulating clones. Despite significant integration bias for active genes and in
569 particular for histone H3K36me3, the clonal size distribution and repopulation of human cell in hu-BLT
570 mice remained normal. These in vivo finding signify that both the vectors are safe and that vector
571 integration near H3K36me3 appears less detrimental to engraftment and in vivo differentiation of vector
572 transduced HSPC.
573
574 **Discussion:**
575 In this study we presented LoVIS-Seq, a new longitudinal clonal tracking method requiring a mere 25µl
576 blood or less to monitor clonal behavior of gene-modified cells in small-animal models. LoVIS-Seq
577 quantitatively captured the clonality of both control (mCherry-H1/H5) and anti-HIV (EGFP-WT)
578 populations in whole blood. We provide the first-ever demonstration of simultaneous polyclonal
579 repopulation of therapeutic vector-modified and control vector-modified cell populations in hu-BLT mouse
580 blood. The polyclonal expansion resembles normal post-transplant HSPC clonal repopulation in mice [13],
581 nonhuman primates [12, 17, 39] and humans [11, 40]. Notably, the clonal frequency data recapitulated the
582 flow cytometric measurements. Persistent clones are major contributors in blood, BM, and spleen. The
583 multi-omics data from uncultured FL-HSPC revealed that vector integration in VIS clones that repopulated

584 in mouse environment is significantly biased toward H3K36me3 and/or H3K79me2 enriched regions.
585 Remarkably, this vector integration bias appears inconsequential with respect to clonal repopulation, as
586 gene-modified HSPC differentiated normally in vivo; this confirms the safety of our therapeutic and control
587 lentiviral vectors.
588 We recently showed that in hu-BLT mice, monitoring clonal expansion of gene-modified and control
589 populations within the same animal gives an unbiased analysis and allows a more direct assessment of
590 therapeutic vectors [18]. In the current study, we longitudinally tracked both anti-HIV gene-modified and
591 control vector-modified populations in the same hu-BLT mouse. Interestingly, we found a competitive
592 growth pattern between two populations, with few clones from each population leading the expansion. The
593 clonal profiles in both populations resemble typical after transplant clonal repopulation confirming safety
594 of both vectors. Since safety and efficacy of multiple vectors can be tested in the same humanized mouse,
595 LoVIS-Seq can reduce both cost and time of vector development.
596 Previous studies propose that in myeloablated mice, hematopoiesis tends to stabilize around 22 weeks post-
597 transplant [15, 48] while a recent study suggested 16 to 24 weeks [49]; our data indicates clonal
598 stabilization between 17 to 19 weeks post-transplant. Overall, the time scale of clonal repopulation of
599 human HSPC in hu-BLT mice resembles that of mouse HSPC after transplant in mice. Cord blood HSPC
600 transplanted in NGS mice also showed similar clonal behavior, with clonal stabilization starting near week
601 18 to 20 post-transplant [50, 51]. Although the timescales compare well with other studies, caution is due
602 considering high incidence of graft versus host-disease related illnesses in xenograft mouse models.
603
604 HIV-1 and HIV-1 based vectors are known to favor transcriptionally active gene dense regions [44, 46],
605 with preference for histone modification H3 acetylation, H4 acetylation and H3K4 methylation [47]. Some
606 studies found no preference for DNase I hypersensitive sites [52] and H3K4 methylation being disfavored
607 [53] and preference for H3K36me3 in Jurkat cells [53]. However, the implications of HIV-1 integration on
608 cell fate are compounded by infection induced cytotoxicity. Another important study provides a link
609 between HIV-1 integrations in genes related to viral replication and in vivo expansion and persistence of
610 HIV-1 infected cells [45]. However, influence of epigenetic features proximal to HIV-1 IS on clonal
611 expansion and persistence remained unexplored. The HIV-1 integration occurs with assistance from nuclear
612 pore complex and targets the active genes closer to the nuclear pore and disfavor heterochromatin regions
613 and active regions located centrally in the nuclease [54]. In partial agreement with previous studies, our
614 analysis of repopulating VIS clones shows that clones with vector integration in transcriptionally active
615 genes in FL-HSPC have higher probability of in vivo expansion however, RNA-Seq data indicates bias
616 against highly expressed genes. For in vitro activated cord blood and BM derived CD34+ HSC, the
617 lentiviral vector showed no preference for highly active genes [22]. This is likely due to either obstruction

618 by transcriptional machinery or detrimental effect of vector integration on survival of the cell. Further
619 investigation of 10 chromatin features showed that vector integration had strong bias toward actively
620 transcribed regions marked by histone modifications H3K36me3 and H3K79me2. This bias could be
621 attributed to LEDGF/p75, a chromatin binding protein essential for efficient HIV-1 integration [55, 56],
622 that binds to integrases of HIV-1 [57, 58] and protects the pre-integration complex from degradation [59],
623 whereas the N-terminal PWWP domain of LEDGF is known to preferentially interact with H3K36me3 [60,
624 61]. A recent study showed that inhibiting binding of LEDGF to integrases leads to increased distance
625 between IS and H3K36me3 regions. H3K79me2 and H3K36me3 mark the gene body and H3K36me3
626 marks exons and is positioned near the 5` end of the exon and is correlated with alternative splicing [62,
627 63]. Thus, VIS in proximity of H3K36me3 are likely to influence co-transcriptional splicing of the
628 proximal gene as well as expression of the vector itself. Human HSPC clones reconstituting the mice blood
629 have significant vector integration bias that suggest a link between in vivo differentiation of VIS clones and
630 proximity of vector integration to H3K36me3. However, tracking of hundreds of single HSPC clones
631 indicated that such transcription events have no adverse impact on the stemness of repopulating vector-
632 modified HSPC.
633 Another recent study demonstrated use of CRISPR/Cas to introduce barcodes in the long-term HSPC and
634 longitudinally tracked a very limited number of HSPC clones [51]. Comparatively, using LoVIS-Seq we
635 have tracked ~10 times more HSPC clones per animal with high accuracy and reproducibility. It is
636 pertinent to note that to enable insertion of barcoded donor DNA into the host genome, HSPC need to
637 undergo in vitro preconditioning and incubation before transplantation. The double stranded breaks
638 introduced by CRISPR/Cas activate DNA damage responses causing significant delays in HSPC
639 proliferation and affects their in vivo repopulation [64]. Additionally, off-target gene-editing by
640 CRISPR/Cas remains a concern. In contrast, LoVIS-Seq does not require preconditioning of HSPC and
641 provides a ready to use high-throughput clonal tracking assay for small-animal models. Furthermore,
642 LoVIS-Seq has wider applicability owing to its adaptability to many lentiviral vectors commonly used to
643 insert transgenes or reporter gene such as GFP.
644 LoVIS-Seq with whole genome amplification allows for quantitative assessment of clonal behavior in
645 small-animal models. However, the accuracy and reproducibility of our assay depends on the initial number
646 of cells used for MDA (Figure 1d). To minimize sampling errors, it is important to have a sufficient
647 number of gene-marked human cells in each 25µl of blood or 10,000 cells to represent each clone in similar
648 proportions as in the bulk population. Higher human reconstitution and gene marking are often desirable
649 and necessary conditions wherein our assay provides optimal results.
650 **Conclusions:**

