BARCODING OF MACAQUE HEMATOPOIETIC STEM AND PROGENITOR CELLS: A ROBUST PLATFORM TO ASSESS VECTOR GENOTOXICITY

A Dissertation
Submitted to the Faculty of the
Graduate School of Arts and Sciences
Of Georgetown University
In partial fulfillment of the requirements for the
Degree of
Doctor of Philosophy
In Microbiology and Immunology

By

Idalia Montano Yabe, M.S.

Washington, DC
March 1, 2019
BARCODING OF MACAQUE HEMATOPOIETIC STEM AND PROGENITOR CELLS: A ROBUST PLATFORM TO ASSESS VECTOR GENOTOXICITY

Idalia Montano Yabe, M.S.

Thesis Advisor: Cynthia E. Dunbar, M.D.

ABSTRACT

Gene therapies using integrating retroviral vectors to modify hematopoietic stem and progenitor cells have shown great promise for treatment of immune system and hematologic diseases. However, activation of proto-oncogenes via insertional mutagenesis has resulted in the development of leukemia. I have utilized cellular “barcoding” to investigate the impact of different vector designs on the clonal behavior of HSPC during in vivo expansion as a quantitative surrogate assay for genotoxicity in a non-human primate model with high relevance for human biology. I transplanted two rhesus macaques with autologous CD34+ HSPC transduced with three lentiviral vectors containing different promoters/enhancers of a predicted range of genotoxicities, each containing a high diversity barcode library that uniquely tags each individual transduced HSPC. Analysis of clonal output from thousands of individual HSPC transduced with these barcoded vectors revealed sustained clonal diversity, with no progressive dominance of clones containing any of the three vectors for up to almost three years post-transplantation. This data supports a low genotoxic risk for lentiviral vectors in HSPC, even those containing strong promoters/enhancers. Additionally, this flexible system can be used for testing of future vector designs.
The research and writing of this thesis is dedicated to my family. My husband Keisuke Yabe who went above and beyond in his support, encouragement and patience throughout this journey beginning with my graduate school application, listening to my research presentations, editing my writing, and helping keep my sanity in check. Our daughter Mela Yabe who taught me how to be more efficient with my time and improved my ability to multi-task. My mom Lilia Gutierrez De Montano who was an advocate for learning, exploring, and believed Mexican women were more than just cooks in the kitchen. My father Jose Montano who worked hard in the farm and brought us to the US for better opportunities. And last but not least, to our “son” Melo Yabe, who brought peace and comfort during stressful periods.

First and foremost, I would like to thank my thesis advisor, Cynthia Dunbar for being an incredibly supportive mentor. Thank you for all the time you spent with me, I appreciated each and every one of our discussions, and I never once walked out of your office without learning something new. Thank you for always making yourself available, no matter how busy you were and even when it was last minute. It is your passion for science, wealth of knowledge, and incredible writing skills which inspired me since the day I joined your lab.

And finally, thank you for your genuine caring nature and expert guidance which had a tremendous impact in my growth and development in becoming a better scientist. From the very first time we met in the interview, I knew you were the type of individual who would ensure that anyone under your watch would not fail, and in fact, thrive and succeed. I cannot imagine having a better mentor.

Second, I would like to thank Dr. Sharon Milgram from OITE, NIH for her invaluable support throughout the program, mentorship, funding that provided me the opportunity to continue in the program, and successfully find my thesis advisor.

Next, I would like to thank my lab mates for their friendship and scientific advice. In particular, I would like to thank Lauren Truitt, who patiently helped me with the bioinformatic perspectives and more. Sandhya Panch for the invaluable long night discussions, bouncing ideas, and fond memories in the lab. To Kyung-Rok Yu, Marcus Corat, Thomas Winkler, Chuanfeng Wu, and SoGun Hong for productive discussions, protocol troubleshooting, and great companionship in the lab. To Stephanie Sellers for sharing protocols, making time for questions, and for creating a supportive environment. And to all my other lab members who although I may not have worked closely with, each played a role in providing a healthy and enjoyable scientific environment.

