Title:-

Longitudinal tracking reveals sustained polyclonal repopulation of human-HSPC in humanized mice 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:

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 revealed that in vivo repopulating clones have significant vector integration bias 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]. Due to low errors and high coverage, MDA can be used for various sequence sensitive application such as single nucleotide polymorphism (SNP) and next generation sequencing studies[11]. MDA-amplified DNA has been used to detect retroviral IS[12] and to sequence full-length HIV-1 proviruses including the IS[13,14].

Although self-inactivating lentiviral vectors are low risk, a strong promoter within the vector can upregulate the expression of endogenous genes where the vector integrated[15,16]. HIV-1 and HIV-1 based vectors are known to favor transcriptionally active gene dense regions[17,18], with preference for histone modification H3 acetylation, H4 acetylation and H3K4 methylation[19]. Other studies found no preference for DNase I hypersensitive sites[20] and H3K4 methylation being disfavored[21] and preference for H3K36me3 in Jurkat cells[21]. The HIV-1 integration occurs with assistance from nuclear pore complex and targets the active genes closer to the nuclear pore and disfavor heterochromatin regions and active regions located centrally in the nuclease[22]. However, the implications of HIV-1 integration on cell fate are compounded by infection induced cytotoxicity. An in vitro study using activated human CD34+ HSC also found lentiviral vector integration preference for active genes[23]. These in vitro studies have tracked impact of vector integration on cell fate over short time however long-term impact is only partially explored. Moreover, in human HSPC—more specifically human fetal HSPC (FL-HSPC)— vector integration preference for epigenetic features is

unexplored. Importantly, influence of VIS-proximal epigenetic features on in vivo survival, proliferation, and differentiation of gene-modified HSPC is unknown.

In this study, we present LoVIS-Seq, a combined MDA and VIS assay for low-volume samples. Using this assay, we longitudinally track hundreds of clones in 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 that vector integration in vivo detected clones 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:

Minimum 10,000 bone marrow cells or 25µl blood is sufficient for LoVIS-Seq:

To test our new assay, we collected bone marrow (BM) cells from hu-BLT (bone marrow-liver-thymus) mouse (m860). Fetal liver CD34+ cells transduced with sh1005(anti-CCR5 shRNA)-EGFP vector or control mCherry vector were mixed in equal ratio and transplanted in the mouse (Figure 1a-b, details in Methods). We estimated clonal composition of the EGFP-WT (WT LTR-index) and mCherry-H1 (H1 LTR-index) populations using unamplified bulk DNA, in triplicate, as described previously[8]. A total of 300 ±42 SD (216 ±22 SD mCherry-H1 VIS and 84 ±20 SD EGFP-WT) VIS were detected. The polyclonal profile in mouse bone marrow (Figure 1c) resembled that found in hu-BLT mice[8] and in autologously transplanted mice[6], nonhuman primates[5,24], and humans[4,25]. Next, to test our LoVIS-Seq assay (Figure 1d) that combines MDA with VIS assays, we first performed MDA directly on 81,000, 27,000, 9,000, 3,000, and 1,000 bone marrow cells, each in duplicate (Supplementary figure 1A, details in Methods section). Equal amounts of MDA-amplified DNA and unamplified bulk DNA were used for the VIS assay (Supplementary table 1).

We found high reproducibility of clonal profiles in different MDA samples and within-MDA replicates of 81,000 to 9,000 cells (avg. Pearson's r value >0.91) (Figure 1e and Supplementary figure 1B-F); for less than 9,000 cells, the reproducibility dropped (avg. Pearson's r=0.87 for 3,000 and 0.73 for 1,000 cells). Importantly, reduced cell numbers caused a modest reduction in VIS detection (Supplementary figure 1G). These data validate that MDA-amplified DNA from >10,000 bone marrow cells is sufficient for LoVIS-Seq.

Next, to test accuracy of LoVIS-Seq with hu-BLT mouse blood, we collected 100µl blood at week 13, 15, 17 and ~1ml of whole blood at week 19 post-transplant. The hu-BLT mice were transplanted with an equal mix of human CD34+ cells transduced with anti-HIV EGFP-WT vector and control mCherry-H5 vector (Figure 1f). Cells from 50µl blood were used for flow cytometry and the remaining cells were used for MDA duplicates; each 25µl of blood (>10,000 human cells). High correlation (median Pearson's r =0.93) of mCherry-H5 and EGFP-WT VIS clonal frequencies between unamplified and MDA-amplified DNA from blood cells (Figure 1g) suggests that the clonal profile of entire mouse blood can be captured with 25µl of

blood. Importantly, the MDA replicates also showed high reproducibility (median Pearson's r>0.95, Supplementary figure 3A). In conclusion, our LoVIS-Seq assay accurately captured the clonality of two vector-modified cell populations in hu-BLT mouse blood using mere 25µl of blood or as few as 10,000 cells.

LoVIS-Seq for simultaneous clonal tracking of therapeutic vector-modified and control vector-modified populations:

After demonstrating accuracy and reproducibility of MDA amplified DNA samples, we used 25µl of hu-BLT mouse blood for LoVIS-Seq and LTR-indexes (Figure 1a, e) to track change in the relative frequencies of 792 mCherry-H5 and 105 EGFP-WT VIS (897 total) clones over 6 weeks (Figure 2a). High correlations between the total mCherry-H5 VIS clonal frequency and mCherry+ cell percentage by flow cytometry (Pearson's r = 0.8) as well as between total EGFP-WT VIS clonal frequency and EGFP+ cells percentage (Pearson's r = 0.9, Supplementary figure 3B) are consistent with our previous study[8]. Notably, the expansion in EGFP-WT clones coincided with reduction of mCherry-H5 contribution and vice-versa (Figure 2b; solid lines); these changes closely match the change in $EGFP^+$ and $mCherry^+$ cell percentages measured by flow cytometry (Figure 2b; dashed lines). Furthermore, repopulation in both EGFP-WT and mCherry-H5 populations is largely driven by expansion of a few HSPC clones—a characteristic feature of after-transplant repopulation. These results present a first-ever demonstration of clonal expansion of two populations, therapeutic vector-modified and control vector-modified, simultaneously repopulating in hu-BLT mice. Importantly, clonal data indicate two population competing to repopulate the mouse blood.