651 Our study revealed the in vivo dynamics of clonal expansion, emergence of stable stem cell clones, and
652 consequences of vector integration bias on repopulation of human HSPC in mouse environment. Using a
653 mere 25µl blood and LTR-indexed vectors, we provide first ever demonstration of dynamic polyclonal
654 expansion in both control and therapeutic vector-modified populations in the same hu-BLT mice. Our study
655 provides insights into clonal dynamics of human HSPC revealing a faster yet normal repopulation of
656 human HSPC in mouse environment. In repopulating clones, we found significant VIS bias for actively
657 transcribed regions specifically for H3K36me3-enriched regions indicative of vector integration in these
658 regions being less detrimental to HSPC. These results provide a framework for understanding the clonal
659 behavior of human HPSC repopulating in a mouse environment, critical for translating results from
660 humanized mice models to the human settings. Moreover, our new assay provides an efficient tool for
661 multifaceted analysis of clonal dynamics in murine and humanized-mouse models used extensively in
662 infectious disease, cancer, gene-therapy, and stem cell research.
663
664 **Data availability.**
665 Raw RNA-seq, ATAC-seq, ChIP-seq data of uncultured FL-HSPC from published reference is available in
666 Gene Expression Omnibus (GEO) with the accession code GSE111484.[30]
667
668 **Acknowledgements:** We thank the technical support from the UCLA Center for AIDS research (CFAR)
669 research cores including the CFAR Gene and Cellular Therapy Core and the CFAR Humanized Mouse
670 Core. We would like also to thank Anna Sahakyan and Ruth Cortado for their help with making lentiviral
671 vectors, transducing human CD34+ HSPC, and performing humanized mouse experiments. **Funding:** This
672 work was supported by grants from the NIH (5U19AI117941, AI110297, AI145038, HL125030,
673 HL126544), CIRM (DR1-01431, TRX-01431-1), the James B. Pendleton Charitable Trust, and the
674 McCarthy Family Foundation (I.S.Y.C).
675 **Competing interests:** Dr. Irvin S.Y. Chen has a financial interest in CSL Behring and Calimmune Inc. No
676 funding was provided by these companies to support this work; Dr. Dong Sung An has a financial interest
677 in Calimmune Inc and CSL Behring that the University of California Regents have licensed intellectual
678 property invented by Dong Sung An, that is being used in the research, to Calimmune Inc. No funding was
679 provided by these companies to support this work. All other authors declare no competing interests.
680
681 Reference:
682 1. Shultz LD, Ishikawa F, Greiner DL: **Humanized mice in translational biomedical research**. *Nat*
683 *Rev Immunol* 2007, **7**(2):118-130.
684 2. Saito Y, Shultz LD, Ishikawa F: **Understanding Normal and Malignant Human Hematopoiesis**
685 **Using Next-Generation Humanized Mice**. *Trends Immunol* 2020, **41**(8):706-720.

686 3. Garcia-Beltran WF, Claiborne DT, Maldini CR, Phelps M, Vrbanac V, Karpel ME, Krupp KL, Power
687 KA, Boutwell CL, Balazs AB *et al*: **Innate Immune Reconstitution in Humanized Bone Marrow-**
688 **Liver-Thymus (HuBLT) Mice Governs Adaptive Cellular Immune Function and Responses to**
689 **HIV-1 Infection**. *Front Immunol* 2021, **12**:667393.
690 4. Melkus MW, Estes JD, Padgett-Thomas A, Gatlin J, Denton PW, Othieno FA, Wege AK, Haase AT,
691 Garcia JV: **Humanized mice mount specific adaptive and innate immune responses to EBV and**
692 **TSST-1**. *Nat Med* 2006, **12**(11):1316-1322.
693 5. Wen J, Wu D, Qin M, Liu C, Wang L, Xu D, Vinters HV, Liu Y, Kranz E, Guan X *et al*: **Sustained**
694 **delivery and molecular targeting of a therapeutic monoclonal antibody to metastases in the**
695 **central nervous system of mice**. *Nature Biomedical Engineering* 2019, **3**(9):706-716.
696 6. Heipertz EL, Walker WE: **Creation of BLT Humanized Mice for Sepsis Studies**. In: *Methods in*
697 *Molecular Biology.* Springer US; 2021: 137-154.
698 7. Lavender KJ, Williamson BN, Saturday G, Martellaro C, Griffin A, Hasenkrug KJ, Feldmann H,
699 Prescott J: **Pathogenicity of Ebola and Marburg Viruses Is Associated With Differential**
700 **Activation of the Myeloid Compartment in Humanized Triple Knockout-Bone Marrow, Liver,**
701 **and Thymus Mice**. *The Journal of Infectious Diseases* 2018, **218**(suppl_5):S409-S417.
702 8. Crawford LB, Tempel R, Streblow DN, Kreklywich C, Smith P, Picker LJ, Nelson JA, Caposio P:
703 **Human Cytomegalovirus Induces Cellular and Humoral Virus-specific Immune Responses in**
704 **Humanized BLT Mice**. *Sci Rep* 2017, **7**(1):937.
705 9. Khamaikawin W, Shimizu S, Kamata M, Cortado R, Jung Y, Lam J, Wen J, Kim P, Xie Y, Kim S *et*
706 *al*: **Modeling Anti-HIV-1 HSPC-Based Gene Therapy in Humanized Mice Previously Infected**
707 **with HIV-1**. *Mol Ther Methods Clin Dev* 2018, **9**:23-32.
708 10. Maldini CR, Claiborne DT, Okawa K, Chen T, Dopkin DL, Shan X, Power KA, Trifonova RT, Krupp
709 K, Phelps M *et al*: **Dual CD4-based CAR T cells with distinct costimulatory domains mitigate**
710 **HIV pathogenesis in vivo**. *Nat Med* 2020, **26**(11):1776-1787.
711 11. Biasco L, Pellin D, Scala S, Dionisio F, Basso-Ricci L, Leonardelli L, Scaramuzza S, Baricordi C,
712 Ferrua F, Cicalese MP *et al*: **In Vivo Tracking of Human Hematopoiesis Reveals Patterns of**
713 **Clonal Dynamics during Early and Steady-State Reconstitution Phases**. *Cell Stem Cell* 2016,
714 **19**(1):107-119.
715 12. Kim S, Kim N, Presson AP, Metzger ME, Bonifacino AC, Sehl M, Chow SA, Crooks GM, Dunbar
716 CE, An DS *et al*: **Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-**
717 **long clonal-tracking study**. *Cell Stem Cell* 2014, **14**(4):473-485.
718 13. Verovskaya E, Broekhuis MJ, Zwart E, Ritsema M, van Os R, de Haan G, Bystrykh LV:
719 **Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative**
720 **clonal analysis using cellular barcoding**. *Blood* 2013, **122**(4):523-532.
721 14. Cheung AM, Nguyen LV, Carles A, Beer P, Miller PH, Knapp DJ, Dhillon K, Hirst M, Eaves CJ:
722 **Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord**
723 **blood cells in NSG mice**. *Blood* 2013, **122**(18):3129-3137.
724 15. Brewer C, Chu E, Chin M, Lu R: **Transplantation Dose Alters the Differentiation Program of**
725 **Hematopoietic Stem Cells**. *Cell Rep* 2016, **15**(8):1848-1857.
726 16. Thielecke L, Aranyossy T, Dahl A, Tiwari R, Roeder I, Geiger H, Fehse B, Glauche I, Cornils K:
727 **Limitations and challenges of genetic barcode quantification**. *Sci Rep* 2017, **7**:43249.
728 17. Barber-Axthelm IM, Barber-Axthelm V, Sze KY, Zhen A, Suryawanshi GW, Chen IS, Zack JA,
729 Kitchen SG, Kiem HP, Peterson CW: **Stem cell-derived CAR T cells traffic to HIV reservoirs in**
730 **macaques**. *JCI Insight* 2021, **6**(1).
731 18. Suryawanshi GW, Khamaikawin W, Wen J, Shimizu S, Arokium H, Xie Y, Wang E, Kim S, Choi H,
732 Zhang C *et al*: **The clonal repopulation of HSPC gene modified with anti–HIV-1 RNAi is not**
733 **affected by preexisting HIV-1 infection**. *Science Advances* 2020, **6**(30):eaay9206.
734 19. Modlich U, Navarro S, Zychlinski D, Maetzig T, Knoess S, Brugman MH, Schambach A, Charrier
735 S, Galy A, Thrasher AJ *et al*: **Insertional Transformation of Hematopoietic Cells by Self-**

