Figure S1

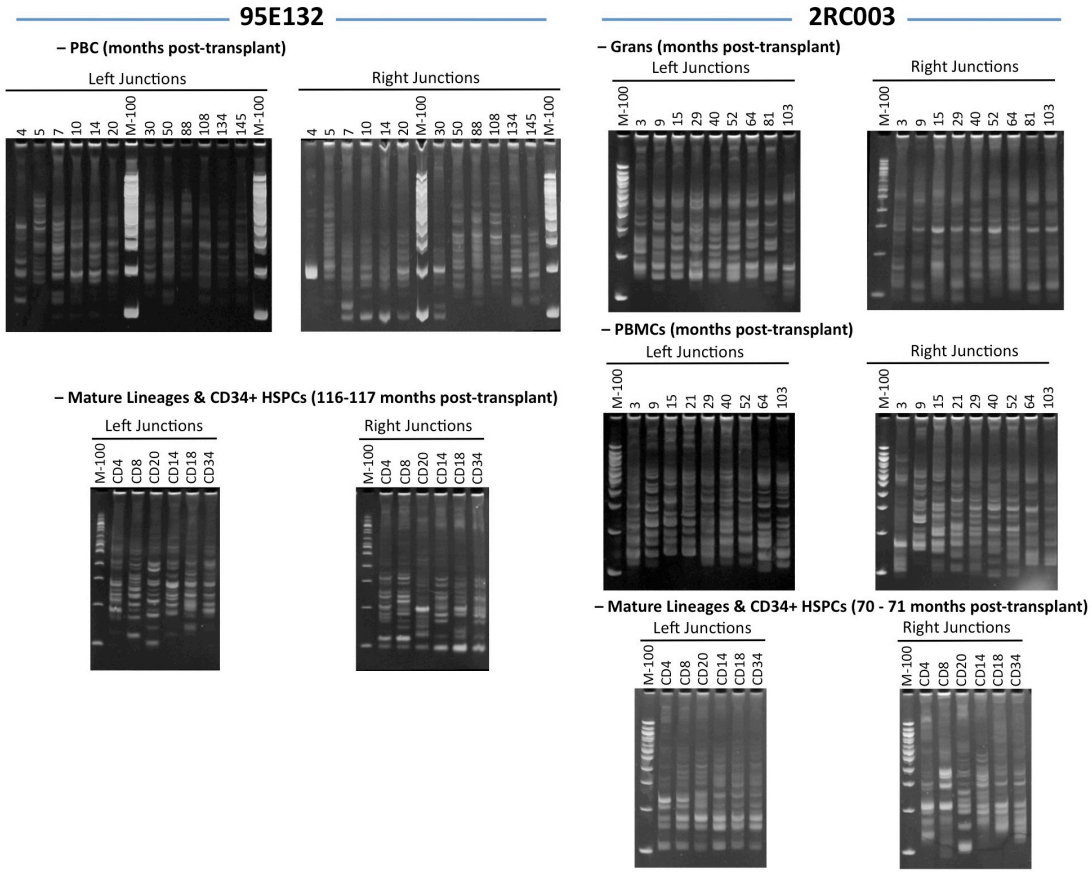

95E132
– PBC (months post-transplant)
Left Junctions
Right Junctions
4 5 7 10 14 20 M-100 30 50 88 108 134 145 M-100
– Mature Lineages & CD34+ HSPCs (116-117 months post-transplant)
Left Junctions
Right Junctions
M-100 CD4 CD8 CD20 CD14 CD18 CD34
2RC003
– Grans (months post-transplant)
Left Junctions
Right Junctions
M-100 3 9 15 29 40 52 64 81 103
M-100 3 9 15 29 40 52 64 81 103
– PBMCs (months post-transplant)
Left Junctions
Right Junctions
M-100 3 9 15 21 29 40 52 64 103
M-100 3 9 15 21 29 40 52 64 103
– Mature Lineages & CD34+ HSPCs (70 - 71 months post-transplant)
Left Junctions
Right Junctions
M-100 CD4 CD8 CD20 CD14 CD18 CD34
M-100 CD4 CD8 CD20 CD14 CD18 CD34
RQ5427
– Grans (months post-transplant)
Left Junctions
Right Junctions
M-100 2 8 19 25 32 43 50 56 67
M-100 2 8 19 25 32 43 50 56 67
– PBMCs (months post-transplant)
Left Junctions
Right Junctions
M-100 2 8 13 19 25 32 43 50 56 67
M-100 2 8 13 19 25 32 43 50 56 67
– Mature Lineages & CD34+ HSPCs (64 - 65 months post-transplant)
Left Junctions
Right Junctions
CD4 CD8 CD20 CD14 CD18 CD34 M-100
CD4 CD8 CD20 CD14 CD18 CD34 M-100
RQ3570
– Grans (months post-trar
Left Junctions
Right Junctions
2 8 13 19 32 38 M-100
M-100 2 8 13 19 32 38
– PBMCs (months)
Left Junctions
Right Junctions
2 8 13 19 25 32 38 M-100
M-100 2 8 13 19 25 32 38
– Mature Lineages & CD34+ cells (36-37 months)
Left Junctions
Right Junctions
M-100 CD4 CD8 CD20 CD14 CD18 CD34
M-100 CD4 CD8 CD20 CD14 CD18 CD34

Figure S1. Gel Electrophoresis of Vector Integration Site PCR amplicons (related to Figure 2).
Acrylamide gel electrophoresis shows various lengths of vector insertion site (VIS) PCR amplicons. Both the left and right DNA junctions of vector integrants in repopulating cell samples from animals 95E132, 2RC003, RQ5427, and RQ3570 were amplified by the bi-directional vector integration site assay. Total peripheral blood cells (PBCs), peripheral blood mononuclear cells (PBMCs), and granulocytes (Grans) – collected at indicated time points (months) –and mature lineages (CD4, CD8, CD20, CD14, and CD18) and CD34+ HSPC – collected at the experimental end-point – were analyzed. Arrows indicate PCR amplicons of internal vector DNA. 100-bp markers (M-100, New England Biolab) were used as size markers.

Figure S2

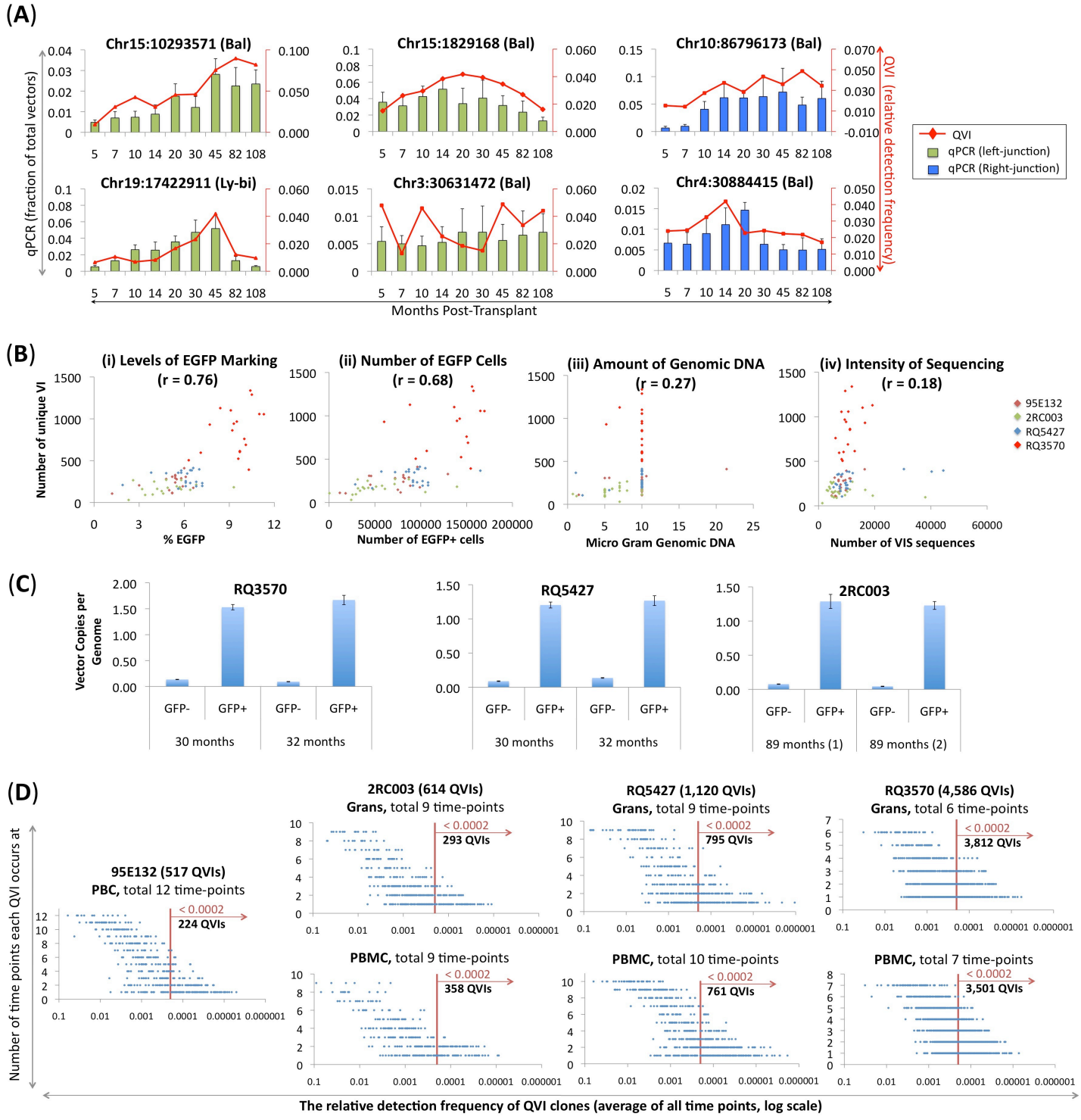


Figure S2. Lentiviral Vector-Mediated Clonal Tracking Assay (related to Figure 2).