Sustained polyclonal repopulation with rapid clonal expansion and stabilization:

After identifying polyclonal repopulation of human cells in mice, we investigated the properties of its clonal dynamics. The maximum number of VIS were detected at week 13 and on average, only 13% (±5%) fewer VIS were detected at week 19 compared to week 13. Also, the number of total VIS contributing to repopulation at each timepoint decreased with time (Figure 2c). On average, 61% (±12.8%) of persistent VIS clones (m599: 235 clones, m598: 150 clones, and m591: 165 clones; total 550 clones) consistently contributed for 6 weeks and provided stable polyclonal repopulation (Figure 2a area plots). While the number of VIS clones steadily dropped over time in both the mCherry-H5 and EGFP-WT populations, their clonal profiles became increasingly similar (Supplementary figure 4A-B), comparable to polyclonal repopulation patterns that have been reported in nonhuman primates[5]. Moreover, correlation between time points indicates clonal distribution at week 13 differs from week 19, with clonal expansion stabilizing around week 17 (Supplementary figure 4B). We also examined Rényi's diversity[26] profiles for each animal at each time point (details in Methods section). The clonal diversity at week 13 was highest and subsequently decreased with time (Figure 2d). Diversities were similar between weeks 17 and 19 as indicated by their overlapping diversity profiles. The Shannon[27] and Simpson[28] indices dropped between weeks 13 and 17, indicating expansion of fewer clones (Supplementary figure 4C). The similarity of the indices between

weeks 17 and 19 suggests stabilization of clones. Contribution by the most frequent clone rose ~2.2 times, from 0.078 (±0.005, n=3) at week 13 to 0.174 (±0.030, n=3) at week 17 (Figure 2e), signaling rapid expansion of a few clones. Overall, the clonal repopulation remained normal despite faster expansion and earlier stabilization of human HSPC clones in hu-BLT mice compared to humans, wherein stable clones appear >1 year post-transplant[4].

Clonal sharing between organs reveals normal repopulation and unique tissue distribution pattern:

Our VIS data show normal clonal repopulation in blood of hu-BLT mice; however, in nonhuman primates, the early post-transplant clonal expansion patterns differ between blood and organs[29]. We performed VIS analysis on bulk cells from bone marrow (BM) and spleens of our hu-BLT mice to investigate whether the tissue/organ clonal expansion pattern differed from blood. We found a very similar clonal expansion pattern (avg. Pearson's r =0.94) between blood and spleen (Figure 3); however, clonal expansion in bone marrow differed from blood and spleen. Interestingly, we observed that in all three tissue compartments, persistent clones contributed the most to repopulation these results show normal clonal repopulation among the three tissue compartments with substantial clonal sharing.

Influence of genomic location and proximal genes on clonal growth:

For each VIS, our assay provided both relative frequency and genomic location of integration allowing us to monitor abnormal growth arising due to mutagenic insertions. Similar to the HIV-1 integration pattern[17], our VIS data from in vivo repopulating clones showed preference for high gene density chromosomes (Supplementary figure 5A). Additionally, similar genomic distribution of low, medium, and high frequency in vivo repopulating VIS clones (Figure 4a) suggests clonal expansion is unrelated to genomic location of integration. We found that in ~80% of the total detected clones, VIS occurred within ±1Kb of protein-coding genes, significantly higher than 46% of 1000 random IS ($p< 0.001$, Supplementary figure 6A). About 8% of VIS were within ±1Kb of long non-coding RNA (lncRNA) and 10% were outside ±1Kb of any genes (distal VIS, Figure 4b inside pie chart). Persistent and top 10 VIS clones (top 10 most frequent VIS clones from each timepoint) showed similar preference for gene biotypes (Figure 4B inside pie chart). Out of all VIS, only 66 (including 1 out of 42 top 10 VIS) were proximal to known cancer consensus genes (Figure 4a Circos plot). Gene ontology analysis of proximal genes and their mouse orthologs showed significant enrichment ($P < 0.01$) in various biological pathways such as cell-cell interaction, viral process and transcription regulation (Supplementary figure 5B). In vitro data for VIS-proximal gene in vector transduced human CD34+ HSC showed enrichment in similar biological processes[23]suggesting that biological function of VIS-proximal genes is unrelated to in vivo clonal expansion. Taken together, our results showed no clear link between in vivo clonal expansion and genomic location of vector integration or biological function of VIS-proximal genes.

Integration bias for transcriptionally active genes in human fetal liver HSPC:

Our data show VIS preference for genic regions, other studies using cell line and primary cells have reported similar vector integration bias for active genes with low to moderate expression[18]. However, the transcriptional state and expression level of the VIS-proximal genes is unknown in human FL-HSPC prior to vector integration. To address this, we analyzed transcriptomic (RNA-seq) and functional genomic (ATAC-seq and ChIP-seq) data from uncultured human FL-HSPC[30] isolated and processed using protocol identical to one used in our study (see Methods). Owing to the direct biological relevance of human FL-HSPC to our humanized BLT mice models, the multi-omics data is well suited to investigate the impact of vector integration on stemness of vector-modified HSPC.

The gene expression (RNA-seq) data show that of all detected clones, including the persistent clones and top 10 VIS clones, >77% VIS are within ±1Kb of transcriptionally active genes (FPKM>1) (Figure 4b outer donut chart). This is significantly higher than the ~27% of random IS proximal to active genes ($p < 0.001$, Supplementary figure 6A). The level of transcriptional activity of VIS-proximal genes (based on the FPKM value) was slightly higher than the median expression level of all active genes (Supplementary figure 5C). Moreover, the median gene activity level (FPKM) varied based on gene biotype, with protein coding proximal genes of all, persistent, and top 10 VIS clones having higher activity than lncRNA or pseudogenes of proximal genes. Similar to in vitro observations[23] our in vivo clonal tracking data show VIS prefer active genes but not highly active genes. These findings suggest that similar to in vitro, in vivo stability and expansion of the HSPC clone is likely linked to expression level of VIS proximal gene.

Vector integration favors actively transcribed regions:

We speculate that the chromatin structure of active genes strongly influences vector integration. Previous in vitro studies in cell lines suggested vector integration preference for genomic features such as select histone modification and DNase I hypersensitivity sites using data from different cell lines. However, such analysis is not available for human FL-HSPC and the effect of vector integration on stemness of these cells remains unexplored. To investigate this, we analyzed functional genomic data for 10 different chromatin features in uncultured FL-HSPC[30] (listed in Figure 5). These features include chromatin accessibility (ATAC-seq), active RNA polymerase II, and 8 histone modifications (ChIP-seq). For VIS-proximal active genes (FPKM>1), we found active transcription markers such as open chromatin region (ATAC-seq peaks) and histone marks (e.g. H3K4me3) near TSS as well as H3K36me3 and H3K79me2-enriched regions within the gene body. However, these marks were less prominent in inactive (FPKM<1) VIS-proximal genes (Figure 5a profile plots and heatmaps). Random IS-proximal genes showed similar enrichment profiles in active genes and their lack in inactive genes (Supplementary figure 6B). Further analysis found that out of 897 repopulating clones, 420 VIS (46% of total VIS) were within actively transcribed regions identified by enrichment for histone-modification marks H3K36me3 and/or H3K79me2. This is significantly higher

compared to only 10% random IS within actively transcribed regions ($p<0.001$, Supplementary figure 6C). Principal component analysis (PCA) on normalized enrichment levels (RPKM values) over a ±1Kb region of VIS (Figure 5b data) clearly separated H3K36me3 and H3K79me2 from all other features and showed no bias for random IS (Supplementary figure 7). Importantly, VIS avoided open DNA and transcription-regulator histone marks H3K4me3 and RNApolII as well as repression marker H3K27me3 (Figure 5b and 5c top panel). The top 10 most frequent VIS clones also displayed similar preference for all chromatin features (Figure 5c middle panel) and comparatively, random IS were evenly distributed across all 10 histone marks (Figure 5c bottom panel). Overall, the analysis reveals that vector integration in repopulating clones is a significantly biased for H3K36me3 and/or H3K79me2 enriched regions. It should be noted that despite this bias the clonal expansion of human HSPC in hu-BLT mice remained normal.