736 **inactivating Lentiviral and Gammaretroviral Vectors**. *Molecular Therapy* 2009, **17**(11):1919-
737 1928.
738 20. Montini E, Cesana D, Schmidt M, Sanvito F, Bartholomae CC, Ranzani M, Benedicenti F, Sergi LS,
739 Ambrosi A, Ponzoni M *et al*: **The genotoxic potential of retroviral vectors is strongly modulated**
740 **by vector design and integration site selection in a mouse model of HSC gene therapy**. *J Clin*
741 *Invest* 2009, **119**(4):964-975.
742 21. Espinoza DA, Fan X, Yang D, Cordes SF, Truitt LL, Calvo KR, Yabe IM, Demirci S, Hope KJ, Hong
743 SG *et al*: **Aberrant Clonal Hematopoiesis following Lentiviral Vector Transduction of HSPCs**
744 **in a Rhesus Macaque**. *Mol Ther* 2019, **27**(6):1074-1086.
745 22. Cattoglio C, Facchini G, Sartori D, Antonelli A, Miccio A, Cassani B, Schmidt M, von Kalle C, Howe
746 S, Thrasher AJ *et al*: **Hot spots of retroviral integration in human CD34+ hematopoietic cells**.
747 *Blood* 2007, **110**(6):1770-1778.
748 23. Cattoglio C, Pellin D, Rizzi E, Maruggi G, Corti G, Miselli F, Sartori D, Guffanti A, Di Serio C,
749 Ambrosi A *et al*: **High-definition mapping of retroviral integration sites identifies active**
750 **regulatory elements in human multipotent hematopoietic progenitors**. *Blood* 2010,
751 **116**(25):5507-5517.
752 24. Esteban JA, Salas M, Blanco L: **Fidelity of phi 29 DNA polymerase. Comparison between protein-**
753 **primed initiation and DNA polymerization**. *J Biol Chem* 1993, **268**(4):2719-2726.
754 25. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler
755 D *et al*: **Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand**
756 **displacement whole genome amplification**. *Nucleic Acids Res* 2004, **32**(9):e71.
757 26. He F, Zhou W, Cai R, Yan T, Xu X: **Systematic assessment of the performance of whole-genome**
758 **amplification for SNP/CNV detection and beta-thalassemia genotyping**. *J Hum Genet* 2018,
759 **63**(4):407-416.
760 27. Einkauf KB, Lee GQ, Gao C, Sharaf R, Sun XM, Hua S, Chen SMY, Jiang CY, Lian XD, Chowdhury
761 FZ *et al*: **Intact HIV-1 proviruses accumulate at distinct chromosomal positions during**
762 **prolonged antiretroviral therapy**. *J Clin Invest* 2019, **129**(3):988-998.
763 28. Patro SC, Brandt LD, Bale MJ, Halvas EK, Joseph KW, Shao W, Wu XL, Guo S, Murrell B, Wiegand
764 A *et al*: **Combined HIV-1 sequence and integration site analysis informs viral dynamics and**
765 **allows reconstruction of replicating viral ancestors**. *P Natl Acad Sci USA* 2019, **116**(51):25891-
766 25899.
767 29. Shimizu S, Hong P, Arumugam B, Pokomo L, Boyer J, Koizumi N, Kittipongdaja P, Chen A, Bristol
768 G, Galic Z *et al*: **A highly efficient short hairpin RNA potently down-regulates CCR5 expression**
769 **in systemic lymphoid organs in the hu-BLT mouse model**. *Blood* 2010, **115**(8):1534-1544.
770 30. Calvanese V, Nguyen AT, Bolan TJ, Vavilina A, Su T, Lee LK, Wang Y, Lay FD, Magnusson M,
771 Crooks GM *et al*: **MLLT3 governs human haematopoietic stem-cell self-renewal and**
772 **engraftment**. *Nature* 2019, **576**(7786):281-286.
773 31. Kim S, Kim N, Presson AP, An DS, Mao SH, Bonifacino AC, Donahue RE, Chow SA, Chen IS:
774 **High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones**. *J*
775 *Virol* 2010, **84**(22):11771-11780.
776 32. Suryawanshi GW, Xu S, Xie Y, Chou T, Kim N, Chen ISY, Kim S: **Bidirectional Retroviral**
777 **Integration Site PCR Methodology and Quantitative Data Analysis Workflow**. *J Vis Exp*
778 2017(124).
779 33. Zhao M, Lee WP, Garrison EP, Marth GT: **SSW library: an SIMD Smith-Waterman C/C++**
780 **library for use in genomic applications**. *PLoS One* 2013, **8**(12):e82138.
781 34. Rényi A: **On measures of entropy and information**. *Proceedings of the Fourth Berkeley Symposium*
782 *on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* 1961.
783 35. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR:
784 **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