I would like to thank each and every one of my committee members, Drs. John Casey, Richard Calderone, and Radhakrishnan Padmanabhan from the Department of Microbiology and Immunology at Georgetown University and Drs. Cynthia Dunbar, Harry Malech, and Andre Larochelle from the National Institutes of Health.

Finally, I am grateful for the support of Dr. William Fonzi from Georgetown University, along with my fellow classmates and many others that are too many to mention here.

Many thanks,
Idalia Montano Yabe
# Table of Contents

**Introduction** ......................................................................................................................... i

**Chapter 1. Background** .......................................................................................................... 4

- Hematopoietic stem cells ........................................................................................................ 4
- Animal models and autologous transplantation ..................................................................... 8
- Retrovirus characteristics ........................................................................................................ 12
- Replication and reverse transcription ..................................................................................... 17
- Chimeric HIV1-based lentiviral vector system design ............................................................. 19
- Gene therapy .......................................................................................................................... 22
- Safety of integrating vectors ..................................................................................................... 29
- Figures ..................................................................................................................................... 41

**Chapter 2. Materials and Methods** ......................................................................................... 46

- Construction and production of barcoded lentivirus vectors .................................................. 46
- Transduction and autologous transplantation of rhesus macaque CD34+ HSPC ..................... 52
- Cell processing and flow cytometric sorting ........................................................................... 55
- Barcode retrieval ..................................................................................................................... 56
- Barcode retrieval and data analysis ......................................................................................... 57
- K562 validation of quantitative simultaneous barcode retrieval from 3 lentiviral libraries ...... 58
- Figures and Tables .................................................................................................................. 60

**Chapter 3. Results I: Development of Barcoded Vectors** ..................................................... 76

- Development of an in vivo platform to study relative vector toxicity .................................... 76
- Figures and Tables .................................................................................................................. 79

**Chapter 4. Results II: In vivo Data from Autologous Transplantation** .................................. 84

- Post-transplantation analysis of vector clonal contributions over time ................................. 84
- Clonal diversity and clonal dynamics over time ..................................................................... 85
- Clonal stability of vectors by their top-contributing clones .................................................... 87
- Figures ..................................................................................................................................... 90

**Chapter 5: Discussion** ............................................................................................................ 97

**References** .............................................................................................................................. 102
List of Figures

Figure 1. HSC hierarchical model .................................................................41
Figure 2. Schematic of the structure of a retrovirus particle .........................43
Figure 3. Lentiviral entry into a target cell ....................................................44
Figure 4. Schematic of lentiviral third generation vector ..........................45
Figure 5. Schematic of the construction of lentiviral vector libraries ............60
Figure 6. Virus serial dilutions .................................................................62
Figure 7. Virus titer calculation .................................................................63
Figure 8. Barcoded library vector diversity ...............................................64
Figure 9. HSPC purity ...............................................................................65
Figure 10. Cell sorting strategy for T, B, and granulocyte cell populations ....66
Figure 11. Cell sorting strategy for monocyte and NK cell populations ..........68
Figure 12. Threshold comparisons ............................................................69
Figure 13. Impact of threshold variation on clone numbers and top clone contribution plots ........................................................................70
Figure 14. Southern blot for single clone identification ...............................72
Figure 15. Design of barcoded vectors .......................................................79
Figure 16. Analysis of comparative promoter strength of each lentiviral vector ...80
Figure 17. Experimental design and GFP marking post-transplantation .........81
Figure 18. Predicted single copy transduction efficiency of CD34+ HSPC ....82
Figure 19. Vector contributions and clonal diversity of EF1-α, MSCV, and SFFV barcoded vectors ....................................................................90
Figure 20. Validation of barcode retrieval to assess clonal contributions ..........92
Figure 21. Diversity of barcoded vectors in hematopoiesis ..........................94
Figure 22. Clonal stability of top contributing clones ....................................95
Figure 23. Heatmaps showing lineage contributions of top contributing clones ....96

List of Tables

Table 1. Viral titers for MSCV vector library used in ZJ48 ........................73
Table 2. Summary of virus titrations of EF1a, MSCV, and SFFV for animals ZJ41 and ZJ48 ...............................................................73
Table 3. Primer sequences used for multiplexing illumina high-throughput sequencing ..............................................................................74
Table 4. ZJ41 CBC results for the various steps during apheresis and the final apheresis product .......................................................................74
Table 5. ZJ48 CBC results for the various steps during apheresis and the final apheresis product .......................................................................75
Table 6. Fluorochrome-conjugated antibodies .............................................75
Table 7. Macaque information ..................................................................83
Introduction