(**A**) These charts compare VIS sequencing-based clonal quantification and clone-specific real-time PCR. The relative frequencies of six QVIs from animal 95E132 (red line, right y-axis) were compared with clone-specific real-time PCR (left y-axis) targeting either the left (green bars) or right (blue bars) vector-host DNA junctions. Error bars refer to standard errors that occurred in the three independent experiments. All QVI clones were Bal type, except Chr19:17422911 (Ly-bi). (**B**) Data show that the sample-to-sample variations in the total number

of unique vector integrants among samples were highly associated with levels of EGFP marking [Pearson's correlation r=0.68, see (i)] and the number of EGFP cells [r=0.76, (ii)] in each sample, and less associated with the amount of genomic DNA [r=0.27, (iii)] and the intensity of sequencing [r=0.18, (iv)]. Data from animals 95E132, 2RC003, RQ5427, and RQ3570 were shown with brown, green, blue, and red dots, respectively, in figures (i)–(iv). While such association was unclear when tested for each animal separately, a high correlation value was observed for the % EGFP+ cell and the EGFP+ cell number data sets when all animals were considered. We also showed that the total number of clones in each animal was proportionally correlated with the total number of EGFP+ CD34+ cells infused in each animal at the beginning of the experiments (see Fig.2A). These results indicate that sample-to-sample variations in our VIS data are due more to biological variability than technical variability because samples were analyzed under optimized assay conditions, including the amount of genomic DNA, the intensity of sequencing, and analysis standards to restrict detection biases. (**C**) The number of vector copies per cell was calculated based on quantitative real-time PCR using GFP+ sorted cells (see suppl.method). For animal 2RC003, two sets of GFP+ and GFP- CD4+ cells were isolated one week apart at approximately 89 months post-transplant [the two sets are denoted as (1) and (2)]. For animals RQ5427 and RQ3570, two sets of GFP+ and GFP- CD4+ cells were isolated at 30 and 32 months post-transplant. An average of 1.26, 1.23, and 1.60 vector copies per cell (i.e., per genome) was detected for animals 2RC003, RQ5427, and RQ3570 respectively. Given that approximately 40.3% of integrated vectors are QVI, the chance that each QVI represents a unique clone was 82.5% for animal 2RC003, 84.3% for animal RQ5427, and 66.7% for animal RQ3570. This range was estimated by modeling the source distribution of vectors per clone as Poisson (drawing a 1000 samples from distributions with means of non-zero copy numbers: 1.60, 1.26, and 1.23), sampling 40.3% of the vectors as QVIs, estimating the fraction of QVIs that represent a unique clone in this sample, and then repeating this process 5000 times (see Suppl. Methods for more details). The number of vector copies per cell was not tested using GFP+ sorted cells in animal 95E132. The total vector copy number in unsorted peripheral blood cell samples in this animal was approximately 1.8 times higher than the EGFP+ cell percentage (data not shown). (**D**) The average detection frequencies of individual QVI clones (x-axis: average of all time-points in the serial blood analysis) and the number of time points at which these clones were detected (y-axis) are shown. Clones ≥0.0002 frequency often occur at more than 3-4 time points, thereby revealing a traceable trajectory of their longitudinal behaviors, while the clones <0.0002 frequency – comprising 43-83% of all QVIs – were undetectable or detected mostly at a single time point and are thus uninformative for kinetics analysis. As the detection frequencies of low-frequency clones are expected to be less reliable than higher-frequency clones – due to the stochastic process of PCR amplification during the assay processes (Peccoud and Jacob, 1996)–, a frequency cutoff of 0.0002 (vertical brown lines) was imposed for cluster analysis; without this cutoff, the results of the analysis would have been unreliable due to the presence of a large number of uninformative, low-frequency clones.

Figure S3

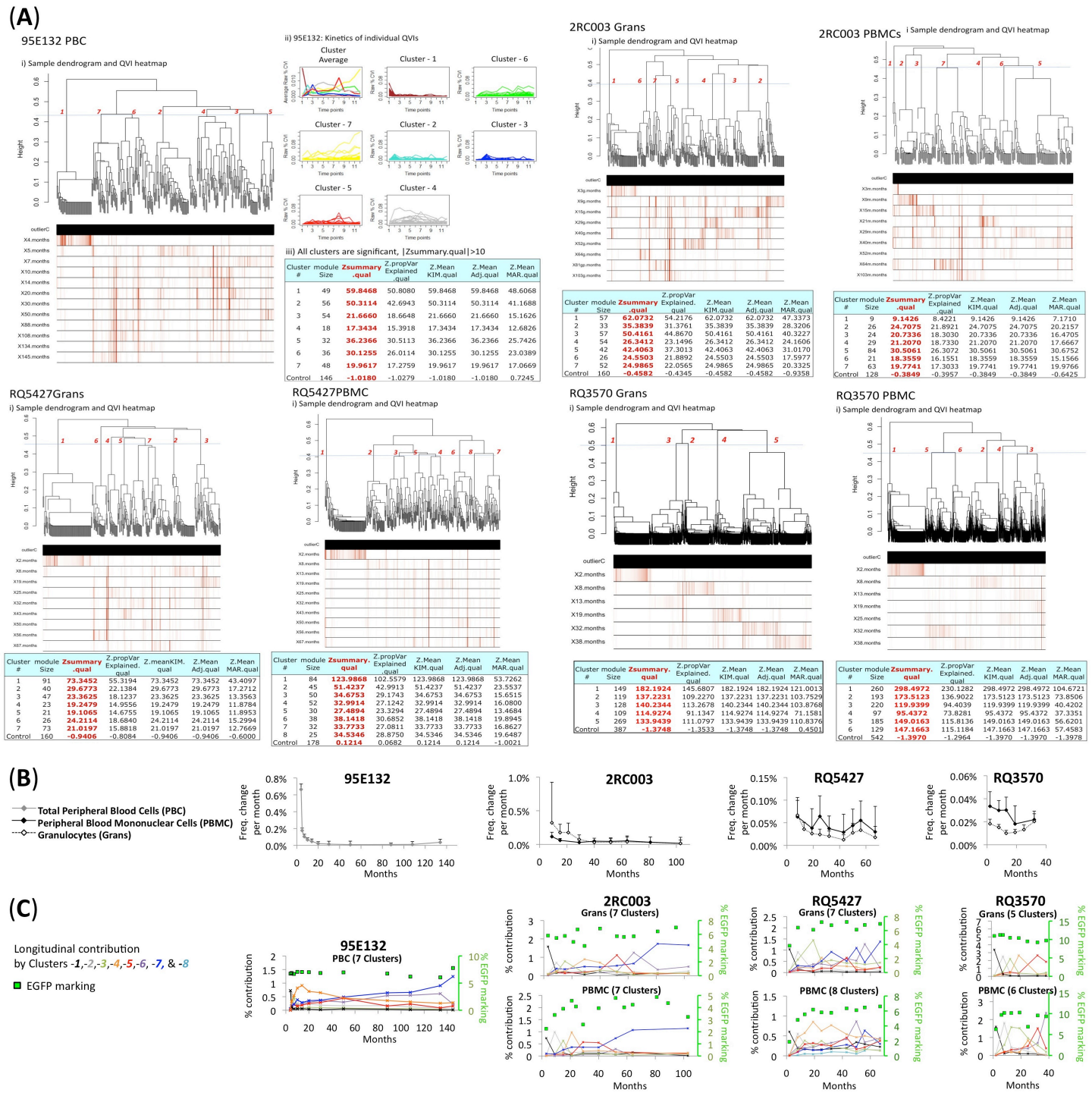


Figure S3. Analysis of Major Clonal Kinetics Patterns (related to Figures 3).

(**A**) Major clonal kinetics patterns for the peripheral blood cells (PBC) in animal 95E132, and for peripheral blood mononuclear cells (PBMCs) and granulocytes (Grans) in animals 2RC003, RQ5427, and RQ3570, were derived using average linkage hierarchical clustering where one minus the correlation matrix was used as the distance measure between clones using the WGCNA R package v.1.20[1]. **i)** Clusters were defined using a threshold line specific to each data set, where each threshold was selected to yield clusters with visually distinct kinetics profiles. The relative frequencies for individual QVIs are shown in a red-white color scheme. The

following static cut heights (h) were used to generate clusters – 95E132, h = 0.43, 7 clusters, 2RC003 Grans, h = 0.46, 7 clusters, 2RC003 PBMC, h = 0.46, 7 clusters, RQ5427 Grans, h = 0.45, 7 clusters and RQ5427 PBMC, h = 0.41, 8 clusters, RQ3570 Grans, h = 0.5, 5 clusters, and RQ3570 PBMC, h = 0.45, 6 clusters – where these heights were found to yield clusters with visually distinct kinetics profiles. Cluster numbers are indicated next to the corresponding branch. The significance of these clusters was quantified using the modulePreservation function in the WGCNA package[2] using parameters including networkType = "signed", dataIsExpr = FALSE and default parameters nPermutations = 200, randomSeed = 1, quickCor = 0. Cluster or "module" significance was evaluated using |Zsummary.qual| > 10. (**B**) Kinetically more stable clones emerged after the initial phase of clonal instability during the first 2-5 months post-transplant. The average changes in frequency per month by individual QVIs (y-axis) are plotted over time (x-axis). The rates of change were calculated based on the frequency difference between identical QVIs at two adjacent time points. Error bars indicate standard errors. **(C)** Sequential and balanced clonal repopulation maintained relatively constant EGFP marking in peripheral blood. The relative contributions by each cluster to the EGFP marking (left y-axis) are shown over time. The clusters and color codes for each cluster are the same as those in Fig. 3A. Different QVI clones contributed sequentially, starting with cluster-1 (black), followed by clusters -2 (grey), -3 (light green), -4 (orange), -5 (red), -6 (purple), -7 (blue), and -8 (light blue). The levels of EGFP markings (green squares) and trend lines (green lines), both on the right y-axis, for granulocytes were based on the granulocyte EGFP data closest to the sample collection time point (Table S1). EGFP markings for PBMCs were based on the average EGFP marking of lymphocytes and monocytes.