785 36. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL,
786 Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with**
787 **TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-578.
788 37. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T:
789 **deepTools2: a next generation web server for deep-sequencing data analysis**. *Nucleic Acids Res*
790 2016, **44**(W1):W160-165.
791 38. Koo TK, Li MY: **A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for**
792 **Reliability Research (vol 15, pg 155, 2016)**. *J Chiropr Med* 2017, **16**(4):346-346.
793 39. Wu C, Li B, Lu R, Koelle SJ, Yang Y, Jares A, Krouse AE, Metzger M, Liang F, Lore K *et al*: **Clonal**
794 **tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer**
795 **cells**. *Cell Stem Cell* 2014, **14**(4):486-499.
796 40. Aiuti A, Biasco L, Scaramuzza S, Ferrua F, Cicalese MP, Baricordi C, Dionisio F, Calabria A,
797 Giannelli S, Castiello MC *et al*: **Lentiviral hematopoietic stem cell gene therapy in patients with**
798 **Wiskott-Aldrich syndrome**. *Science* 2013, **341**(6148):1233151.
799 41. Shannon CE: **A Mathematical Theory of Communication**. *Bell Syst Tech J* 1948, **27**(3):379-423.
800 42. Simpson EH: **Measurement of Diversity**. *Nature* 1949, **163**(4148):688-688.
801 43. Wu C, Espinoza DA, Koelle SJ, Potter EL, Lu R, Li B, Yang D, Fan X, Donahue RE, Roederer M *et*
802 *al*: **Geographic clonal tracking in macaques provides insights into HSPC migration and**
803 **differentiation**. *J Exp Med* 2018, **215**(1):217-232.
804 44. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F: **HIV-1 integration in the human**
805 **genome favors active genes and local hotspots**. *Cell* 2002, **110**(4):521-529.
806 45. Haworth KG, Schefter LE, Norgaard ZK, Ironside C, Adair JE, Kiem HP: **HIV infection results in**
807 **clonal expansions containing integrations within pathogenesis-related biological pathways**. *JCI*
808 *Insight* 2018, **3**(13).
809 46. Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD:
810 **Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences**. *PLoS*
811 *Biol* 2004, **2**(8):E234.
812 47. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD: **HIV integration site selection: analysis by**
813 **massively parallel pyrosequencing reveals association with epigenetic modifications**. *Genome*
814 *Res* 2007, **17**(8):1186-1194.
815 48. Lu R, Czechowicz A, Seita J, Jiang D, Weissman IL: **Clonal-level lineage commitment pathways**
816 **of hematopoietic stem cells in vivo**. *Proc Natl Acad Sci U S A* 2019, **116**(4):1447-1456.
817 49. Rodriguez-Fraticelli AE, Weinreb C, Wang SW, Migueles RP, Jankovic M, Usart M, Klein AM,
818 Lowell S, Camargo FD: **Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis**.
819 *Nature* 2020, **583**(7817):585-589.
820 50. Belderbos ME, Jacobs S, Koster TK, Ausema A, Weersing E, Zwart E, de Haan G, Bystrykh LV:
821 **Donor-to-Donor Heterogeneity in the Clonal Dynamics of Transplanted HumanCord Blood**
822 **Stem Cellsin Murine Xenografts**. *Biol Blood Marrow Transplant* 2020, **26**(1):16-25.
823 51. Ferrari S, Jacob A, Beretta S, Unali G, Albano L, Vavassori V, Cittaro D, Lazarevic D, Brombin C,
824 Cugnata F *et al*: **Efficient gene editing of human long-term hematopoietic stem cells validated by**
825 **clonal tracking**. *Nat Biotechnol* 2020.
826 52. Lewinski MK, Yamashita M, Emerman M, Ciuffi A, Marshall H, Crawford G, Collins F, Shinn P,
827 Leipzig J, Hannenhalli S *et al*: **Retroviral DNA integration: viral and cellular determinants of**
828 **target-site selection**. *PLoS Pathog* 2006, **2**(6):e60.
829 53. Lelek M, Casartelli N, Pellin D, Rizzi E, Souque P, Severgnini M, Di Serio C, Fricke T, Diaz-Griffero
830 F, Zimmer C *et al*: **Chromatin organization at the nuclear pore favours HIV replication**. *Nat*
831 *Commun* 2015, **6**.
832 54. Marini B, Kertesz-Farkas A, Ali H, Lucic B, Lisek K, Manganaro L, Pongor S, Luzzati R, Recchia
833 A, Mavilio F *et al*: **Nuclear architecture dictates HIV-1 integration site selection**. *Nature* 2015,
834 **521**(7551):227-231.

835 55. Llano M, Saenz DT, Meehan A, Wongthida P, Peretz M, Walker WH, Teo W, Poeschla EM: **An**
836 **essential role for LEDGF/p75 in HIV integration**. *Science* 2006, **314**(5798):461-464.
837 56. Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, Ecker JR, Bushman F: **A role for**
838 **LEDGF/p75 in targeting HIV DNA integration**. *Nat Med* 2005, **11**(12):1287-1289.
839 57. Vanegas M, Llano M, Delgado S, Thompson D, Peretz M, Poeschla E: **Identification of the**
840 **LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin**
841 **tethering**. *J Cell Sci* 2005, **118**(Pt 8):1733-1743.
842 58. Cherepanov P, Sun ZYJ, Rahman S, Maertens G, Wagner G, Engelman A: **Solution structure of the**
843 **HIV-1 integrase-binding domain in LEDGF/p75**. *Nat Struct Mol Biol* 2005, **12**(6):526-532.
844 59. Llano M, Delgado S, Vanegas M, Poeschla EM: **Lens epithelium-derived growth factor/p75**
845 **prevents proteasomal degradation of HIV-1 integrase**. *J Biol Chem* 2004, **279**(53):55570-55577.
846 60. Eidahl JO, Crowe BL, North JA, McKee CJ, Shkriabai N, Feng L, Plumb M, Graham RL, Gorelick
847 RJ, Hess S *et al*: **Structural basis for high-affinity binding of LEDGF PWWP to**
848 **mononucleosomes**. *Nucleic Acids Res* 2013, **41**(6):3924-3936.
849 61. Pradeepa MM, Sutherland HG, Ule J, Grimes GR, Bickmore WA: **Psip1/Ledgf p52 binds**
850 **methylated histone H3K36 and splicing factors and contributes to the regulation of alternative**
851 **splicing**. *PLoS Genet* 2012, **8**(5):e1002717.
852 62. Hon G, Wang W, Ren B: **Discovery and annotation of functional chromatin signatures in the**
853 **human genome**. *PLoS Comput Biol* 2009, **5**(11):e1000566.
854 63. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T: **Regulation of**
855 **alternative splicing by histone modifications**. *Science* 2010, **327**(5968):996-1000.
856 64. Schiroli G, Conti A, Ferrari S, Della Volpe L, Jacob A, Albano L, Beretta S, Calabria A, Vavassori
857 V, Gasparini P *et al*: **Precise Gene Editing Preserves Hematopoietic Stem Cell Function following**
858 **Transient p53-Mediated DNA Damage Response**. *Cell Stem Cell* 2019, **24**(4):551-565 e558.