The first clinical trials demonstrating clear improvement following transplantation of genetically-modified hematopoietic stem and progenitor cells (HSPC) were reported 15 years ago, utilizing \(\gamma\)-retroviral (\(\gamma RV\)) vectors based on the Moloney murine leukemia virus (MLV) to treat congenital immunodeficiency disorders 1,2. However, soon thereafter patients developed both myeloid and lymphoid acute leukemias due to insertional activation of nearby proto-oncogenes by vector enhancer sequences 3-7. Lentiviral (LV) vectors with deletion of viral enhancers were already in development and have subsequently been used in promising HSPC gene therapy clinical trials, but have also been linked to at least one instance of clonal expansion 8-10.

A number of murine models and *in vitro* assays have been explored for prediction of HSPC genotoxicity, but each has potential limitations. Tracking of insertion sites following transplantation of transduced normal murine HSPC requires large numbers of mice to track relatively low numbers of clones, and very long-term follow-up and/or secondary transplants are required to uncover clonal dominance and transformation 11-13. Murine tumor-prone models have more rapid and penetrant tumor onset, however the genes linked to tumors in these models do not closely match the responsible loci in human clinical trials, potentially decreasing preclinical utility 14, 15. *In vitro* immortalization of murine HPSC has been developed as a more practical and rapid screening approach, however this assay is difficult to quantitate, captures myeloid but not lymphoid transformation, and identifies only insertions activating Evi1 pathways 16-18.

Both *in vivo* and *in vitro* murine assays may lack predictive value for humans due to differences in HSPC properties between species, including lifetime replicative demand,
HSPC frequency, ease of immortalization, telomere length, and spectrum of common hematologic tumors\textsuperscript{19}. Non-human primates (NHP) are phylogenetically closely related to humans, and have been predictive of human clinical results regarding gene transfer efficacy and insertion site patterns\textsuperscript{19, 20}. While NHP models are expensive and technically-demanding, they can be used to study the behavior of thousands of vector insertion sites in a small cohort of animals over time. However, development of tumors in NHP is rare and requires long-term follow-up\textsuperscript{21, 22}, thus an approach to monitor genotoxic clonal expansion prior to actual malignant transformation would be of utility.

Insertion site identification and retrieval methodologies have evolved over time from detection of restriction fragment length polymorphisms to PCR-based approaches for retrieval of genomic sequences flanking proviral insertions. Linear amplification-mediated PCR (LAM-PCR) and modifications avoiding bias associated with restriction enzyme fragment generation have been the preferred methods for retrieving proviral integrations in both preclinical and clinical studies, and have proven invaluable for identifying insertions associated with tumors, as well as providing an overview of insertion patterns with various vector classes\textsuperscript{4, 23, 24}. However, these insertion retrieval approaches require large amounts of DNA and are semi-quantitative, thus unable to reproducibly quantitate several fold changes in individual clonal contributions\textsuperscript{25}. An alternative for clonal tracking is the inclusion of high diversity barcodes within the vector backbone. In both murine\textsuperscript{26} and non-human primate\textsuperscript{27} studies, HSPC clonal dynamics have been mapped in a quantitative and efficient manner using this approach.

In this dissertation, we now utilize barcoding to investigate the impact of specific promoter and enhancer elements within HIV-derived lentiviral vectors (LV) on clonal
dynamics. We focused on LV because they are the current standard for efficient
transduction of HSPC in the clinical setting, having replaced γRV due to higher
efficiency and lack of overt genotoxicity to date. The specific aim was to assay vectors
containing one of three promoter/enhancers with a range of predicted genotoxicity. The
broader goal was to develop and validate an approach for screening genotoxicity in an
animal model that is predictive of results in human HSPC gene transfer applications.
Chapter 1. Background