Discussion:

In this study we presented LoVIS-Seq, a new longitudinal clonal tracking method requiring a mere 25µl blood or less to monitor clonal behavior of gene-modified cells in small-animal models. LoVIS-Seq quantitatively captured the clonality of both control (mCherry-H1/H5) and anti-HIV (EGFP-WT) populations in whole blood. We provide the first-ever demonstration of simultaneous polyclonal repopulation of therapeutic vector-modified and control vector-modified cell populations in hu-BLT mouse blood. The polyclonal expansion resembles normal post-transplant HSPC clonal repopulation in mice[6], nonhuman primates[5,24] and humans[4,25]. Notably, the clonal frequency data recapitulated the flow cytometric measurements. Persistent clones are major contributors in blood, BM, and spleen. The multi-omics data from uncultured FL-HSPC revealed that vector integration in VIS clones that repopulated in mouse environment is significantly biased toward H3K36me3 and/or H3K79me2 enriched regions. Remarkably, this vector integration bias appears inconsequential with respect to clonal repopulation, as gene-modified HSPC differentiated normally in vivo; this confirms the safety of our therapeutic and control lentiviral vectors. We recently showed that in hu-BLT mice, monitoring clonal expansion of gene-modified and control populations within the same animal gives an unbiased analysis[8] and allows a more direct assessment of therapeutic vectors. In the current study, we longitudinally tracked both anti-HIV gene-modified and control vector-modified populations in the same hu-BLT mouse. Interestingly, we found a competitive growth pattern between two populations, with few clones from each population leading the expansion. The clonal profiles in both populations resemble typical after transplant clonal repopulation confirming safety of both vectors. Since safety and efficacy of multiple vectors can be tested in the same humanize mouse, LoVIS-Seq can reduce both cost and time of vector development.

Previous studies propose that in myeloablated mice, hematopoiesis tends to stabilize around 22 weeks post-transplant[7,31] while a recent study suggested 16 to 24 weeks[32]; our data indicates clonal stabilization between

17 to 19 weeks post-transplant. Overall, the clonal repopulation of human HSPC in hu-BLT mice resembles that of mouse HSPC after autologous transplant. Cord blood HSPC transplanted in NGS mice also showed similar clonal behavior, with clonal stabilization starting near week 18 to 20 post-transplant[33,34]. Although the timescales compare well with other studies, caution is due considering high incidence of graft versus host-disease-related illnesses in xenograft mouse models.

Analysis of repopulating VIS clones shows vector integration preference for transcriptionally active genes in FL-HSPC however, RNA-Seq data indicates bias against highly expressed genes. For in vitro activated cord blood and BM derived CD34+ HSC, the lentiviral vector showed no preference for highly active genes[23]. This is likely due to either obstruction by transcriptional machinery or detrimental effect of vector integration on survival of the cell. Investigation of 10 chromatin features showed strong VIS bias toward actively transcribed regions marked by histone modifications H3K36me3 and H3K79me2. This bias could be attributed to LEDGF/p75, a chromatin binding protein essential for efficient HIV-1 integration[35,36], that binds to integrases of HIV-1[37,38] and protects the pre-integration complex from degradation[39], whereas the N-terminal PWWP domain of LEDGF is known to preferentially interact with H3K36me3[40,41]. H3K79me2 and H3K36me3 mark the gene body[42] and H3K36me3 marks exons and is positioned near the 5` end of the exon and is correlated with alternative splicing[43,44]. Thus, VIS in proximity of H3K36me3 are likely to influence co-transcriptional splicing of the proximal gene as well as expression of the vector itself. In vivo repopulating clones having significant vector integration bias for H3K36me3 may be linked to survival of clones in vivo requiring further investigation. However, tracking of hundreds of single HSPC clones suggests that such transcription events have miniscule to no impact on the stemness of repopulating vector-modified HSPC.

A recent study demonstrated use of CRISPR/Cas to introduce barcodes in the long-term HSPC and longitudinally tracked a very limited number of HSPC clones[34]. Comparatively, using LoVIS-Seq we have tracked ~10 times more HSPC clones per animal with high accuracy and reproducibility. It is pertinent to note that to enable insertion of barcoded donor DNA into the host genome, HSPC need to undergo in vitro preconditioning and incubation before transplantation. The double stranded breaks introduced by CRISPR/Cas activate DNA damage responses causing significant delays in HSPC proliferation and affects their in vivo repopulation[45]. Additionally, off-target gene-editing by CRISPR/Cas remains a concern. In contrast, LoVIS-Seq does not require preconditioning of HSPC and provides a ready to use high-throughput clonal tracking assay for small-animal models. Furthermore, LoVIS-Seq has wider applicability owing to its adaptability to many lentiviral vectors commonly used to insert transgenes or reporter gene such as GFP.

LoVIS-Seq with whole genome amplification allows for quantitative assessment of clonal behavior in small-animal models. However, the accuracy and reproducibility of our assay depends on the initial number of cells used for MDA (Figure 1d). To minimize sampling errors, it is important to have a sufficient number of

gene-marked human cells in each 25µl of blood or 10,000 cells to represent each clone in similar proportions as in the bulk population. Higher human reconstitution and gene marking are often desirable and necessary conditions wherein our assay provides optimal results.

Overall, using a mere 25µl blood and LTR-indexed vectors, we explored polyclonal expansion in both control and therapeutic vector-modified populations in the same hu-BLT mice. LoVIS-Seq revealed the in vivo dynamics of clonal expansion, emergence of stable stem cell clones, and consequences of vector integration bias on repopulation. Thus, LoVIS-Seq provides an efficient tool for multifaceted analysis of clonal dynamics in murine and humanized-mouse models that are used extensively in HIV, cancer, gene-therapy, and stem cell research.

Methods:

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

Human fetal thymus and livers were obtained from Advanced Bioscience Resources (ABR) and the UCLA CFAR Gene and Cellular Therapy Core. Human fetal liver CD34+ HSPC and thymus pieces were processed as previously described[46]. Briefly, a single cell suspension of fetal liver cells was strained through 70 µm mesh and layered onto density gradient separation media (Ficol Paque PLUS, GE Healthcare). After 20 minutes of centrifugation, the mononuclear cells layer was collected. Anti-CD34+ microbeads (Miltenyi Biotech) were used for magnetic isolation of CD34+ cells from mononuclear cells. Calvanese et al.[30] also obtained fetal liver from the UCLA CFAR Gene and Cellular Therapy Core and followed identical CD34+ magnetic sorting to isolate uncultured FL-HSPC for RNA-seq, ATAC-seq, and ChIP-seq assay.