859

860

861

862

863

864

865

866

867

868

869 **Figures-**

870 **Figure 1: LoVIS-Seq reproduces clonal distribution of entire mouse blood using 25µl blood: a**)
871 Diagram showing Anti-HIV-EGFP-WT and control mCherry-H1/H5 vectors having WT, H1, or H5 LTR-
872 index and strategy for VIS assay with LTRi-seq. LTR index sequences are shown in magenta box. **b**) Hu-
873 BLT mouse model: Fetal liver CD34+ cells were separately transduced with either anti-HIV or control
874 vectors and transduced cells were mixed 1:1 for transplant. The mix of transduced cells was transplanted in
875 myeloablated NSG mice with a fetal thymus tissue implant. **c**) Stacked bar plot showing clonal frequencies
876 of VIS in BM of hu-BLT mouse. Clones from mCherry-H1 and EGFP-WT cells were identified by

877 corresponding LTR barcodes. In the stacked bar plot, each band represents a unique VIS (HSPC clone) and
878 thickness of the band shows clonal frequency or abundance of that HSPC clone. Percentage of mCherry+ or
879 EGFP+ cells within human cell (hCD45+) population are shown on top of the corresponding stacked-bar.
880 **d**) Plot showing Pearson's r for correlations of mCherry-H1 (red dots) and EGFP-WT (green dots) VIS
881 clonal profiles between unamplified DNA replicates and replicates of MDA-amplified DNA samples for
882 different cell numbers. **e**) Experimental protocol for longitudinal clonal tracking in humanized BLT mice.
883 **f**) Scatter plot showing VIS clonal frequencies between unamplified whole blood DNA and two replicates
884 of MDA-amplified DNA from 25µl blood at week 19 (r= Pearson's r, diagonal line is r=1) for m599 and
885 m598. Clonal frequency of mCherry-H5 (red dots) and EGFP-WT (green dots) VIS clones in unamplified
886 DNA samples (y-axis) and MDA replicates (x-axis).
887 **Figure 2: Longitudinal clonal tracking in hu-BLT mice: a**) Area plots show clonal repopulation in
888 whole blood over time from week 13-19. Each colored band is a unique VIS clone and thickness of the
889 band corresponds to frequency of the VIS clone. Dashed black line separates mCherry-H5 VIS clones
890 (below) and EGFP-WT VIS clones (above). **b**) Line plots show changes over time in the total frequency of
891 mCherry-H1 VIS clones (solid red line) and total frequency of EGFP-WT VIS clones (solid green line) as
892 well as percentages of mCherry+ cells (dashed red line) and EGFP+ cells (dashed green line). **c**) Heatmaps
893 showing percentage change in shared clones between two timepoints. Digits inside white tiles on the
894 diagonal show number of VIS detected at each time point. Colors of each heatmap tile correspond to
895 percentage of clones shared and color key is provided on the right. Digits in each tile show the number of
896 VIS shared between two timepoints. **d**) Renyi's diversity profiles evaluated using raw count data from two
897 replicates at each timepoint and by varying value of alpha. Renyi's diversity profiles are arranged with
898 highest diversity at the top to lowest at the bottom. Topmost curve with no overlap or intersection with any
899 other curve has the highest overall diversity. Diversity of curves that overlap or intersect is undefined. **e**)
900 Line plot showing contribution of highest contributing clone at different timepoints. Values reported are
901 exp(H(alpha)) at alpha=∞.
902 **Figure 3: Unique clonal sharing pattern between different tissues:** Polar area plots of clonal expansion
903 and sharing in peripheral blood, spleen, and bone marrow (BM). There are three axes, one for each tissue.
904 Stacked bar plot on each axis shows size distribution of clones in the tissue. Each colored stack represents a
905 VIS clone and its thickness shows abundance of the clone. Clones shared between tissues are connected
906 using ribbons with colors matching the clone's stack color in the bar plot. Black line encompasses total size
907 distribution of persistent clones that continuously contributed from week 13 till 19. Clones did not
908 continuously contribute for 6 weeks and were detected at week 19 in blood, spleen, and BM are outside
909 black line. Pearson's r values are shown in black.

910 **Figure 4: Chromosomal distribution of VIS and its bias for transcriptionally active genes: a**) Circos
911 plot shows genomic location of all 897 VIS from mice m599, m598, and m591. Box plots in the center
912 show maximum frequency of mCherry-H5 (filled red dots) and EGFP-WT (filled green dots) VIS clones in
913 mice m599, m598, and m591 over 6 weeks. Genomic location of mCherry-H5 (filled red dots) and EGFP-
914 WT (filled green dots) VIS clones are plotted on three concentric circles depending on their maximum
915 frequency over 6 weeks: Low frequency clones with maximum frequency below the 1st quartile value
916 (innermost concentric circle), High frequency clones above 3rd quartile value (outermost concentric circle),
917 and Medium frequency clones between 1st and 3rd quartile (middle concentric circle). Top 10 high
918 frequency VIS clones are shown in darker colors. Color coded short line segments under each VIS show
919 gene biotypes of VIS-proximal genes. Gene symbols above ideograms represent genes proximal to the top
920 10 VIS clones from mice m599 (blue), m598 (brown), and m591 (light pink). **b**) Classification of all,
921 persistent, and Top 10 VIS clones based on biotype of VIS-proximal gene. Inner pie chart shows clones
922 classified based on gene biotype of the most proximal gene. Outer Donut plots show number of VIS (in
923 bracket % of VIS) proximal to active (dark color) or inactive (faded colors) genes. Active proximal genes
924 have FPKM >1.
925 **Figure 5: Epigenetic determinants of vector integration: a**) Profile plots and heatmap for 10 chromatin
926 features and input sample in active and inactive VIS-proximal genes in uncultured FL-HSPC. Profile plots
927 show mean score for active (blue line) and inactive (red line) proximal genes. Score is calculated from
928 normalized read count (RPKM) for each sample. Each row in heatmap shows expression level of 10
929 chromatin features in proximal genes from TSS to TES with 2Kb flanks upstream and downstream. Color
930 scale key shows range of normalized expression. **b**) Profile plots and heatmaps for 10 chromatin features in
931 region flanking ±1Kb of each VIS. Profile plots showing mean scores over ±1Kb region flanking VIS.
932 Each row in the heatmap shows the expression level of 10 chromatin features in regions flanking ±1Kb of
933 VIS. Individual color scale key shows the range of normalized expression for corresponding features. **c**)
934 Bar plots show number of VIS within ±1Kb of enriched region (peak) of different chromatin features. VIS
935 clones and random IS are binned by absolute distance in base pairs (bp) between the enriched region and
936 IS. Bars show number of VIS in each bin. Top panel shows binning for all VIS, middle panel shows top 10
937 VIS clones, and bottom panel shows random IS falling within ±1Kb of enriched region.
938
939
940
941
942
943