Hematopoietic stem cells

Hematopoietic stem cells (HSC) are committed tissue-specific multipotent stem cells located at the very top of the hematopoietic hierarchy and are distinguished by their unique ability to self-renew and sustain production of all blood cell lineages for the life of the organism\textsuperscript{28,29}. HSC have the potential to differentiate into multipotent progenitor (MPP) cells that have reduced self-renewal activity then further differentiate into oligopotent progenitors and finally, unipotent progenitors giving rise to mature and terminally differentiated blood cells. As new information on HSC differentiation is uncovered, the classical HSC hierarchical tree is challenged, and the exact pathway of HSC differentiation remains unknown and yet to be defined. However, based on the combination of the classical HSC hierarchical dogma and various hypothesis, Figure 1 demonstrates that MPP give rise to oligopotent progenitors that include the common myeloid progenitors (CMP) which give rise to the myeloid lineages, megakaryocytes, and erythroid lineage, natural killer cell progenitor (NKP) that give rise to natural killer cells, and lymphoid progenitors (LP) which give rise to T cells and B cells. The classical model describes T and B cells evolving from a common lymphoid progenitor, however, studies including those from our lab, have challenged this dogma by suggesting that T and B cells do not share ancestry despite both being lymphocytes functionally. Instead, B cells have more ancestry with the myeloid lineage suggesting that T and B cells do not come from a shared progenitor\textsuperscript{27}. The CMP gives rise to oligopotent megakaryocyte-erythroid progenitor (MEP) and granulocyte-monocyte progenitor (GMP). The MEP generate megakaryocytes and erythrocytes, although there have been studies suggesting
that megakaryocytes are derived directly from a multipotent cell such as HSC/MPP\textsuperscript{30}
while the GMP produce granulocytes (neutrophil, eosinophil, basophil) and monocytes. Megakaryocyte in turn gives rise to platelets and monocytes further differentiate into
macrophages or dendritic cells (Figure 1).

Self-renewal is a process defining HSC in which each cell has the potential to
regenerate itself to theoretically replenish all hematopoietic cells for the lifespan of an
organism. Blood cells are typically lost as a result of normal physiologic function, injury,
disease, or from the basic process of aging and HSC are responsible for restoring these
lost cells. HSC homeostasis is achieved through highly regulated and sophisticated
mechanisms provided within the bone marrow niche, which is a specialized
microenvironment that regulates HSC via a complex network of cytokines, cell-cell
interactions, and cell-extracellular matrix interactions, controlling metabolic functions,
signaling pathways, transcription factor expression and function, and epigenetic
pathways. The bone marrow niche is tightly regulated in order to achieve and maintain a
fine balance between HSC proliferation, differentiation, survival, self-renewal, and
trafficking\textsuperscript{31, 32}, all of which continue to be heavily researched areas.

HSC are a heterogeneous population\textsuperscript{33-39} consisting of short-term HSC (ST-HSC)
and long-term HSC (LT-HSC) that home to the bone marrow via cell signaling pathways.
ST-HSC are an abundant population that make up the majority of HSPC, they are derived
from LT-HSC, are defined by their committed nature, loss of their self-renewal ability,
engraft transiently, and fail to restore the immune repertoire. LT-HSC are much less
abundant and only make up a small fraction of HSPC population. However, LT-HSC
have the ability to self-renew and have the potential for long-term reconstitution. These
characteristics are what makes LT-HSC great targets for gene therapies. The total number of HSC can be predicted based on transplantation analysis and for non-human primates (NHP) the number of HSC predicted is approximately 1 ST-HSC for every 5,000 HSPC and 1 LT-HSC for every 20,000 HSPC and in murine models is approximately 1-8 HSC in 100,000 nucleated BM cells\textsuperscript{40-46}. There are also HSC that are released into the bloodstream, however, peripheral blood HSC make up a very small number of HSC and little is known about the physiologic function of HSC that circulate in the peripheral blood. Data from our lab suggests under most conditions, egress from the marrow is an HSC death pathway\textsuperscript{47}.