Figure S4.

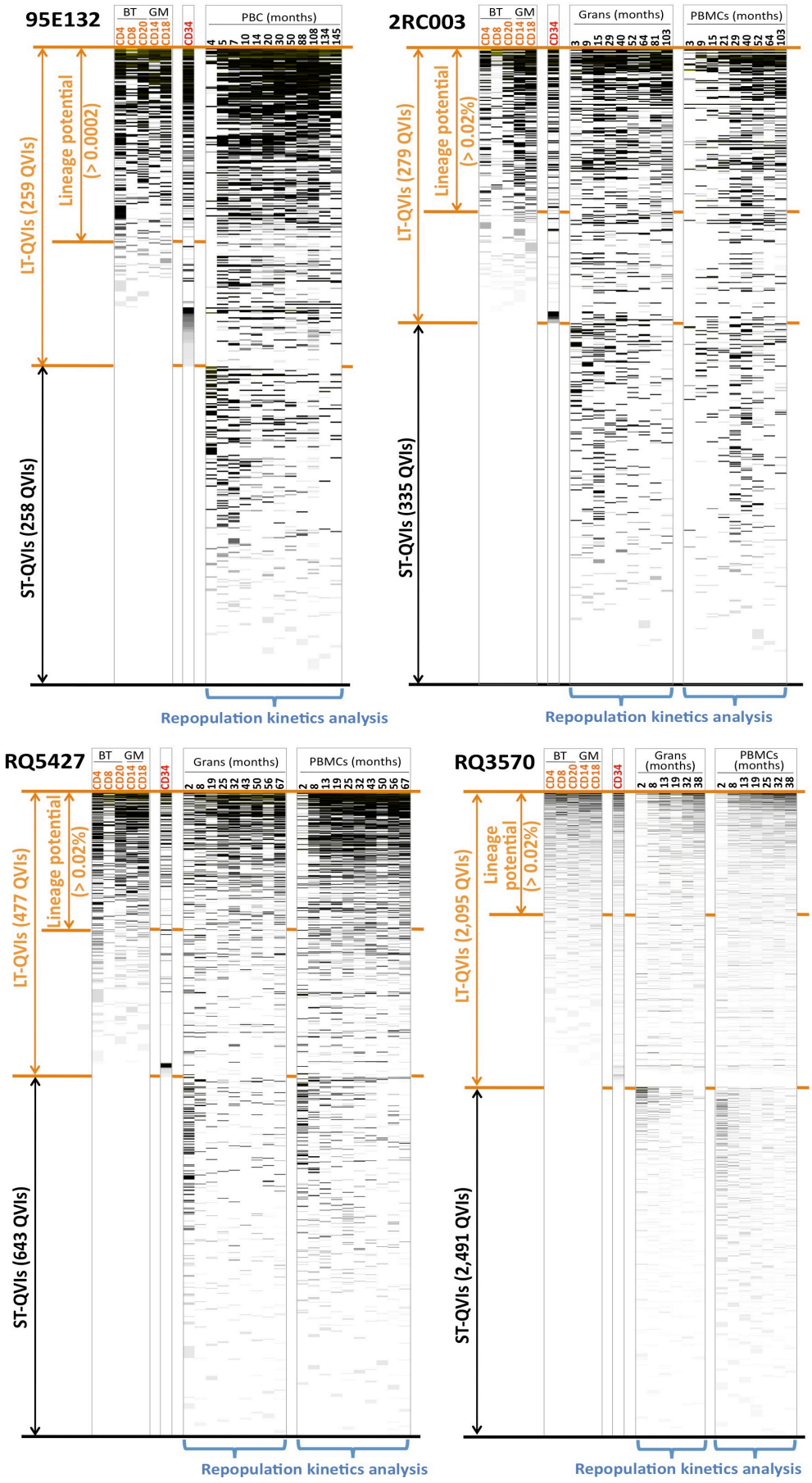
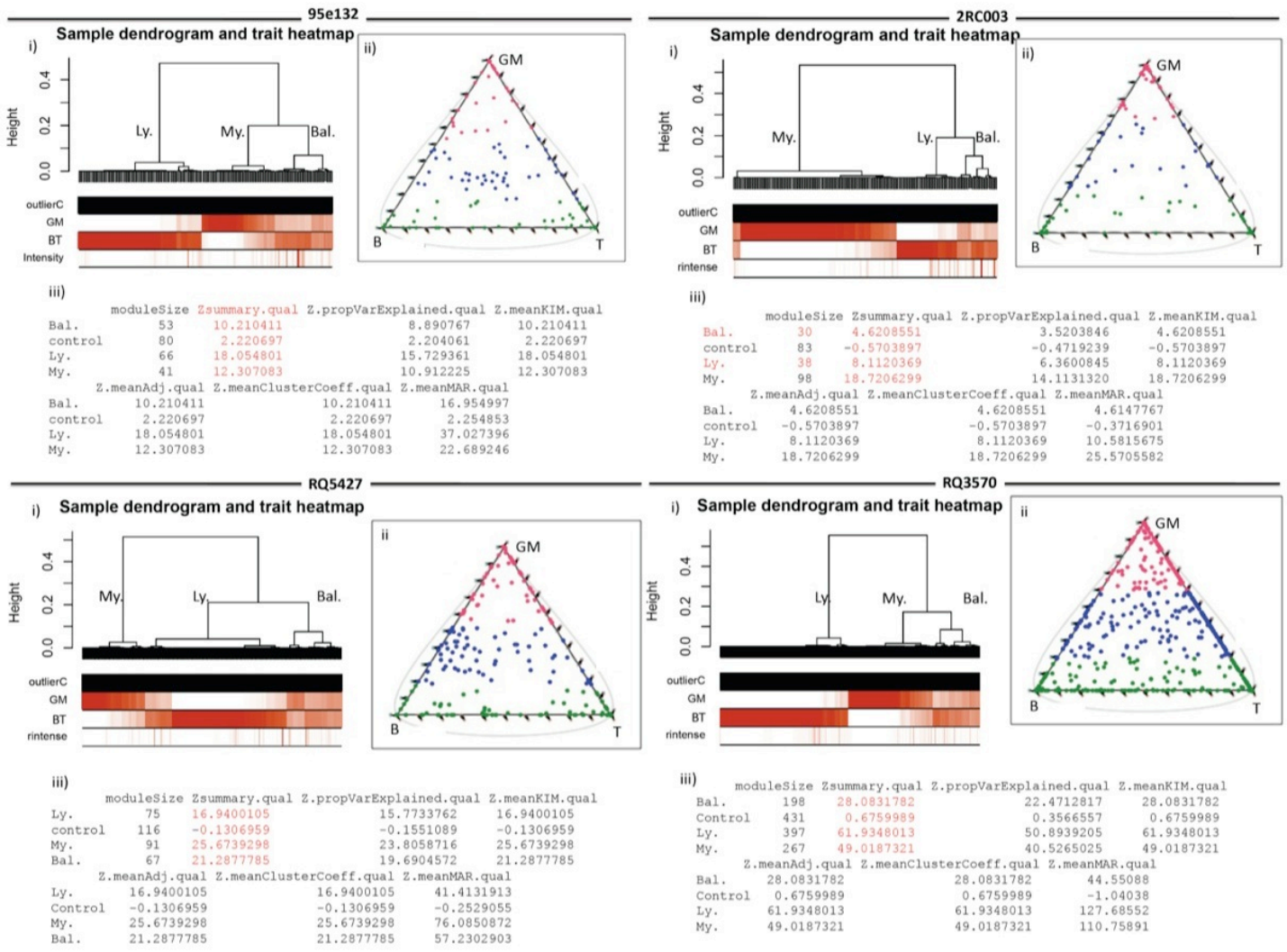

95E132
BT GM
CD4 CD8 CD20 CD14 CD18
CD34
PBC (months)
4 5 7 10 14 20 30 50 88 108 134 145
LT-QVIs (259 QVIs)
Lineage potential (> 0.0002)
ST-QVIs (258 QVIs)
Repopulation kinetics analysis
2RC003
BT GM
CD4 CD8 CD20 CD14 CD18
CD34
Grans (months)
3 9 15 29 40 52 64 81 103
PBMCs (months)
3 9 15 21 29 40 52 64 103
LT-QVIs (279 QVIs)
Lineage potential (> 0.02%)
ST-QVIs (335 QVIs)
Repopulation kinetics analysis
RQ5427
BT GM
CD4 CD8 CD20 CD14 CD18
CD34
Grans (months)
2 8 19 25 32 43 50 56 67
PBMCs (months)
2 8 13 19 25 32 43 50 56 67
LT-QVIs (477 QVIs)
Lineage potential (> 0.02%)
ST-QVIs (643 QVIs)
Repopulation kinetics analysis
RQ3570
BT GM
CD4 CD8 CD20 CD14 CD18
CD34
Grans (months)
2 8 13 19 32 38
PBMCs (months)
2 8 13 19 25 32 38
LT-QVIs (2,095 QVIs)
Lineage potential (> 0.02%)
ST-QVIs (2,491 QVIs)
Repopulation kinetics analysis