Figure 1

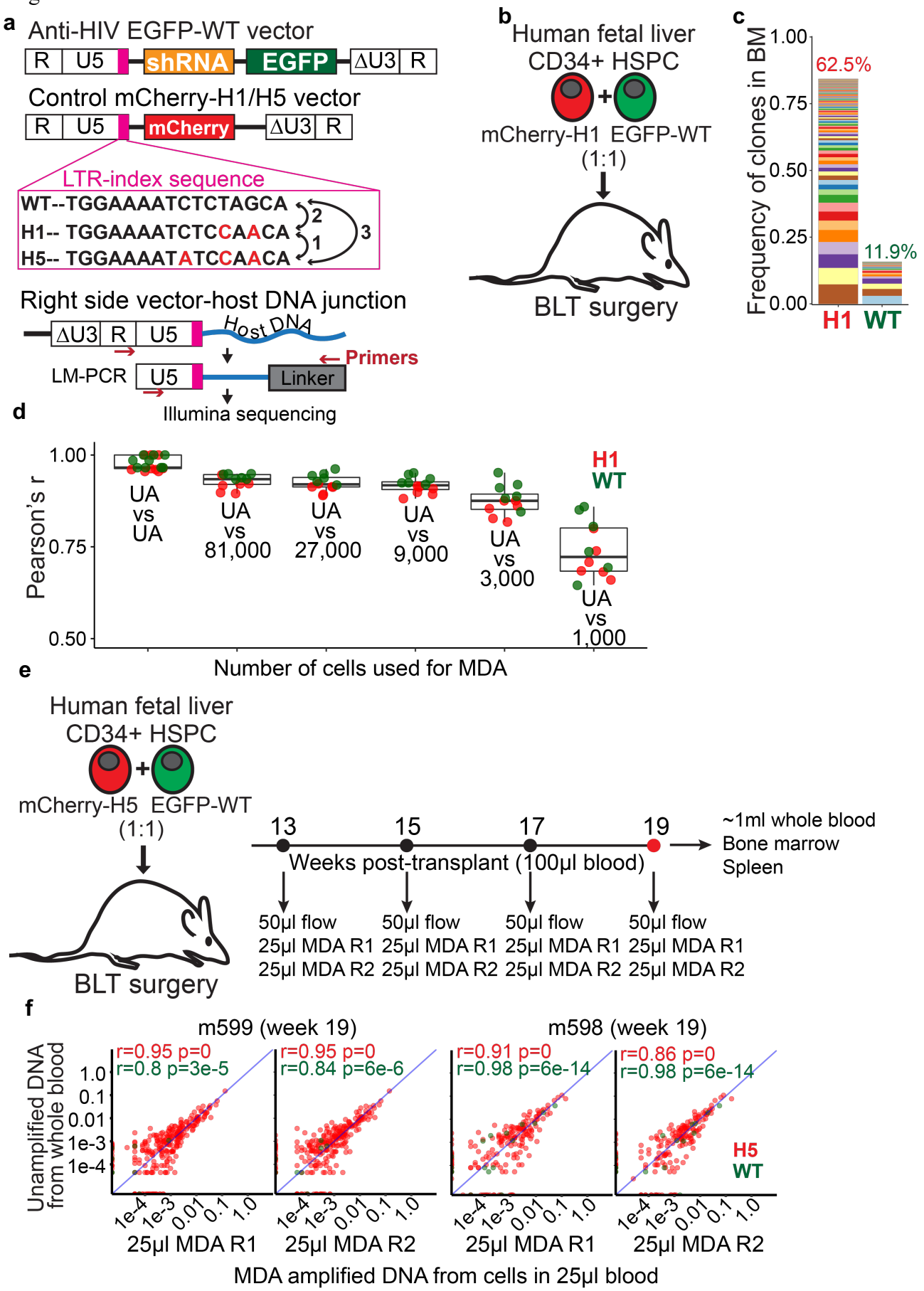

a
Anti-HIV EGFP-WT vector
R U5 shRNA EGFP ΔU3 R
Control mCherry-H1/H5 vector
R U5 mCherry ΔU3 R
LTR-index sequence
WT--TGGAAAATCTCTAGCA
H1-- TGGAAAATCTCCAACA
H5-- TGGAAAATATCCAACA
2
1
3
Right side vector-host DNA junction
ΔU3 R U5
Host DNA
Primers
LM-PCR
U5
Linker
Illumina sequencing
b
Human fetal liver
CD34+ HSPC
mCherry-H1 EGFP-WT
(1:1)
BLT surgery
c
Frequency of clones in BM
1.00
0.75
0.50
0.25
0.00
62.5%
11.9%
H1
WT
d
Pearson's r
1.00
0.75
0.50
UA vs UA
UA vs 81,000
UA vs 27,000
UA vs 9,000
UA vs 3,000
UA vs 1,000
H1
WT
Number of cells used for MDA
e
Human fetal liver
CD34+ HSPC
mCherry-H5 EGFP-WT
(1:1)
BLT surgery
13
15
17
19
Weeks post-transplant (100µl blood)
~1ml whole blood
Bone marrow
Spleen
50µl flow
25µl MDA R1
25µl MDA R2
f
m599 (week 19)
m598 (week 19)
Unamplified DNA from whole blood
r=0.95 p=0
r=0.8 p=3e-5
r=0.95 p=0
r=0.84 p=6e-6
r=0.91 p=0
r=0.98 p=6e-14
r=0.86 p=0
r=0.98 p=6e-14
1.0
0.1
0.01
1e-3
1e-4
H5
WT
25µl MDA R1
25µl MDA R2
MDA amplified DNA from cells in 25µl blood