Humanized BLT mouse and sample collection

NOD.Cg-*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/SzJ (NSG) mice, 6-8-week-old, were myeloablated 1 day before transplant by intraperitoneal (i.p.) injection with 10 mg/kg of 6-thioguanine (6TG) (Sigma-Aldrich, Saint Louis, MO) or 35mg/kg of Busulfan for mouse m860. Myelo-preconditioned mice were transplanted with human fetal liver CD34+ HSPCs transduced with Anti-HIV vector (EGFP+) (0.5×10^6 cells/mouse) and mixed with HSPCs transduced with the control (mCherry+) vector (0.5×10^6 cells/mouse). The mice were transplanted with a two-step procedure: half the mixture of EGFP+ and mCherry+ transduced cells was solidified by matrigel (BD Bioscience, San Jose, CA), mixed with CD34- cells as feeder cells ($4.5x10^6$ cells), and implanted with a piece of human thymus under the mouse kidney capsule. Then, mice were injected with the other half of the mixed EGFP+ and mCherry+ transduced cells via retro-orbital vein plexus using a 27-gauge needle on the same day. Bone marrow cells for MDA were harvested from mouse m860 at week 25 post-transplant. For longitudinal clonal tracking and to monitor human leukocyte reconstitution and percentage of the EGFP+ and mCherry+ marked cells, 100µl of mouse blood was collected from the retro-orbital vein every two weeks from weeks 13-19 post-transplant. Plasma was removed and peripheral blood cells were stained with

monoclonal antibodies for 30 minutes. Red blood cells were lysed with red blood cell lysis buffer (4.15 g of NH4Cl, 0.5 g of KHCO3, and 0.019 g of ethylenediaminetetraacetic acid in 500 mL of H2O) for 10 minutes and washed with FACS buffer (2% fetal calf serum in phosphate-buffered saline [PBS]). Stained cells were resuspended in 16µl PBS, of which 8µl was split equally into two tubes for MDA replicates. The remaining 8µl was mixed with 300µl of 1% formaldehyde in PBS and examined with Fortessa (BD Biosciences) flow cytometers. Flow cytometry data was utilized to monitor human reconstitution (Supplementary table 2) and count human cells, mCherry+ cells, and EGFP+ cells as well as human T and B cells in blood (Supplementary figure 2). The following monoclonal antibodies with fluorochromes were used: human CD45-eFluor 450 (HI30, eBioscience), CD3-APC-H7 (SK7: BD Pharmingen), and CD19-BV605 (HIB19: BioLegend). Data were analyzed on FlowJo (TreeStar, Ashland, OR) software.

The UCLA Institutional Review Board has determined that fetal tissues from diseased fetuses obtained without patient identification information are not human subjects. Written informed consent was obtained from patients for use of these tissues for research purposes. All mice were maintained at the UCLA Center for AIDS Research (CFAR) Humanized Mouse Core Laboratory in accordance with a protocol approved by the UCLA Animal Research Committee.

LoVIS-Seq workflow with whole genome amplification and quantitative VIS assay

Multiple displacement amplification for whole genome amplification: To estimate the minimum number of cells required for LoVIS-Seq, we collected 81,000, 27,000, 18,000, 9,000, 3,000, and 1,000 bone marrow cells from mouse m860 by serial dilution and stored in 4µl of PBS at -20ºC. MDA was done directly on cells using the REPLI-g Single Cell Kit from Qiagen (Cat #150343) following kit-specific protocol. For longitudinal clonal tracking in the blood compartment, 100 µl blood was drawn at weeks 13, 15 & 17. At end point (week 19), max blood (≈ 1ml) was collected, out of which 100µl was used for flow cytometric analysis along with MDA; the remainder was used to isolate unamplified whole blood DNA. Cells for MDA were isolated as described above and stored in 4µl of PBS at -20ºC. MDA-amplified DNA was then used for quantitative VIS assay. A Qiagen DNeasy Blood & Tissue Kit was used to extract unamplified DNA from max blood cells, splenocytes, and bone marrow cells.

Quantitative VIS assay and data analysis workflow: For VIS sequencing, we followed the procedures described in our previous publication[5,8,47,48] and focused on analyzing only the right LTR junctions using CviQI and RsaI restriction enzymes. For our VIS assay we used one microgram MDA-amplified or unamplified genomic DNA for animal m860 samples and two micrograms MDA-amplified or unamplified genomic DNA for different time point samples, with a few exceptions (see Supplementary Tables 1 and 2). DNA samples were subject to extension PCR using LTR specific biotinylated primers

/5BiotinTEG/CTGGCTAACTAGGGAACCCACT 3' and /5BiotinTEG/CAGATCTGAGCCTGGGAGCTC 3'. The extension PCR product was then digested using CviQI and RsaI restriction enzymes and biotin primer bound DNAs isolated using streptavidin-agarose Dynabeads using magnetic separator as per manufactures instructions. The vector-host junctions capture on streptavidin beads were processed for linker-mediated PCR (LM-PCR) methods as described previously[47,48]. The linker ligated vector-host junction DNA was subjected to two step PCR. First step amplification was done using primer 5' CTGGCTAACTAGGGAACCCACT 3' and first linker primer GTGTCACACCTGGAGATAT. We removed the internal vector sequence by restriction enzyme (SfoI) digestion. The digested product of first PCR was then amplified using primer 5'ACTCTGGTAACTAGAGATCC 3' and second linker primer 5' GGAGATATGATGCGGGATC 3'. Since the LTR index sequence is included in the vector-host junction the we obtain unbiased amplification all the H1, H5 and/or WT VIS sequences. Lentiviral vectors used in this study as derived from FG12-mCherry lentiviral vector[46] and all the primers are designed accordingly. A detailed protocol for VIS assay is provided in supplementary text. The amplicon libraries prepared using custom made Illumina sequencing primers for Illumina MiSeq (m860 samples) or iSeq100 (m599, m599, and m591 samples) sequencer. Sequences with a virus-host junction with the 3' end LTR, including both the 3'-end U5 LTR DNA and ≥ 25 base host DNA (with ≥ 95% homology to the human genome), were considered true VIS read-outs. The sequence mapping and counting method was performed as described previously[8]. In brief, sequences that matched the 3'end LTR sequence joined to genomic DNA as well as LTR-indexes (H1, H5 or WT) were identified using a modified version of SSW library in C++[49]. Reads were classified as H1, H5, or WT VIS based on the LTR barcodes used in the experiment. VIS sequences were mapped onto the human genome (Version hg38 downloaded from https://genome.ucsc.edu/) using Burrows-Wheeler Aligner (BWA) software. Mapped genomic regions were then used as reference and VIS reads were remapped using BLAST to further remove poorly mapped reads to get an accurate estimate of sequence count. Final VIS counting was done after correcting for VIS collision events and signal crossover as described previously. VIS with a final sequence count less than the total number of samples analyzed per animal were removed. VIS clones with maximum frequency values below 1st quartile were classified as "low frequency", clones with maximum frequency value above 3rd quartile were classified as "high frequency", and clones with maximum frequency between the 1st and 3rd quartiles were designated "medium frequency". VIS clones that were detected with frequency >0 at every week from 13-19 are termed "persistent clones". The 10 high frequency VIS clones at each timepoint were selected as top 10 VIS. All the VIS data and list of VIS-proximal genes is provided in supplementary file.