Three decades ago, murine bone marrow HSC became the first tissue-specific stem cells to be characterized and isolated\textsuperscript{48}. The methods used then and still in use today to purify HSC and downstream progenitors are based on differentially-expressed cell-surface markers and fluorescence-activated cell sorting (FACS) or immunoabsorption methodologies. Over time, techniques have improved based on identification of new cell-surface markers\textsuperscript{49}, and sorting technologies allowing simultaneous detection of multiple parameters. In humans, LT-HSC are generally identified using the following cell-surface markers: Lin\textsuperscript{-}CD34\textsuperscript{+}CD38\textsuperscript{-}CD90\textsuperscript{-}CD45RA\textsuperscript{-}CD49f\textsuperscript{+}\textsuperscript{30, 50, 51}. Lineage negative (Lin\textsuperscript{-}) cells describe cells that have been negatively selected for mature and terminally differentiated cell markers, for example, it depletes T cells, B cells, NK cells, monocytes, dendritic cells, and granulocytes. The cell surface marker, CD34, identifies a heterogenous subset of cells that includes both ST-HSC and LT-HSC as well as many differentiated progenitors\textsuperscript{52-55}. CD38 is a cell surface marker used to further distinguish between committed progenitors and more primitive CD34+...
CD38-expressing cells represent the majority of CD34+ cells and identify more mature and committed progenitors, in contrast to CD38- cells responsible for more engraftment ability. The CD90 molecule is used to identify LT-HSC, it is expressed in various types of stem cells that include HSCs and when combined with CD34 expression, it identifies the most purified primitive LT-HSC. CD49f, an integrin, can further isolate HSC from other progenitors, such as MPP, providing a purer HSC population. The leukocyte common antigen CD45RA isoform is used to further enrich for more primitive HSC.

Blood cells vary in their life span ranging from hours to years, with massive shifts in the demand for mature cells of specific lineages in response to growth, bleeding, infection or other stressors. For instance, myeloid cells may circulate for only a few hours-days, platelets for several days, and red blood cells for up to 4 months in humans. Memory T cells can persist and self-renew peripherally for many years. B cells can differentiate to plasma cells and likewise persist for years. In a healthy adult, the hematopoietic system must produce an average of 1.5E+06 blood cells every second while at the same time, having to maintain HSC self-renewal and protect HSC from acquisition of deleterious mutations and transformation to uncontrolled growth and neoplastic transformation. Not surprisingly, hematopoiesis can change with aging, and studies from our lab in transplanted rhesus macaques suggest that long-term multipotent clones are less potent and more slowly contributing with increased age, with young and aged macaques showing different kinetics and lineage reconstitution patterns.

HSC have developed into great targets for therapeutic applications in gene therapy, due to their unique ability to self-renew and produce all blood cell lineages,
meaning that therapeutic genetic modification of these cells should result in lifelong therapeutic benefit. HSC gene therapy has been used in the treatment of various primary immunodeficiencies and degenerative disorders. However, because HSC reside mostly in a quiescent state and cycle infrequently dividing only every 1-2 weeks in murine and primate models\textsuperscript{64-66}, choosing the right vector that is capable of transducing quiescent cells without the requirement for the target cells to undergo cycling is critical for successful gene delivery\textsuperscript{67, 68}.

Animal models and autologous transplantation

We have chosen to use non-human primate (NHP), rhesus macaque as our experimental animal model as it most closely resembles human physiology, particularly in the hematopoietic and immune systems. Our rhesus macaque are not inbred animals and thus like humans, they are genetically diverse. Mice and other small laboratory animals, on the other hand, are extensively inbred. Consequently, results from studies carried out in smaller mammals can be biased depending on the genetic background chosen. Although mice have been critical in our understanding in gene therapy, they simply do not provide an accurate prediction of human outcomes, as became apparent based on results of early human clinical trials, designed based on encouraging experiments in rodents. Mice and humans differ in various aspects including size and lifespan, which results in exponentially different hematopoietic demands, for example, larger animals require more cells so HSC from a large animal will differ from a small animals in HSC proliferation, differentiation, self-renewal, cell cycling and quiescent state\textsuperscript{19}. Rhesus macaques have life-spans of up to 35 years compared to 2 years for mice,
thus they more closely resemble humans and allow for long-term studies. Transduction efficiencies of HSC also differ, for example, murine HSC are more easily transduced than human HSC which can lead to lower marking of gene corrected cells in human trials\textsuperscript{19, 69, 70},

Autologous HSPC transplantation involves use of the individual’s own HSPC for \textit{ex vivo} manipulation followed by infusion back into the same individual. Advantages of autologous HSPC over allogeneic transplants include bypassing the need for an HLA-matched donor, avoiding the use of immune suppressants, and importantly, obviating concern for immune complications such as graft-versus-host disease (GVHD) and rejection. However, genetic correction or modification of autologous cells will be necessary to treat target disorders.