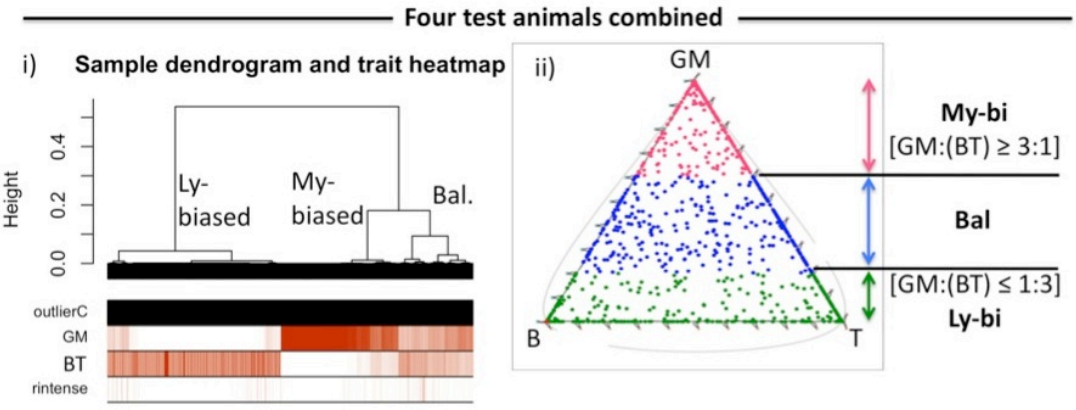
Figure S4. Determining Long-term and Short-term Repopulating Clones (related to Figures 3).
The chart shows how long-term (LT-QVI) and short-term (ST-QVI) repopulating clones were defined in animals 95E132, 2RC003, RQ5427, and RQ3570. The frequency profiles of individual QVIs are shown for mature lineages [lymphoid (CD4+, CD8+, and CD20+) and myeloid (CD14+, and CD18+)], mPB CD34+ HSPCs, and serial peripheral blood cells [PBC(95E132) or Grans/PBMC(the rest)] using a white-black-yellow color scheme. LT-QVI clones were determined based on the presence of QVI clones in mature lineages [lymphoid(BT) and myeloid (GM) cells] and in mPB CD34+ HSPCs near the end-point (36, 64, 70, or 116 months for animals RQ3570, RQ5427, 2RC003, and 95E132, respectively). The remainder were defined as ST-QVIs. The patterns of clonal repopulation kinetics were determined based on the clonal frequency profile of serial blood samples (Fig.S3). The lineage output potentials of LT-QVIs were determined based on the clonal frequency profile for each of the mature lineages tested (Fig.S5). LT-QVI clones with a detection frequency of ≥ 0.0002 (average of all mature lineages) were used.

Figure S5

(A)



(B)



iii)

	moduleSize	Zsummary.qual	Z.propVarExplained.qual	Z.meanKIM.qual	Z.meanAdj.qual	Z.meanMAR.qual
My-bi	458	54.258227	46.740551	54.258227	54.258227	159.7532423
Bal.	287	38.441980	32.977508	38.441980	38.441980	91.4427267
control	710	1.404442	1.349091	1.404442	1.404442	0.5919257
Ly-bi	676	75.869326	63.808215	75.869326	75.869326	218.9295941

Figure S5. Distinctive Lineage Output Potentials of My-bi, Bal, and Ly-bi LT-QVIs (related to Figure 4). **(A)** LT-QVIs in test animals showed three clearly distinct subgroups with myeloid-biased, lymphoid-biased, and balanced lineage output patterns. LT-QVIs from animals 95E132, 2RC003, RQ5427, and RQ3570 were clustered into three groups based on the average QVI frequencies in myeloid [GM (CD14+ and CD18+ cells)] and lymphoid lineages [BT: T-cells (CD4+ and CD8+)+B-cells (CD20+). i) Clones were clustered using the WGCNA package. Relative frequencies for GM and BT are shown in a red-white (100-0%) color scheme. Each of the three subtypes appears next to its corresponding branch. ii) A ternary diagram shows the three groups identified from cluster analysis. iii) The significance of the myeloid (My.)-biased, lymphoid (Ly.)-biased, and balanced (Bal.) was quantified using the module Preservation function in the WGCNA package. Cluster or "module" significance was evaluated using |Zsummary.qual| > 10. **(B)** The LT-QVIs of all four animals were clustered. Two borderline "GM:(BT)" ratios were determined to divide these LT-QVIs into the three subtypes. My-bi, Bal, and Ly-bi QVIs were defined by GM:(BT) ratios of ≥ 3:1, between < 3:1 and >1:3, and ≤ 1:3, respectively. Cluster or "module" significance was evaluated using |Zsummary.qual| > 10.

Figure S6.

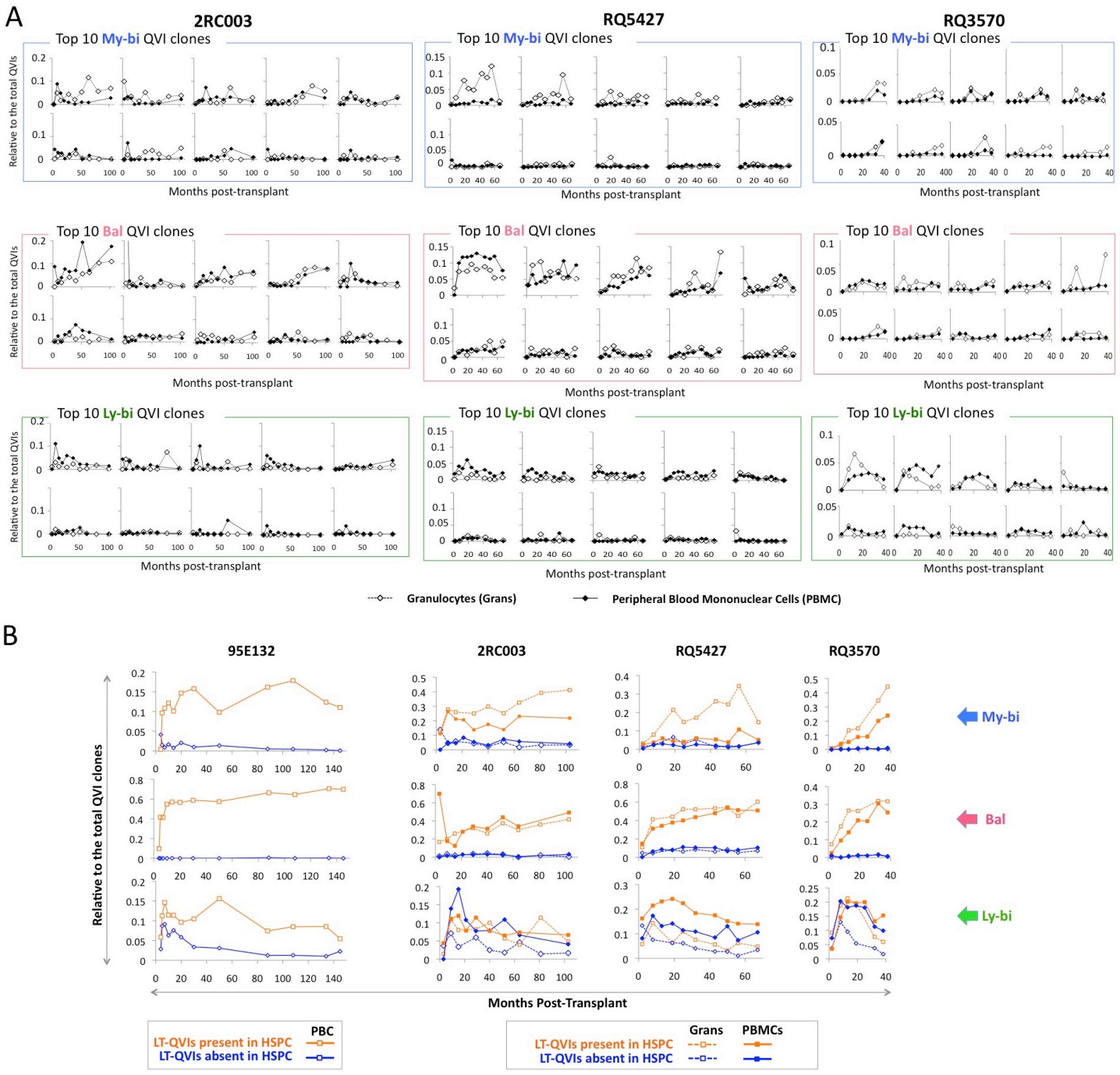
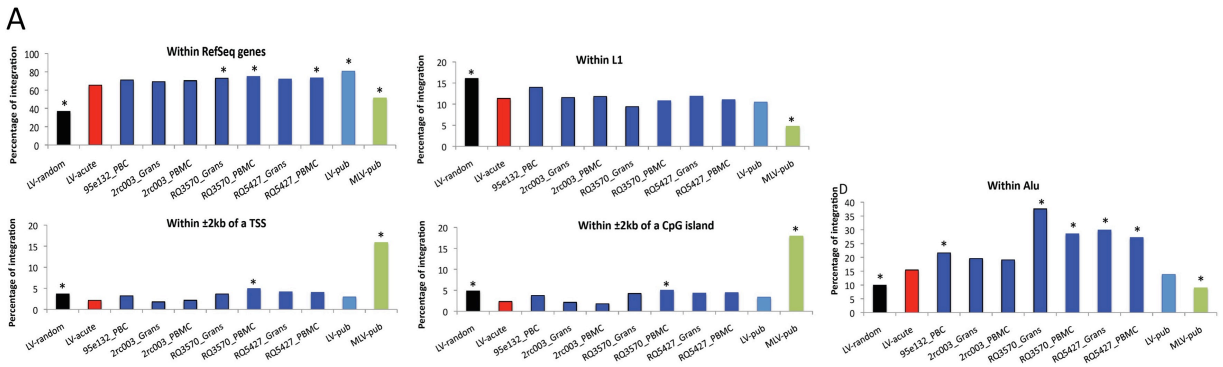
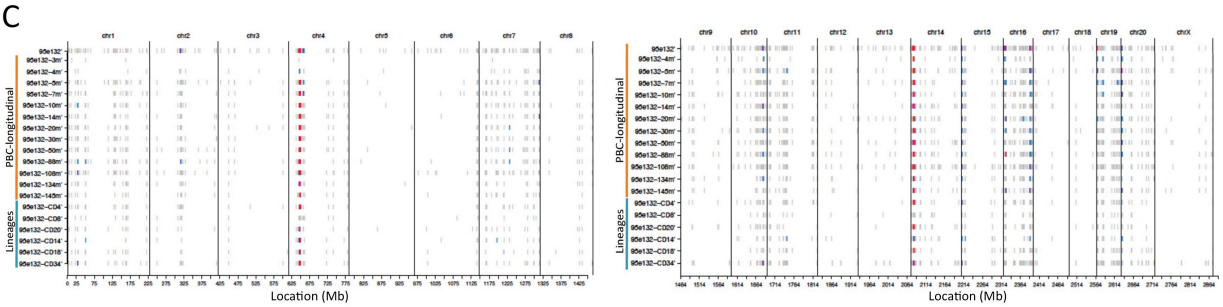


Figure S6. Repopulation Kinetics of My-bi, Bal, and Ly-bi LT-QVIs (related to Figure 4).