Random integration sites

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

Clonal diversity analysis

For diversity analysis, we used Rényi's diversity/entropy[26] of order α defined as follows

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

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

RNA-seq data analysis

Raw sequence data of uncultured FL-HSPC (in triplicate) was pre-processed for quality using Fastqc. Trimmomatic was used to remove adaptors and for quality trimming. After this, reads were aligned onto human genome hg38 using RNA STAR aligner[50]. SAMtools was used to remove reads with low mapping scores (< 20) and to generate BAM files. Cufflinks[51] was used to calculate FPKM values for all genes. The human cancer consensus gene list is from Catalogue of Somatic Mutations In Cancer (https://cancer.sanger.ac.uk/census).

ATAC-seq analysis

Raw sequence data of uncultured FL-HSPC (in triplicate) was pre-processed for quality using Fastqc. Adaptor removal and quality trimming was done using Trimmomatic. After this, reads were mapped onto human genome hg38 using bowtie2 with parameter --very-sensitive -X 2000 -k 1. SAMtools was used to

remove reads with low mapping (<20) scores, blacklisted regions[52], and to generate BAM files. Picard tool kit was used to remove duplicate reads. We used Genrich, a paired end peak caller, to identify ATAC peaks. Software deepTools[53] was used to generate coverage (.bw) files and for visualization of open DNA in genes and VIS-proximal regions.

ChIP-seq data analysis

Raw sequence data of uncultured FL-HSPC for histones, RNAploII, and input were pre-processed for quality using Fastqc. Trimmomatic was used to remove adaptors and for quality trimming. After this, reads were mapped onto human genome hg38 using bowtie2 with parameter --local. SAMtools was used to remove reads with low mapping (<20) scores, blacklisted regions[52], and to convert SAM to BAM format. Picard tool kit was used to remove duplicates. MACS2 tool was used to call peaks for all histone marks and RNApolII using input sample as control. Software deepTools[53] was used to generate coverage .bw files and for visualization of histone/RNApolII in genes and VIS proximal regions.

Statistical analysis

Clonal frequencies are summarized as means ± standard deviations (SDs). Pearson correlations are used to compare the similarity and reproducibility of clonal profiles between two samples and replicates, respectively. Pearson's r values and p values are calculated using statistical software R (version 3.6, https://www.r-project.org/). To determine if VIS preference for genomic and epigenetic features differs significantly from random IS, we used Pearson's chi-squared test with Yate's continuity correction (function chisq.test() in software R). We use Principle component analysis (PCA) to reduce the complexity of read coverage data of multiple chromatin feature in proximity to VIS. The dimensionality reduction by PCA method is similar to clustering and allows detection of patterns in the data. In this study, PCA was done using software deepTools[53].

Data availability.

Raw RNA-seq, ATAC-seq, ChIP-seq data of uncultured FL-HSPC from published reference is available in Gene Expression Omnibus (GEO) with the accession code GSE111484[30].

Acknowledgements: We thank the technical support from the UCLA Center for AIDS research (CFAR) 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 He, F., Zhou, W., Cai, R., Yan, T. & Xu, X. Systematic assessment of the performance of whole-genome amplification for SNP/CNV detection and beta-thalassemia genotyping. *J Hum Genet* **63**, 407-416, doi:10.1038/s10038-018-0411-5 (2018).

12 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).

13 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).

14 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).

15 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).

16 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).

17 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).

18 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).

19 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).

20 Lewinski, M. K. *et al.* Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog* **2**, e60, doi:10.1371/journal.ppat.0020060 (2006).

21 Lelek, M. *et al.* Chromatin organization at the nuclear pore favours HIV replication. *Nat Commun* **6**, doi:ARTN 648310.1038/ncomms7483 (2015).

22 Marini, B. *et al.* Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**, 227-231, doi:10.1038/nature14226 (2015).

23 Cattoglio, C. *et al.* Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* **110**, 1770-1778, doi:10.1182/blood-2007-01-068759 (2007).

24 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).

25 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).

26 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).

27 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).

28 Simpson, E. H. Measurement of Diversity. *Nature* **163**, 688-688, doi:DOI 10.1038/163688a0 (1949).

29 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).

30 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).

31 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).

32 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).

33 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).

34 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).

35 Llano, M. *et al.* An essential role for LEDGF/p75 in HIV integration. *Science* **314**, 461-464, doi:10.1126/science.1132319 (2006).

36 Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11**, 1287-1289, doi:10.1038/nm1329 (2005).

37 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).

38 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).

39 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).

40 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).

41 Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R. & Bickmore, W. A. Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet* **8**, e1002717, doi:10.1371/journal.pgen.1002717 (2012).

42 Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol* **17**, 1495-1499, doi:10.1038/nsmb.1924 (2010).

43 Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5**, e1000566, doi:10.1371/journal.pcbi.1000566 (2009).

44 Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000, doi:10.1126/science.1184208 (2010).

45 Schiroli, G. *et al.* Precise Gene Editing Preserves Hematopoietic Stem Cell Function following Transient p53-Mediated DNA Damage Response. *Cell Stem Cell* **24**, 551-565 e558, doi:10.1016/j.stem.2019.02.019 (2019).

46 Shimizu, S. *et al.* A highly efficient short hairpin RNA potently down-regulates CCR5 expression in systemic lymphoid organs in the hu-BLT mouse model. *Blood* **115**, 1534-1544, doi:10.1182/blood-2009-04-215855 (2010).

47 Kim, S. *et al.* High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones. *J Virol* **84**, 11771-11780, doi:10.1128/JVI.01355-10 (2010).

48 Suryawanshi, G. W. *et al.* Bidirectional Retroviral Integration Site PCR Methodology and Quantitative Data Analysis Workflow. *J Vis Exp*, doi:10.3791/55812 (2017).

49 Zhao, M., Lee, W. P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One* **8**, e82138, doi:10.1371/journal.pone.0082138 (2013).