Figure 2

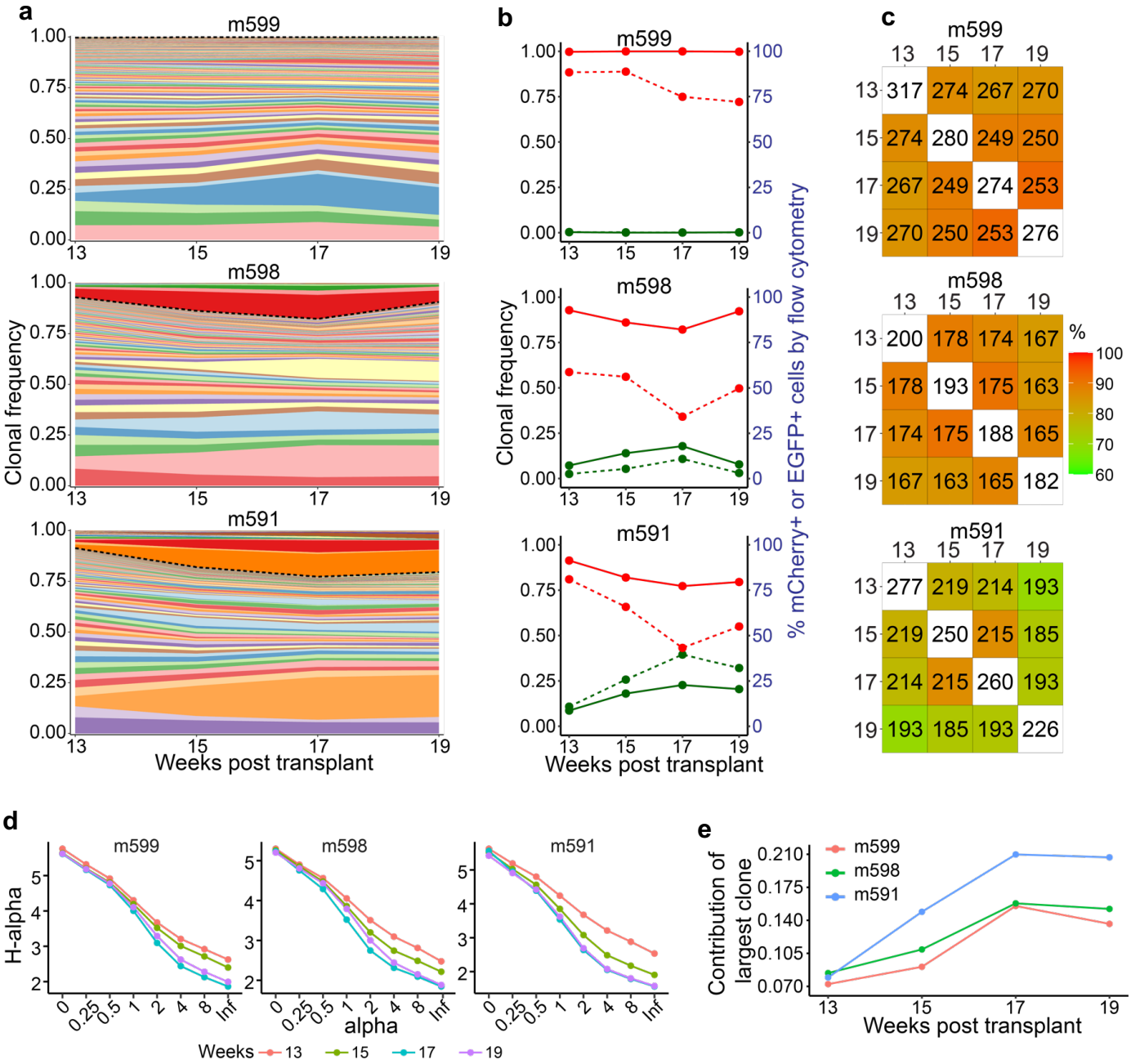
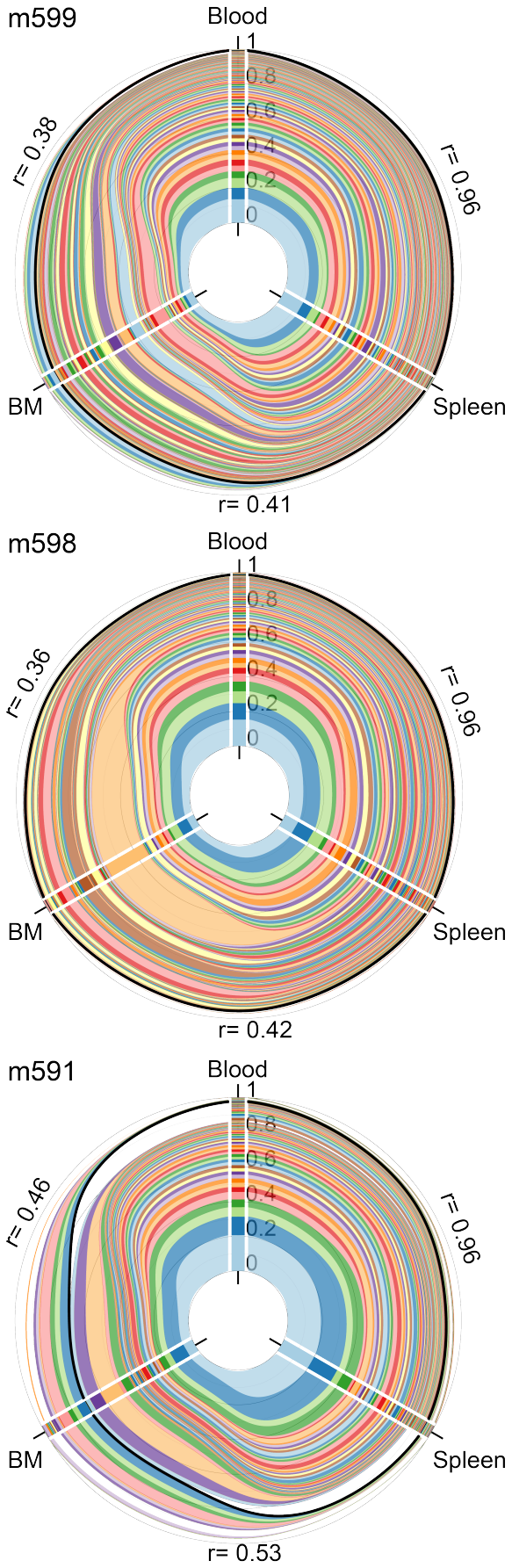
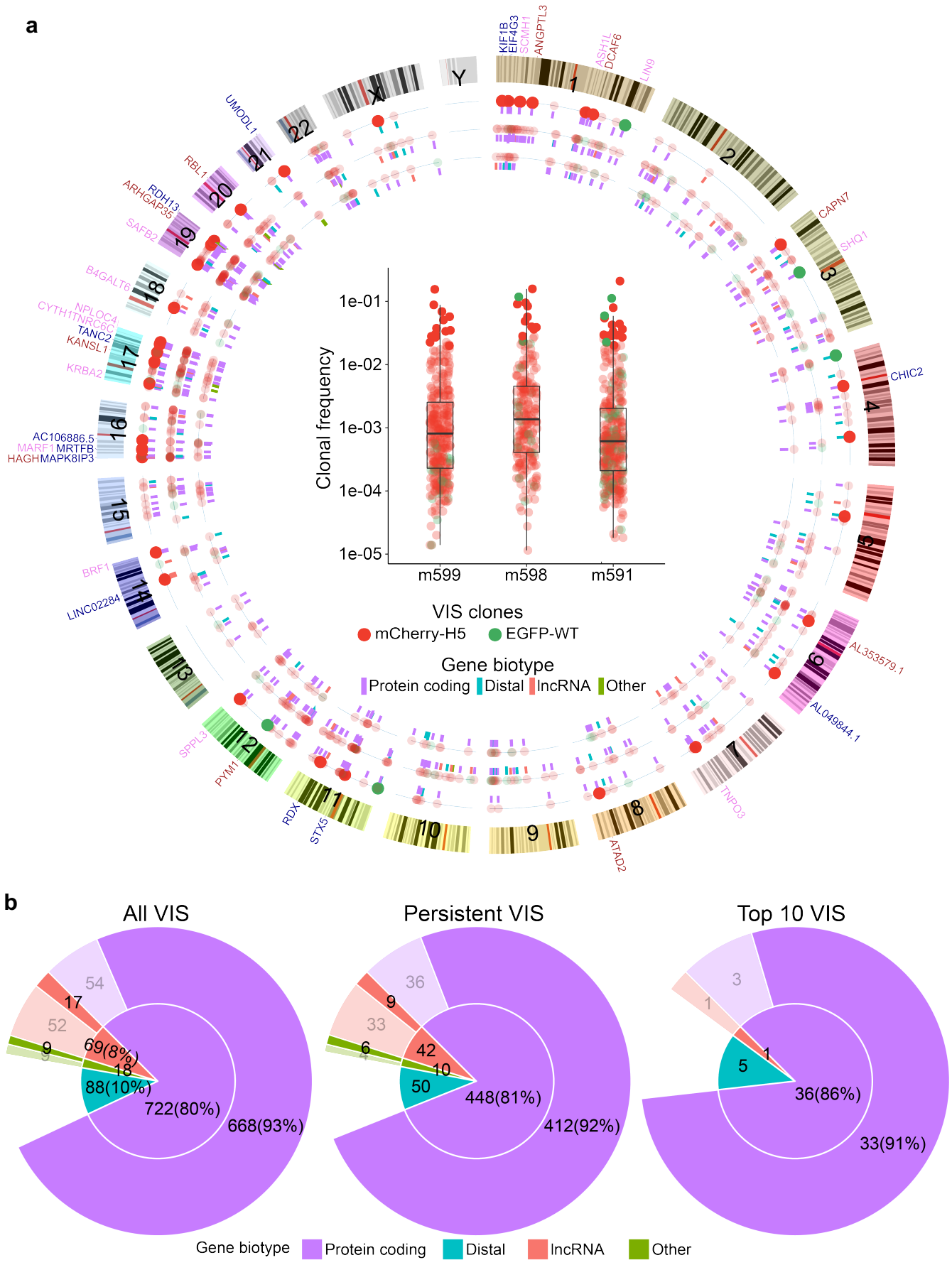