(**A**) The repopulation kinetics of the 10 most frequent QVI clones in each subtype are shown. The relative contributions of these clones clones for the My-bi (blue box), the Bal (magenta box), and the Ly-bi (green box) were plotted. The relative contributions of these clones to total Grans (open diamonds) and PBMCs (filled diamonds) are compared over time. (**B**) Comparison of the relative contributions over time between the LT-QVIs that were present and not present in the underlying CD34+ HSPCs. The total contributions by the My-bi, Bal, and Ly-bi QVI clones detectable in the mPB CD34+ HSPC (orange) are shown in comparison with those not detected in the HSPCs (blue).

Figure S7.





D

Groups	My-bi	Bal	Ly-bi	ST-QVI	Whole genome
Total genes	311	223	361	1,710	24,980
Overlapping genes with ST-QVI genes	134	115	150	1,710	1,710
Odds ratios (OR)	11.1	15.5	10.5	-	-
95% Confidence Interval (CI)	8.7-14.0	11.7-20.4	8.4-13.1	-	-

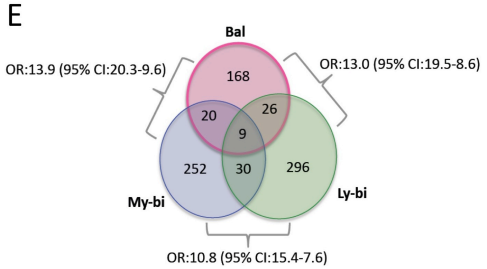


Figure S7. VIS analysis Indicates Benign Integration Bias Without Selection for Any Particular Cell Type (related to Figure 6)

(**A**) The percentage of VIS (y-axis) located within five known genomic features, including Refseq genes, ±2kb of transcription start sites (TSS), ±2 kb of CpG islands, Alu repeats, and L1 repeats. 10,000 random sites (LV-random) were generated *in silico* by selecting random integrants positioned within ±1 kb of a *Taqα I* site (TCGA) in the rhesus macaque genome, in order to simulate the results expected from the bi-directional VIS assay. Published data for lentiviral vectors (LV-pub [Hematti, 2004 #6934; Beard, 2007 #6933]) and murine leukemia virus vectors [MLV-pub [Beard, 2007 #6933;Calmels, 2005 #7229]] were re-analyzed using our standards and are included for comparison. Samples with P values <0.0005 [chi square test in comparison to acute infection data (LV-acute)] are indicated by an asterisk (*). **(B)** Magnification of two major VIS hot-spots located on chr 4 and chr 14 in rhesus macaques shows that VIS were spread throughout the multi-genic, mega-base scale hot-spots. Each set of data appears in a different row, starting from lentiviral VIS in freshly infected naïve CD34+ cells (LV-CD34-acute; black) and followed by lentivirus VIS from published data (LV-pub; gray); VIS in animals 95E132 (magenta), 2RC003 (blue), RQ3570 (green) and RQ5427 (brown); and RefSeq genes (dark blue). These results are displayed using custom tracks in the UCSC genome browser (genome.uscs.edu). The two major BCP hot-spots in chr 4 and 14 are magnified (red rectangles). (**C**) These figures show lentivirus VIS colored by BCP hot-spots for different time points and cell types in our test animals. Genomic VIS hot-spots for animal 95E132 are displayed. Each row shows the complete genomic VIS pattern for a given data set, where data set names are indicated on the y-axis, and the x-axis gives the genomic location of the VIS in Mb units. Vertical black bars indicate chromosome boundaries. Color definitions have been assigned for each data set independently, based on relative VIS density. Hot-spot regions appear in red. Data indicate a conservation of hot-spots over time and across cell types. Animals 2RC003, RQ5427, and RQ3570 showed similar patterns (data not shown). (**D**) The genes in the LT-QVI subtype (My-bi, Bal, and Ly-bi) data sets are compared with those in the ST-QVI data set. While hosting different VIS, 42–52% of the genes in My-bi, Bal, and Ly-bi were also found in the gene list of ST-QVIs, with odds ratios (OR) ranging from 10.5 [95% confidence interval (CI):8.4–13.1] to 15.5 [95% CI:11.7–20.4]. (**E**). This chart compares the genes of the My-bi, Bal, and Ly-bi subtype data sets. These genes, shown by a Venn diagram, overlapped significantly. OR and 95% CI are shown.

Table S1. Hematopoietic Reconstitution and Treatment History for Test Animals. (Related to Figure 1)

(**A**) Treatment History. Dates for leukapheresis, bone-marrow mobilization, and measles vaccination are noted. (**B**) Hematopoietic reconstitution. White blood cells (WBC), red blood cells (RBC), platelets (PLT), lymphocytes (Lymphs), monocytes (Monos), segmented neutrophils (Segs), Basophils (Basos), and Eosinophils (Eos) reconstitution are followed over time after CD34+ cell transplant. (**C**) Mature lineage and mPB CD34+ cells isolated near the end-point. The total number of cells and percent EGFP markings for CD14+, CD18+, CD4+, CD8+, CD20+, and CD34+ cells are noted. Table S1 is provided as a separate data file (excel file).

Table S2. Summary of VIS Sequence Analysis (Related to Table 1).

Abbreviations: PBC, peripheral blood cells; Grans, granulocytes; PBMC, peripheral blood mononuclear cells;VIS, Vector integration site; L, left junction; R, right junction; QVI, quantifiable vector integrants (Vector integrants with a TaqaI site within 450bp at either one of two or both junctions.); LT-QVI, long-term QVI (the QVI clones present in blood lineages isolated near the experimental end-point). Table S2 is provided as a separate data file (excel file).

[a] Analyzed vector-cellular DNA junctions

[b] Number of sequences that passed filtering after pyrosequencing.

[c] Number of uncertain sequences that include low quality sequences and vector sequences.

[d] Number of authenticated VIS sequences (vector DNA joined to cellular DNA)

[e] Unique VIS determined after VIS sequence matching and enumeration

[f] maximum sequence counts for a unique VIS

[g] Median sequence counts per unique VIS

[h] Average sequence counts per unique VIS

[i] Standard sequence deviation

[j] Unique VIS with a TaqaI site within 450bp from the junction (PCR amplicon <500bp)

* Data representing vector integrants determined by combining the left and the right junction data

[¶] Re-analyzed VIS sequence data from Kim, S. et al. JVI. 2010, PMID: 21876045

** Percentage EGFP for CD4, CD8, CD20, CD14, CD18 and CD34 were measured by flow-cytometry analysis during the cell isolation.

** Percent EGFP for PBCs are the average of the Lymphocyte, Monocyte, and Granulocyte EGFP marking in the longitudinal EGFP analysis data (see Table S1)

** Percent EGFP for Granulocytes and PBMCs are that of Granulocyte and Lymphocytes, respectively, in the longitudinal EGFP analysis data (see Table S1)

Table S3. Data for BCP Hotspots in the Genome (Related to Figure 6)

The table shows detailed BCP hotspot analysis results for lentivirus vector integration site data sets, including all animal data combined (LV-all animals), previously published data by other groups (LV-pub_data), animals 95E132, 2RC003, RQ5427, RQ3570, three LT-QVI subtypes (My-bi, Ly-bi, and Balanced) and short-term QVIs (ST-QVI). No hot-spot was detected in LV-acute. Table S3 is provided as a separate data file (excel file). Abbreviations: BCP hot-spots, vector integration site hot-spots (row) for each data set were defined by the BCP hot-spot definition (Presson, A. et al BMC Bioinformatics, 2011, 12:367); b.b1 and b.b2, boundaries defined by dividing each chromosome into consecutive 1Mb segments (bins); d.b1 and d.b2, the start (d.b1) and the end (d.b2) position of a hot-spot defined by insertion sites that are closest to b.b1 and b.b2 measured in bp's; VIS, number of insertions within [d.b1,d.b2]; Width, size of segment defined by d.b1 and d.b2; pctVIS (%), (number of VIS / Total VIS)*100; MbDens, VIS per Mbp; % Density (/Mb), pctVIS / Width (Mb).

Table S4. Vector Integration Site Sequences and Sequence Counts for Each Vector Integrants (Related to Table 1).

VIS Sequence data for all time-points and cell types as well as keys to the data table are provided as a separate data file (excel file).

Table S5. The Relative Detection Frequency for Each QVI (Related to Table 1).

The relative detection frequencies of QVIs for all time-points and cell types as well as keys to the data table are provided as a separate data file (excel file).

SUPPLEMENTARY EXPERIMENTAL PROCEDURES

Animal care and peripheral blood cell isolation.