50 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

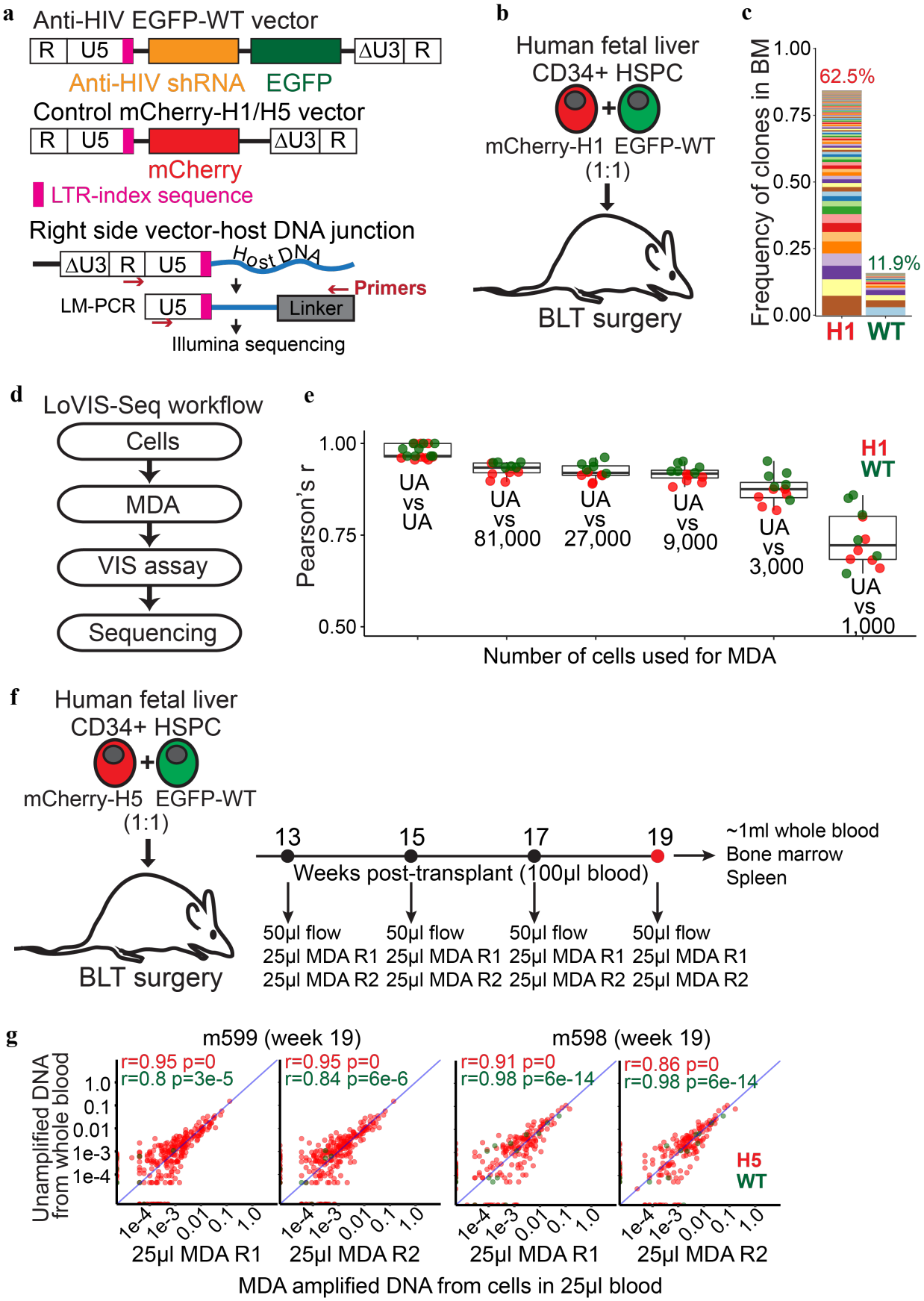
51 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

52 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354, doi:10.1038/s41598-019-45839-z (2019).