a
m599
m598
m591
Clonal frequency
Weeks post transplant
b
% mCherry+ or EGFP+ cells by flow cytometry
c
m599
13 15 17 19
13 317 274 267 270
15 274 280 249 250
17 267 249 274 253
19 270 250 253 276
m598
13 200 178 174 167
15 178 193 175 163
17 174 175 188 165
19 167 163 165 182
%
100 90 80 70 60
m591
13 277 219 214 193
15 219 250 215 185
17 214 215 260 193
19 193 185 193 226
d
H-alpha
alpha
0 0.25 0.5 1 2 4 8 Inf
Weeks 13 15 17 19
e
Contribution of largest clone
m599
m598
m591
Weeks post transplant

Figure 3

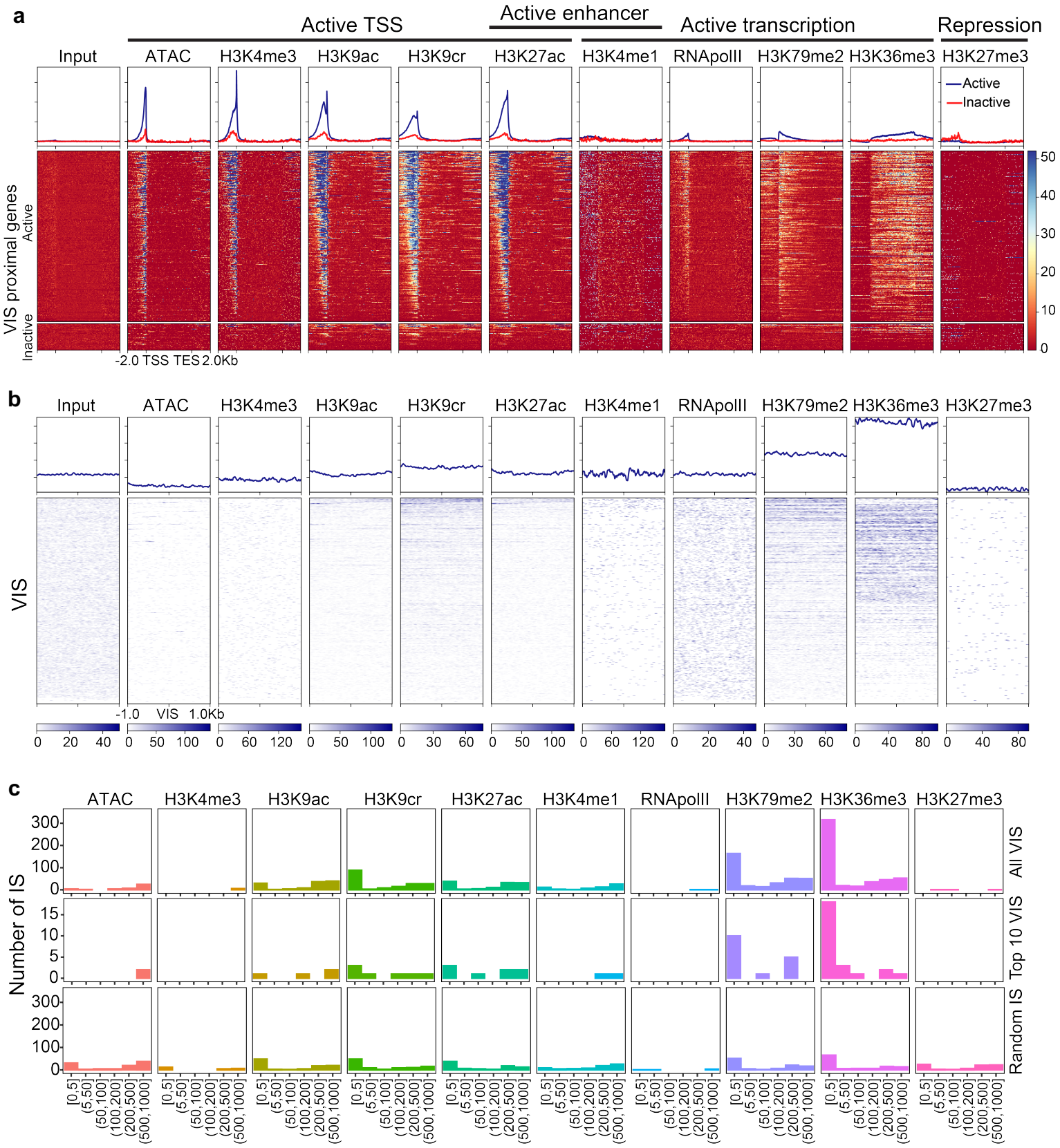

m599
Blood
1
0.8
0.6
0.4
0.2
0
r= 0.38
r= 0.96
BM
Spleen
r= 0.41
m598
Blood
1
0.8
0.6
0.4
0.2
0
r= 0.36
r= 0.96
BM
Spleen
r= 0.42
m591
Blood
1
0.8
0.6
0.4
0.2
0
r= 0.46
r= 0.96
BM
Spleen
r= 0.53

965 Figure 4



966

Figure 5


a
Active TSS
Active enhancer
Active transcription
Repression
Input
ATAC
H3K4me3
H3K9ac
H3K9cr
H3K27ac
H3K4me1
RNApolII
H3K79me2
H3K36me3
H3K27me3
Active
Inactive
VIS proximal genes
Active
Inactive
-2.0 TSS TES 2.0Kb
b
VIS
-1.0 VIS 1.0Kb
c
Number of IS
All VIS
Top 10 VIS
Random IS
[0,5]
(5,50]
(50,100]
(100,200]
(200,500]
(500,1000]