Four rhesus macaques were previously transplanted with autologous CD34+ cells transduced with Self-inactivating lentivirus vector(An et al., 2007; An et al., 2001; Sander et al., 2006) and maintained in accordance with federal guidelines and the policies of the Veterinary Research Program of the National Institutes of Health by the protocols approved by the Animal Care and Use Committee of the National Heart, Lung, and Blood Institute. Peripheral blood cells (PBC) were isolated from EDTA-treated whole blood by ammonium chloride-mediated red blood cell lysis in 2-week to 6-month intervals and cryo-preserved. Peripheral blood mononuclear cells (PBMCs) and Grans from animals 2RC003, RQ5427 and RQ3570 were separated by the ficoll-hypaque gradient centrifugation and cryo-preserved every 6 months. Mature blood cell lineages (CD14, CD18, CD4, CD8, and CD20) were isolated from the leukapheresis cell product (without mobilization) by Ficoll-Hypaque density separation followed by red cell lysis and differential immunomagnetic selection. CD14+ and CD18+ cells were isolated based on size (forward scatter plot) and granularity (side scatter plot) as well as by CD14PE antibody (BD Pharmingen # 555398) or CD18 PE antibody (BD Pharmingen # 555924), respectively. CD20+ and CD8+ cells were enriched using CD8 beads (Dynal Invitrogen catalogue # 113.33D), and CD20 beads (Miltenyi Biotec catalogue # 130-191-105), respectively, utilizing differential magnetic separation, and further sorted using CD8PE (BD Pharmingen catalogue # 557086) and CD20 PE (BD Pharmingen catalogue # 556633), respectively. CD4+ cells were first enriched by using a Macs LS separation column (Miltenyi Biotec catalogue # 120-00-475), CD3 biotin antibody (Miltenyi Biotec catalogue # 130-092-008) and Streptavidin microbeads (Miltenyi Biotec catalogue # 481-01), and further sorted using CD4PE (BD Pharmingen catalogue # 550630). CD34+ HSPCs were isolated from the leukapheresis cell product collected from the peripheral blood after mobilization with granulocyte colony-stimulating factor (G-CSF) and stem cell factor (SCF) as described previously(An et al., 2001) within one month of and at least two weeks apart from the mature lineage isolation dates (Table 1). All the isolated lineages and CD34+ cells showed at least 99.4% purity except for CD4+ cells from RQ5427 and 2RC003 (both showed 94.1%) and CD20 from 2RC003 (98.7%). For animal 2RC003, 6 mg/kg of Ganciclovir IV was administered for two weeks to remove HSV1-sr39tk expressing cells at 23 months post-transplant (Table S1).

Bi-directional vector integration site assay.

Vector integration sites (VISs) in repopulating cell samples were sequenced by *TaqαI*-mediated bi-directional VIS sequencing assay as described previously(Kim et al., 2010) with modifications for different vector types for animals 2RC003, RQ3570, and RQ5427 (see Suppl.Methods for detail). > 5μg of genomic DNA per sample was used, except for CD4+GFP+ and CD8+ cells from 2RC003 and RQ5427 and CD14+ cells from 95E132

(0.7 – 2 μg was used) due to insufficient amount of cells available (Table S2). An average of 35.6 – 214.7 VIS sequence copies per vector integrant was analyzed.

i) Animal 95E132: total peripheral blood cells (PBCs) at 5, 50, 88, and 108 months were previously analyzed(Kim et al., 2010). All the other samples from this animal were processed by the same procedures. Individual samples were uniquely labeled by different 10 bp multiplex identifier (MID) sequences with modified A-4-Rg2 and B-4-Lg2 primers and subjected to 454-pyrosequencing as described previously. VIS PCR amplicons are shown in Fig.S1 (8% acrylamide gel electrophoresis). The summary of VIS sequence analysis is available in Table S2.

ii) Animal 2RC003: The lentivirus vectors (SIN18cpptRhMLV-E) used in this animal(Sander et al., 2006) required specific primer sets for VIS amplification during the bi-directional VIS assay due to modifications in the long terminal repeat (LTR) regions. The Experiment-1 approach of bi-directional VIS amplification(Kim et al., 2010) was used with modifications. Primers 9560R(5'-CAG GCT CAG ATC TGG TCT AA-3'), tB-MID-cppt2e3R (5'-CTA TGC GCC TTG CCA GCC CGC TCA G**NN NNN NNN NN**G CAG ATC TTG TCT TCG TTG G-3'), and tB-9005R (CTA TGC GCC TTG CCA GCC CGC TCA GTC ATT GGT CTT AAA GGT ACC-3') were used instead of Lg1, B-4-Lg2, and B-4-RgR, respectively. PCR conditions were same except that the first PCR was stopped at 25 cycles and the second PCR was stopped at 25 to 30 cycles, depending on the samples. Fig.S1 show VIS PCR amplicons on the 8% acryl amide gel. The summary of VIS sequence analysis is available in Table S2.

iii) Animals RQ3570 and RQ5427: SIV-based lentivirus vectors (SIVGAERhMLV-E) expressing shRNA against the CCR5 gene were used for these animals(An et al., 2007). Both the left and the right vector-host junction DNA were amplified by a modified bi-directional VIS PCR amplification procedure (Kim et al., 2010). First, two primers, SIV6935R (5'-GTC GTG GTT GGT TCC TGC C-3') and SIV2320F2 (5'-GTA AAG CGG CCG CAG ATC AAC ACG AGT TTT ATA AAA-3'), specific to the left and the right junctions, respectively, were bound and linearly extended in a 500 μl reaction solution containing 10 μg of genomic DNA, 1X picomaxx buffer (Agilent Technologies), 0.2mM of deoxynucleoside triphosphates (dNTPs), 0.1μM of S_6935R, 0.1μM of S_2320F2, and 25 units of picomaxx DNA Polymerase (Agilent Technologies) under the following conditions- 94°C for 3 mins, followed by incubation at 55°C for 5 mins and at 72°C for 5 mins. The junction DNA was then purified by the QIAquick PCR purification kit (QIAGEN) and digested with *TaqαI* (T/CGA). *TaqαI*-digestion produces different lengths of junction DNA depending on the location of the *TaqαI* site within the cellular DNA. Digested DNA was purified and then the *TaqαI* cleavage site in cellular DNA was ligated to a Linker DNA containing a 'CG' nucleotide 5' overhang complementary with the *TaqαI*-digested DNA. A Linker DNA was prepared by annealing LinkA (5'-CGG ATC CCG CAT CAT ATC TCC AGG TGT GAC AC-3') with TaqαILinkS (5'-CAC CTG GAG ATA TGA TGC GGG ATC-3'). Linker-ligated DNA was

amplified by a two-step PCR. The first PCR was carried out using 3 primers, including B-SIV6755Fcm(5'-/5BioTEG/CCT GGT CAA CTC GCT ACT CGG TAA TAA GA -3'), B-SIV2431Rcm(5'-/BioTEG/CTT CCT TTT CTA ATT ACA TGT CTA AGA TTC-3'), and Link1(5'-TAA CTG TCA CAC CTG GAG ATA-3'), to amplify both the left and the right vector-cellular DNA junctions in a 150 μl PCR mixture with 0.25 μM of B-SIV6755F, 0.25 μM of B-SIV2431R, 0.5 μM of Link1, 0.2mM of dNTPs, and 12 units of picomaxx DNA polymerase under the following conditions- 2 min of preincubation at 94°C, followed by 12 cycles of 94°C for 25 s, 56°C for 25 s, and 72°C 1.5 mins. DNA was purified using a QIAquick PCR purification kit (Qiagen). Amplified DNA, biotinylated at the 5' end of the vector side DNA, were bound to streptavidin-agarose Dynabeads (Dynal magnetic beads, Invitrogen) and separated from unbound DNA (in the supernatant) using a magnetic separator as described in the company manual. The bead-bound DNA was then stored into a 50 μl 1X picomaxx buffer. The two junctions were separately amplified by the junction-specific, second PCR. 5 μl of bead-bound DNA was used for the second PCR. For PCR amplification of the left junctions, primers SIV2405R (5'-GAT TCT ATG TCT TCT TGC AC-3') and Link2 (5'-GGA GAT ATG ATG CGG GAT C-3') were used. For the right junctions, primers SIV6788F (5'-CCT GGT CTG TTA GGA CCC TT-3') and Link2 were used. The conditions were identical to those for the first PCR, except that the second PCR was conducted with 29 cycles. The amplified DNA for each junction was then purified and processed for 454-pyrosequencing as described in the company's amplicon sequencing manual (Roche). Each sample was uniquely marked using 10 bp MID sequences located between the sequencing primer (A or B) and the target-specific primer (SIV2405R or SIV6788F). Vector integration site PCR amplicons from both the left and the right junctions were visualized in an 8% acrylamide gel (see Fig.S1). The summary of sequence analysis can be found in Table S2.