53 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

621 **Figures-**

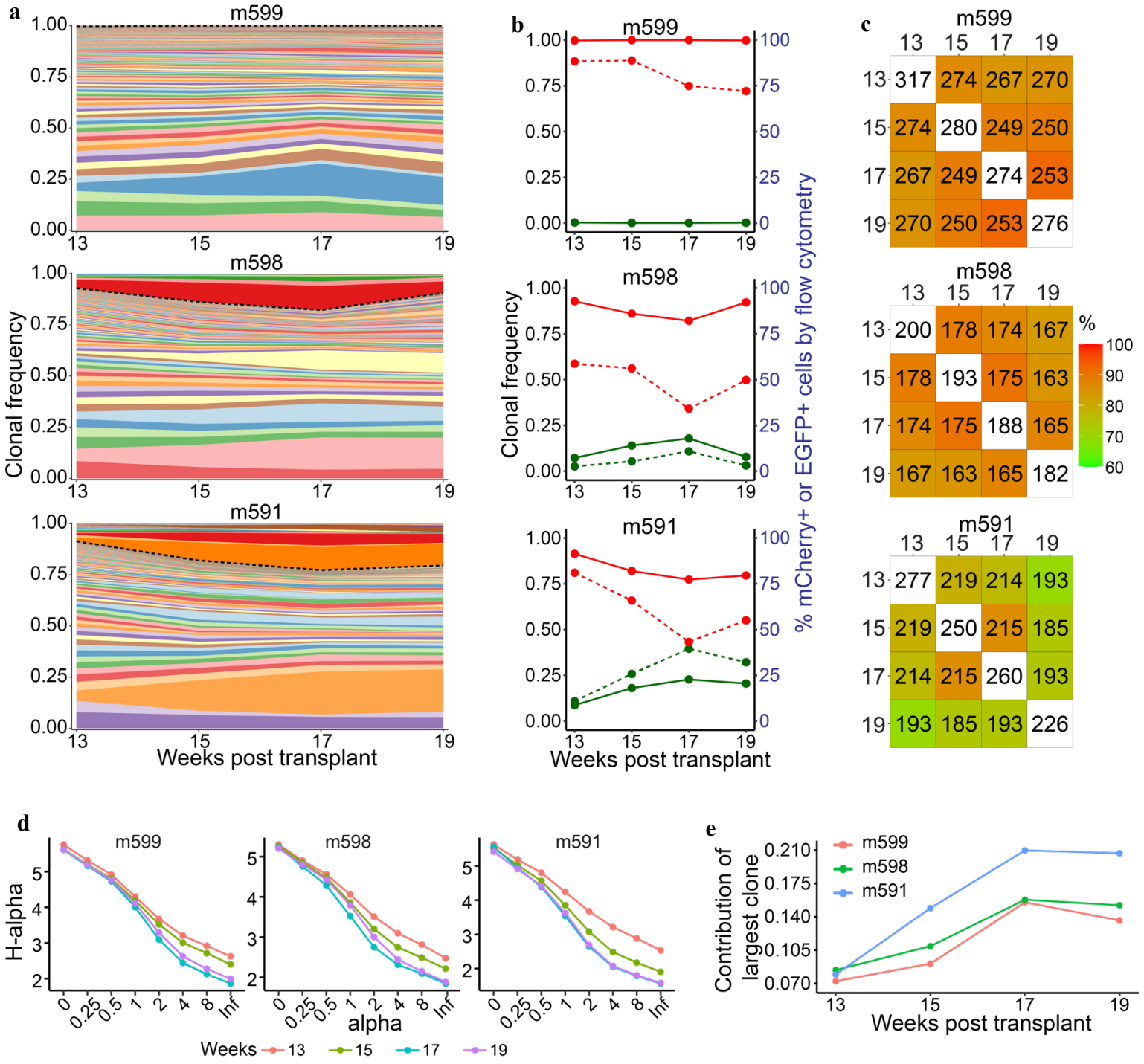
622 Figure 1



623

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

644 Figure 2

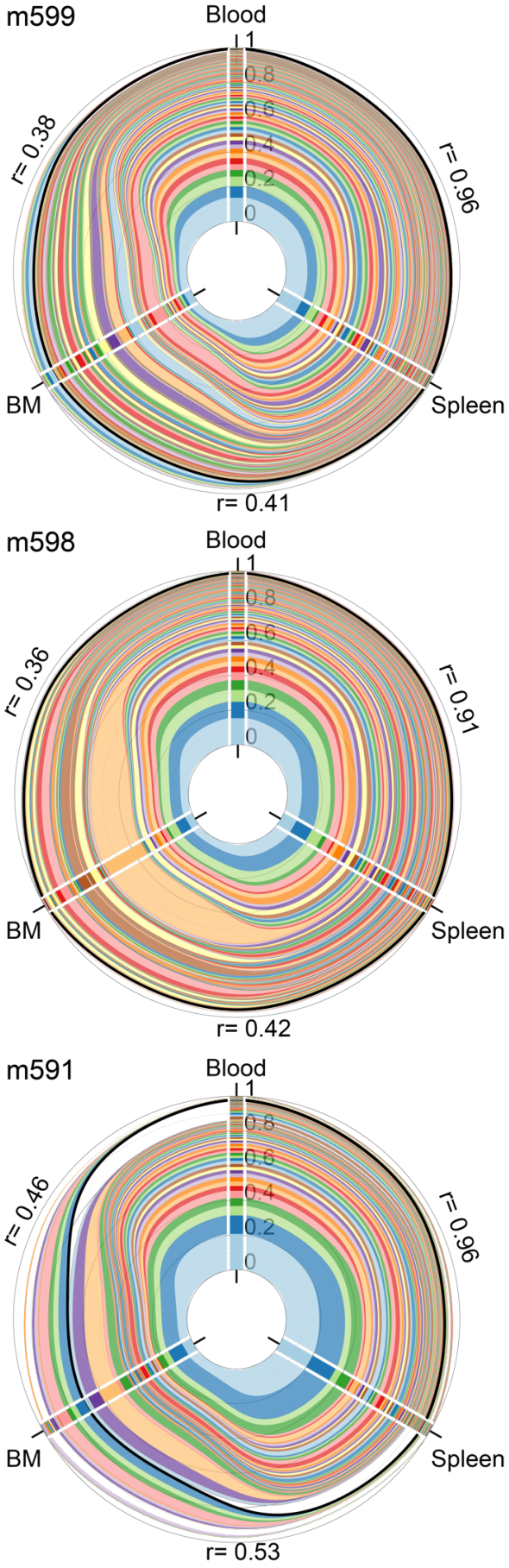


645

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

652 percentage change in shared clones between two timepoints. Digits inside white tiles on the diagonal show
653 number of VIS detected at each time point. Colors of each heatmap tile correspond to percentage of clones
654 shared and color key is provided on the right. Digits in each tile show the number of VIS shared between two
655 timepoints. **d**) Renyi's diversity profiles evaluated using raw count data from two replicates at each
656 timepoint and by varying value of alpha. Renyi's diversity profiles are arranged with highest diversity at the
657 top to lowest at the bottom. Topmost curve with no overlap or intersection with any other curve has the
658 highest overall diversity. Diversity of curves that overlap or intersect is undefined. **e**) Line plot showing
659 contribution of highest contributing clone at different timepoints. Values reported are exp(H(alpha)) at
660 alpha=∞.
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685

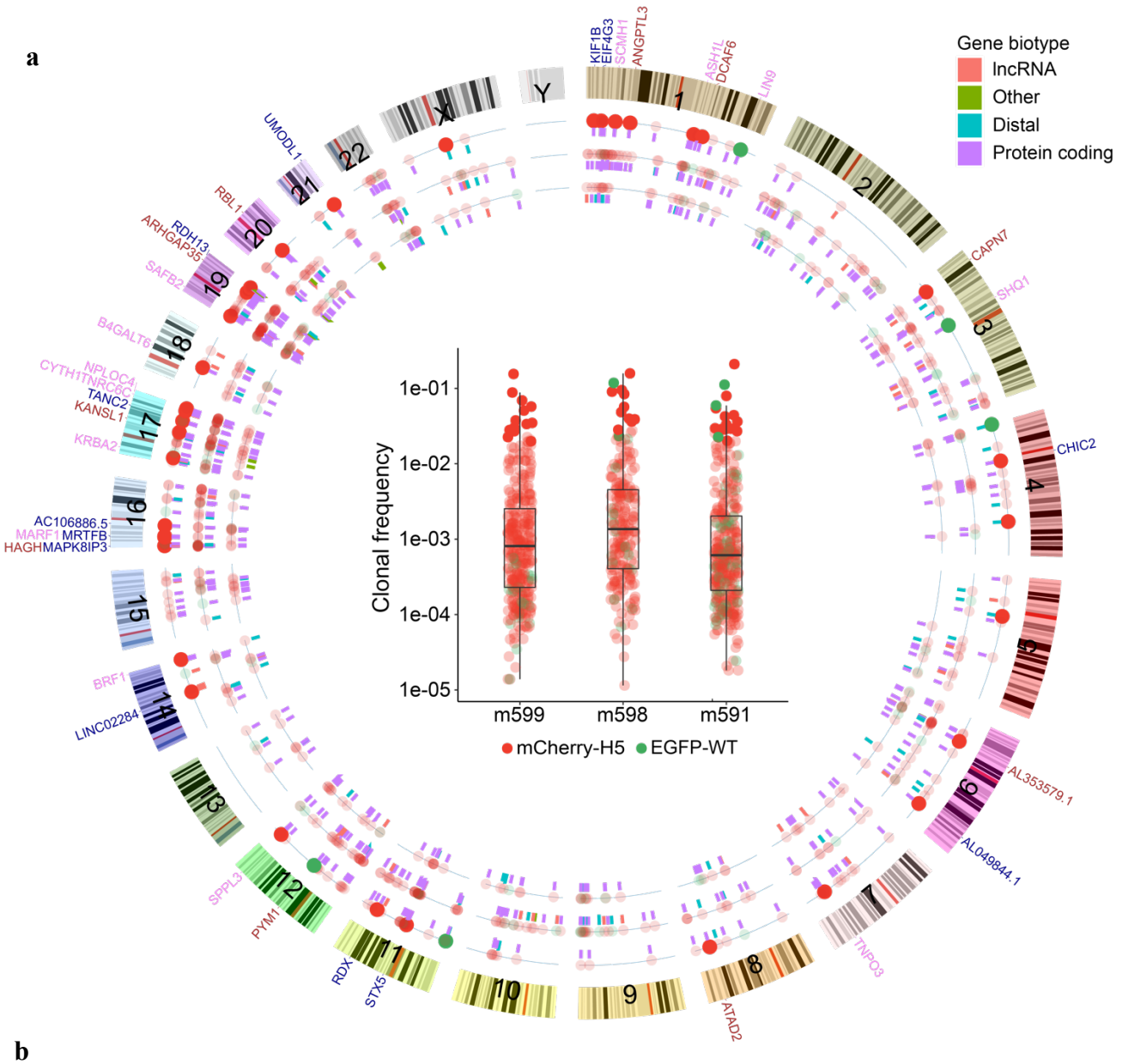
686 Figure 3



687

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.

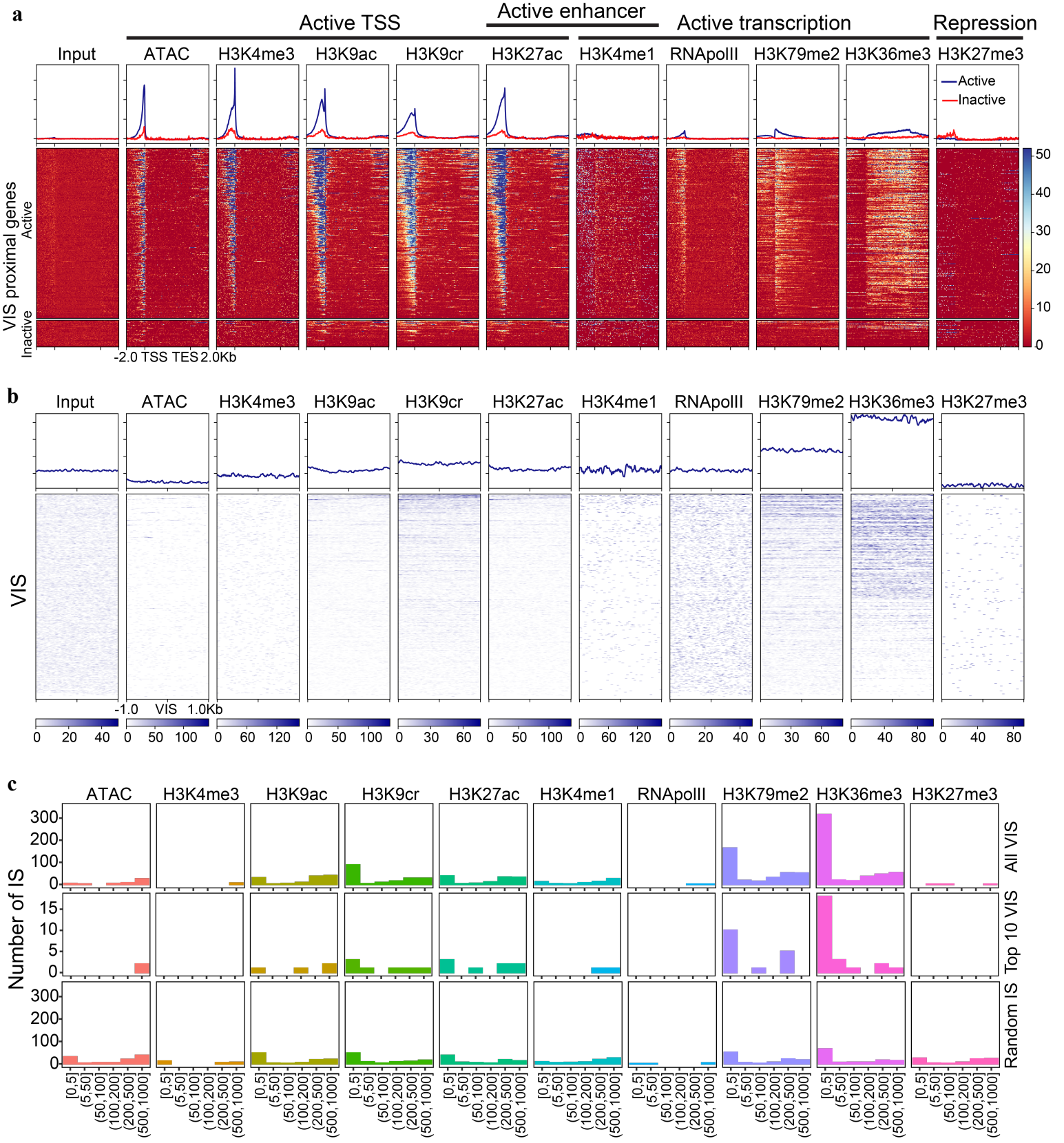
722 Figure 4



723

724 **Figure 4: Chromosomal distribution of VIS and its bias for transcriptionally active genes: a**) Circos
725 plot shows genomic location of all 792 mCherry-H5 (red dots) and 105 EGFP-WT (green dots) VIS from
726 mice m599, m598, and m591. Box plots in the center show maximum frequency of mCherry-H5 (red dots)
727 and EGFP-WT (green dots) VIS clones in mice m599, m598, and m591 over 6 weeks. Genomic location of
728 mCherry-H5 (red dots) and EGFP-WT (green dots) VIS clones are plotted on three concentric circles
729 depending on their maximum frequency over 6 weeks: Low frequency clones with maximum frequency
730 below the 1st quartile value (innermost circle), High frequency clones above 3rd quartile value (outermost
731 circle), and Medium frequency clones between 1st and 3rd quartile (middle circle). Top 10 high frequency
732 VIS clones are shown in darker colors. Functional classification of VIS-proximal genes is shown by short
733 line segments, color coded as in the legend. Gene symbols above ideograms represent genes proximal to the
734 top 10 VIS clones from mice m599 (blue), m598 (brown), and m591 (light pink). **b**) Classification of all,
735 persistent, and Top 10 VIS clones based on biotype of proximal gene. Inner pie chart shows clones classified
736 based on gene biotype of the most proximal gene. Outer Donut plots show number of VIS and numbers in
737 bracket show % of VIS proximal to active (dark color) or inactive (faded colors) genes. Active proximal
738 genes have FPKM >1.
739

740 Figure 5



741

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

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

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