Analysis of VIS sequences

The VIS PCR products were purified and subjected to 454-pyrosequencing (Roche) using a Genome Sequencer FLX according to the company manual for amplicon sequencing, with emPCR conditions as described previously with modifications(Kim et al., 2010). VIS sequences contained homopolymer errors and other various erroneous sequences as seen previously(Kim et al., 2010). In order to identify and enumerate authentic VIS sequences from a pool of sequences that included varying degrees of sequencing errors, we performed VIS authentification analysis, followed by cycles of step-wise sequence homology tests using in-house programs as well as manual procedures. (i) VIS authentication. VIS sequences were authenticated by the existence of a vector-cellular DNA junction. Cellular DNA sequences were confirmed by their alignment with the rhesus macaque genome (Jan. 2006 rheMac2 assembly) and the human genome (Mar. 2006 hg18 assembly; NCBI Build 36.1) using BLAT (www.genome.ucsc.edu) and GMAP (www.gene.com/share/gmap/). (ii) Sequence enumeration for unique VIS. In order to determine VIS sequence counts in the data set, we performed cycles of step-wise sequence homology tests. Briefly, in the first round of tests, VIS sequences were clustered into groups

of similar sequences based on a high-stringency (100-95%) sequence homology, after which point the representative sequence of each group and an initial tally of matching sequences were subjected to a second round of sequence homology tests. In the second-round analysis, first-round products were clustered based on a lower-stringency sequence homology. In each round, we performed step-wise tests to group homologous sequences, enumerate the total number of sequences in each group, and determine the most representative sequence. These processes were repeated, lowering the sequence homology stringency of each successive cycle until we reached 80%. In the final step, to achieve a more complete analysis, we manually assigned certain remaining sequences with <80% homology. More details on these processes can be found in a previous study(Kim et al., 2010). (iii) Combining the Left- and the Right-junction VIS data. VIS sequences from the left and right junctions were matched based on the presence of the 5 bp palindromic overlap of the two VIS sequences aligned on the genome with an opposite orientation. When determining the total number of clones (vector integrants), no threshold was imposed. Once it passed our analysis criteria, each unique VIS sequence was considered a true event in the sample, even if it was detected only once. (iv) Determining QVI clones. As PCR amplicon > 500 bp was inefficiently sequenced in our assay conditions, only vector integrants that generate VIS PCR amplicons ≤500bp at either the left or the right junction were quantitatively analyzed [termed as quantifiable vector integrant (QVI)]. VIS PCR length was determined using the nearest *TaqαI* site in the reference genome (≤450bp from the junction). The relative frequencies of QVIs were determined by combining sequence detection frequencies from the left and right junctions [see details in (Kim et al., 2010)]. We estimated approximately 40.3% of randomly integrated vectors are QVI (Kim et al., 2010). In our data sets, 61–83% of detected unique vector integrants were QVI: this discrepancy may due to the assay limitation to detect vectors located distantly from *TaqαI* sites. As an average of 1.2 to 1.6 vector copies per cell was detected in three of our test animals (Fig.S2C). Therefore, given that approximately 40.3% of integrated vectors are QVI, the chance that each QVI represents a unique clone was 82.5% for animal 2RC003, 84.3% for animal RQ5427, and 66.7% for animal RQ3570.

Identifying and Controlling the Events of "Collisions"

Identical sequences shared by different animals were considered 'collisions'(Cartier et al., 2009). These events accounted for approximately 15% of total identified vector integrants. The correct identifiers (animals) for these sequences were clearly distinguishable from the wrong ones in most cases; the average difference in detection frequencies between the correct and the wrong identifiers ranged from 76- (±89 SD) to 1062- (±1086 SD) fold, depending on the test animal. Emulating previous analysis by Cartier, N. et al(Cartier et al., 2009), the wrong identifiers were removed when at least a 10-fold difference in sequence frequency was observed. About 0.8% of unique vector integrants remained undetermined after this process.

Kinetics Clusters

QVI clones with an average detection frequency ≥ 0.0002 (average of all time points in serial blood analysis) were subject for clonal kinetics analysis (Fig.S2D). Major clonal kinetics patterns were derived using the WGCNA R package v.1.20(Langfelder and Horvath, 2008), where one minus the correlation matrix was used as the distance measure between clones, and average linkage hierarchical clustering was used to cluster the clones within each data set (see Fig.S3A). Clusters were defined by using thresholds that were specific to each data set, where each threshold was selected to yield clusters with visually distinct kinetics profiles. The significance of these clusters was quantified using the modulePreservation function in the WGCNA package (Langfelder and Horvath, 2008) using parameters networkType = "signed", dataIsExpr = FALSE and default parameters (nPermutations = 200, randomSeed = 1, quickCor = 0). Cluster or "module" significance was evaluated using |Zsummary.qual| > 10. The following static cut heights (h) were used to generate clusters: 95e132 (h = 0.43, 7 clusters), 2rc003 Grans (h = 0.46, 7 clusters), 2rc003 PBMC (h = 0.46, 7 clusters), RQ3570 Grans (h = 0.5, 5 clusters), RQ3570 PBMC (h = 0.45, 6 clusters), RQ5427 Grans (h = 0.45, 7 clusters) and RQ5427 PBMC (h = 0.41, 8 clusters); where these heights were found to yield clusters with visually distinct kinetics profiles.

Analysis of Clonal Lineage Potentials

Analogous to previous studies(Dykstra et al., 2007; Lu et al., 2011), the differentiation potentials of individual LT-QVIs – differentiation into granulocytes/monocytes(GM), T-cells, and B-cells – were determined using ternary plots and cluster analysis (Fig.S5). CD14+ and CD18+ cells represented for GM. CD4+ and CD8+ cells were for T-cells. CD20+ cells represented for B-cells. Only LT-QVIs with an average frequency of ≥ 0.0002 (average over GM, T-cells, and B-cells) were used for lineage output potential analysis. It should be noted that since our cutoff value was based on an average across lineages (GM,T-cell,B-cell), we may have excluded some lineage-biased clones near the cutoff. Average linkage hierarchical clustering was used to group these LT-QVIs into different lineage potential subtypes using the WGCNA R package v. 1.20(Langfelder and Horvath, 2008). One minus the correlation matrix was used as the distance measure between clones and the significance of the My-bi, Bal and Ly-bi clusters was quantified using the modulePreservation function in the WGCNA package using parameters networkType = "signed", dataIsExpr = FALSE and default parameters (nPermutations = 200, randomSeed = 1, quickCor = 0). Cluster or "module" significance was evaluated using |Zsummary.qual| > 10. Fig.S5A shows lymphoid-, myeloid-biased, and balanced QVIs in each test animal. Based on cluster analysis on QVIs from all four test animals, we determined the GM:(TB) ratios of ≥ 3:1, between 3:1 & 1:3, and ≤ 1:3, to define My-bi, Bal, and Ly-bi subtypes, respectively (Fig.S5B).

Estimating the portion of long-term stem cells in mPB CD34+ HSPCs transplanted in test animals.

We estimate approximately 40.3% of randomly integrated vectors are QVI. Thus, the total number of long-term repopulating clones in each animal was calculated by multiplying the compensation factor (2.48) to the total number of LT-QVIs in each animal. As the total of 269, 279, 477, and 2,095 LT-QVIs were recovered for animals 95E132, 2RC003, RQ5427, and RQ3570, respectively, therefore a total of 667, 692, 1183, and 5196 clones, respectively, were expected to be repopulating for long-term in these animals. Since approximately 7×10^6, 9×10^6, 11×10^6, 30×10^6 EGFP+ CD34+ HSPCs were transplanted in animals 95E132, 2RC003, RQ5427, and RQ3570, respectively, approximately 0.0095%, 0.0076%, 0.0107%, and 0.0173% of transplanted EGFP+ CD34+ cells in these animals, respectively, were estimated to be long-term stem cells.

Characterization of genomic features hosting VISs.

A chi-square test with a 0.001 significance level was used to compare genomic characteristics between acute infection and repopulating cell VIS data and for other tests of categorical data. Wilcoxon rank sum tests were used to compare gene expression levels among data sets. Z-tests were used to test for a difference in proportions. The genes hosting VIS were determined based on the xenoRefSeqAli data in rheMac2 track in UCSC genome browser (www.genome.ucsc.edu). Only the hg18 RefSeq genes with best alignment onto the rheMac2 genome were chosen. These genes were characterized by canonical pathway analysis using using Ingenuity Pathways Analysis software versions 8 and 9 (Ingenuity® Systems, www.ingenuity.com) or functional annotation clustering using DAVID Bioinformatics Resources 6.7 (http://david.abcc.ncifcrf.gov) (Huang et al., 2008). P-values were adjusted to control for multiple testing using Benjamini-Hochberg's (BH) false discovery rate (Benjamini and Hochberg, 1995 Y (1995). Journal of the Royal Statistical Society Series B (Methodological) 57: 289:300). All reported p-values are two-tailed. As a control analysis, we generated an *in silico* control data set that were consistent with the bi-directional VIS assay data assuming the maximum PCR length that the assay can generate is 1 kb. Briefly, genomic sequences from a randomly selected site to the nearest *TaqαI* site (TCGA) at both upstream and downstream were collected and then those sequences that are < 1 kb were aligned back to the genome. Of these, we chose 10,000 sites that were unambiguously aligned to the genome as a control data set. We compared the overlap of genes hosting a VIS among the data sets including ST-QVI, My-bi, Bal and Ly-bi subtype clones, by calculating odds ratios (OR) in comparison to the total number of genes in the whole genome [a total of 24,980 best matching xenoRefSeqAli data (2010) for the rheMac2].

Genomic hot-spots for vector integration sites.

The Bayesian change-point (BCP) model(Presson et al., 2011) was used to define hot-spots among data sets with varying size. Briefly, the BCP method relies on first partitioning the genome into non-overlapping 1Mb bins and transforming the counts of VIS per bin to z-scores. A Bayesian change-point analysis of the z-scores

was used to determine local changes in z-scores, which segregated the data into regions of high and low VIS density. Megabase bins within high VIS density regions were considered "hot-bins", and the closest VIS to each hot-bin boundary was used to define "hot-spots".

Comparison of hot-spots in HSPC subtypes.

Comparisons among subtypes and between subtypes and LV-all were made using hot-bins to control for size differences between hot-spots. Hot-bins were megabase bins that were considered high density, and then hot-spots were defined from hot-bins by choosing the closest VIS to each hot-bin boundary and then merging overlapping hot-spots. In general there are more hot-bins than hot-spots because consecutive hot-bins were considered distinct and not lumped into a single hot-bin. Hot-bins among subtypes were significantly correlated, with overlap ranging between 29-79% among subtypes and overlap p-values (based on a Fisher's exact test for hot-bin overlap) ranging from 1.7e-07 to 3.3e-22. The strongest overlaps were observed between a) Ly-bi and balanced, where 11/14 (79%) of Ly-bi hot-bins overlapped with the balanced group (24 hot-bins), and b) Ly-bi and Low-Freq where again 11/14 Ly-bi hot-bins overlapped with 24 Low-Freq hot-bins. The weakest overlap was observed between Ly-bi and My-bi, where 4/14 (29%) of Ly-bi hot-bins overlapped with 12 hot-bins in My-bi. There was strong overlap between the subtypes and LV-all, where the overlap ranged between 30-83% and p-values ranged from 5.9e-13 to 9.3e-41. The strongest overlap was observed between LV-all and the Low-Freq subtype, where 19/23 (83%) LV-all hot-bins overlapped with the 12 Low-Freq hot-bins. The weakest overlap was between Lv-all and My-bi where only 7/23 Lv-all hot-bins overlapped with the 12 My-bi hot-bins. However, the overlap of My-bi hot-bins was 50% 6/12 with the LV-all hot-bins. The overlap of LV-all with Ly-bi and LV-all with balanced was 61% (14/23 in both cases).

Predicting the percent QVIs representing a single clone

The percent QVIs representing a single clone in test animals was calculated using R(*http://cran.us.r-project.org/*) *as* shown below;

```
## Average vector copy number (VCN) per cell
lambda1 = 1.60                    # animal RQ3570
lambda2 = 1.23                    # animal RQ5427
lambda3 = 1.26                    # animal 2RC003
## iteration number (into) and sample size (poisSS)
itno = 5000
poisSS = 1000
## generating poisson distribution samples
## we used adjusted lambda (adjlambda = lambda -1) to accommodate "0"s in samples (s1),
 (adjlambda1 = lambda1 -1)
(adjlambda2 = lambda2 -1)
(adjlambda3 = lambda3 -1)
## Start of iteration
pv=NULL
```

```
mv=NULL
for(k in 1:itno){
s1 = rpois(poisSS,adjlambda2)        ### change animals (adjlambda??) here ###
spoiss = s1+1                        # adjusted lambda
tp = table(spoiss)
## generate vectors in each clone
genSamples = function(tpi){
ntpi = as.numeric(names(tpi))
ftpi = as.numeric(tpi)
fv = rep(ntpi,ntpi*ftpi)
bv = rep(1:ftpi,each=ntpi)
sv=paste(fv,bv,sep=".")
return(sv)
}# end geneSamples
svv=NULL
for(i in 1:length(tp)){
svv=c(svv,genSamples(tp[i]))
}#end for
## choose QVIs (40.3 % of vectors)
snum = round(length(svv)*0.403)
ssvv = sample(svv,snum,replace=FALSE)
tabv = table(ssvv)
## Identify QVIs representing a single clone
pv=c(pv,sum(tabv==1)/length(ssvv))
## Identify clones with 2 or more vectors
mv=c(mv,1-sum(tabv==1)/length(tabv))
}#end of iteration
## results:
mean(pv) # avg % QVI from a unique clone
mean(mv) # average number of clones with two or more vectors in it.

## end
```

Quantitative Real-Time PCR to Determine Vector Copies Per Cell.

Genomic DNA of EGFP+ CD4+ cells was subjected to quantitative real-time PCR (qPCR). (i) The total number of vector copies was determined using primers specific to the EGFP gene [GFP-2F (5'-GTG GTG CCC ATC CTG GTC GA-3') and GFP-1R (5'-TCA GGG TCA GCT TGC CGT AG-3')]. qPCR was carried out using a 15 µl reaction mixture containing 1 µM of GFP-2F, 1 µl of GFP-1R, and 1X iTaq SYBR Green super mix (bio-rad) on a CFX96 Real-Time System (bio-rad) under the following conditions: 2 mins of pre-incubation at 95°C, followed by 40 cycles of 95°C for 22s, 60°C for 30s, and 72°C for 15s. Plasmid DNA containing the EGFP gene (FG12 vector) was diluted into 10^6–10^3 copies/µl and used as the standard. (ii) The total number of genome copies (cell copies) was determined using a primer set [P4 (5'-GTG GAG CAT TGT AGA TTC AT-3') and J1R (5'-AAG GCA GGA TTT GGG TAA CT-3')] targeting the spectrin, alpha, non-erythrocytic 1(SPTAN1) gene [the genomic DNA position chr15:10293598-10293713 (Jan.2006 RheMac2 assambly)]. This site was chosen after an extensive test of 5 different genomic sites. A plasmid DNA containing the target

genomic DNA [chr15:10293572-10293754, cloned into TOPO-TA cloning vector (Invitrogen)] was used as a standard control. The reaction conditions were identical to those of (i) vector copy qPCR.

Clone-specific Real-Time PCR.

Six QVI clones in animal 95E132 were subjected to two-step quantitative real-time PCR (qPCR) analysis. In order to use sufficient target copies for qPCR with limited samples, the qPCR target sites in the genomic DNA pool (5–20μl of genomic DNA) were initially amplified by a multiplex PCR then further quantified by the clone-specific 2nd PCR. Primer names, binding sites and sequences are same as those shown in (Kim et al., 2010). The multiplex PCR for the left and the right junction were carried out separately using 1μM of internal vector primers (LgC for the left or RgC for the right junction PCR) and a mixture of clone-specific Lc1 (for the left-side PCR) or Rc1 (for the right-side PCR) primers (0.2μM each) in a final volume of 100μl with and 0.2 mM deoxynucleoside triphosphates and 10U of picomaxx high fidelity PCR system (Agilent Technologies) under the following conditions: 2 min of pre-incubation at 95°C, followed by 16 cycles of 94°C for 25s, 58°C for 25s, and 72°C for 1.5 mins. The product of the multiplex PCR was purified using Qiaquick PCR Purification Kit (Qiagen) and one-tenth of the eluted DNA was subjected to the clone-specific qPCR. For the left junction clone-specific qPCR, a 25 μl reaction mixture containing 1st PCR product, 0.5μM of Lg1, 0.5μM of clone specific Lc2 primers, and 1X iTaq SYBR Green supermix with ROX (Bio-Rad) was prepared for each target site. qPCR was performed on an IQ-5 real-time PCR detection system (Bio-Rad) with following conditions: 1.5 min of preincubation at 95°C, followed by 40 cycles of 94°C for 10s, 58°C for 20s, and 72°C for 30s. The right junction clone-specific qPCR was carried out in the same conditions except using Rg1 and Rc2 primers. For standard copy number assays, the PCR target sites for the chosen clones were clones into TOPO-TA cloning vectors (Invitrogen) using either the primer sets, LcA, LgA, LcB and LgB (the left junctions) or another primer sets, RcA, RgA, RcB, and RgB (the right junctions). Plasmids copy numbers were determined based on the NanoDrop 1000 (Thermo Scientific). The mixture containing an equal molar of these control plasmids were serially diluted into 20 ng/μl rhesus genomic DNA in the range of 10^6–10^1 copies/10μl. Standard copy number qPCR was carried out by the identical two-step PCR procedures described above using the serially diluted, mixture of control plasmids. The clone-specific qPCR results were normalized by the total vector copies per sample. The total vector copies were measured by two-step PCR, where 1st PCR was carried out using either the primer sets LgC and Lv1 for the left junctions or RgC and Rv1 for the right junctions and followed by the qPCR using one-tenth of purified 1st PCR product, 1X iTaq SYBR Green supermix with ROX and the primer sets Lg1 and Lv2 (for the left junctions) or the primer sets Rg1 and Rv2 (for the right junctions). The conditions for the 1st PCR and the qPCR are same as above.

REFERENCES

An, D., Donahue, R., Kamata, M., Poon, B., Metzger, M., Mao, S., Bonifacino, A., Krouse, A., Darlix, J., Baltimore, D*., et al.* (2007). Stable reduction of CCR5 by RNAi through hematopoietic stem cell transplant in non-human primates. Proc Natl Acad Sci U S A *104*, 13110-13115.

An, D., Kung, S., Bonifacino, A., Wersto, R., Metzger, M., Agricola, B., Mao, S., Chen, I., and Donahue, R. (2001). Lentivirus vector-mediated hematopoietic stem cell gene transfer of common gamma-chain cytokine receptor in rhesus macaques. Journal of Virology *75*, 3547-3555.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Cartier, N., Hacein-Bey-Abina, S., Bartholomae, C., Veres, G., Schmidt, M., Kutschera, I., Vidaud, M., Abel, U., Dal-Cortivo, L., Caccavelli, L*., et al.* (2009). Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. Science *326*, 818-823.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S., Brinkman, R., and Eaves, C. (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. Cell Stem Cell *1*, 218-229.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols *4*, 44-57.

Kim, S., Kim, N., Presson, A., An, D., Mao, S., Bonifacino, A., Donahue, R., Chow, S., and Chen, I. (2010). High-throughput, sensitive quantification of repopulating hematopoietic stem cell clones. Journal of Virology *84*, 11771-11780.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lu, R., Neff, N., Quake, S., and Weissman, I. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature Biotechnology *29*, 928-933.

Peccoud, J., and Jacob, C. (1996). Theoretical uncertainty of measurements using quantitative polymerase chain reaction. Biophysical Journal *71*, 101-108.

Presson, A., Kim, N., Xiaofei, Y., Chen, I., and Kim, S. (2011). Methodology and software to detect viral integration site hot-spots. BMC Bioinformatics *12*, 367.

Sander, W., Metzger, M., Morizono, K., Bonifacino, A., Penzak, S., Xie, Y., Chen, I., Bacon, J., Sestrich, S., Szajek, L*., et al.* (2006). Noninvasive molecular imaging to detect transgene expression of lentiviral vector in nonhuman primates. Journal of Nuclear Medicine *47*, 1212-1219.