The methods used to isolate and collect HSPC from patients has changed over the past 39 years from the more invasive approach of surgical collection from the iliac crest and other bones to the less invasive approach of obtaining HSPC from the peripheral blood. Early studies on peripheral blood stem cells (PBSC) involved the collection of stem cells at steady state without the use of cytokine stimulation\textsuperscript{71-73}. These studies validated the concept that HSPC were present in the blood, although at very low level. The realization that the number of HSPC present in the blood could be markedly increased via stimulation with granulocyte colony-stimulating factor (G-CSF)\textsuperscript{74, 75} revolutionized collection of HSPC for transplantation. In the BM, stromal cells anchor CD34+ cells through complex adhesive interactions and G-CSF, through a multistep process involving both protease-dependent and independent mechanisms, disrupts the interactions between CD34+ cells and stromal cells resulting in the release of CD34+
cells into the peripheral blood\textsuperscript{76, 77}. G-CSF is the most common mobilizing agent, its mobilizing activity is based on reducing the chemokine stromal cell-derived factor 1 (SDF1, also known as CXCL12) in the marrow and disrupting other adhesive interactions, resulting in release of HSPC into the blood\textsuperscript{77-82}. G-CSF has been used for many years to mobilize CD34+ cells from the bone marrow to the peripheral blood and is considered the clinical “gold standard” for HSC mobilization. However, there is a small percentage of patients that mobilize poorly with G-CSF, and for this reason, there has been an interest in identifying additional mobilizing agents.

The compound AMD3100 (AMD stands for AnorMeD, chemical name now is plerixafor) is a bicyclam, first discovered as a putative anti-HIV activity\textsuperscript{83, 84} and later used to mobilize stem cells from the bone marrow into peripheral blood. AMD3100 specifically binds to the C-X-C chemokine receptor type 4 (CXCR4) expressed on the surface of cells, and in the case of HIV, where the virus binds to CXCR4 on the surface of cells to initiate viral entry, AMD3100 antiviral activity functions as a potent antagonist of CXCR4, blocking HIV-1 binding to the cell and preventing the virus from fusing with the cell membrane and as a result, blocks viral entry. Its mobilizing ability was discovered upon treating HIV patients where it was observed that not only were virus levels reduced upon administration of AMD3100, but also a sudden increase in the number of white blood cells in the peripheral blood. Further study indicated that CD34+ HSPC were also increased in the blood soon after AMD3100 administration.

Phase I/II/III studies confirmed the safety and efficacy of AMD3100, and the US Food and Drug Administration (FDA) approved AMD3100 (plerixafor; brand name
Mozobil, Genzyme Corp) in 2006 for mobilizing HSC and to this date, AMD3100 is the only FDA-approved CXCR4 antagonist for use in mobilizing HSC.

Mobilized PBSC collections not only involve a less invasive procedure but also result in collection of larger numbers of HSPC compared to bone marrow harvests. Mobilized PBSC are also thought to engraft more rapidly compared to stem cells derived from bone marrow, primarily because there is 3-fold more HSPC in mobilized PBSC compared to bone marrow grafts\textsuperscript{85-87}, that could potentially help in the faster recovery of transplanted patients. The collection of PBSC from the blood involves using an apheresis procedure (also termed leukapheresis) involving a cell separator device in which blood flows from an individual’s vein into a spinning chamber segregating cells by size and density, allowing continuous removal of a blood component of choice, for instance mononuclear cells including HSPC, and returning all other blood components (platelets, red blood cells, plasma) back into the patient’s blood stream via a second vein. In general, the process of autologous transplantation of genetically-modified HSPC involves: (1) HSPC mobilization into peripheral blood with G-CSF or other mobilizing agent(s), (2) collection of autologous mononuclear cells via apheresis, (3) ex-vivo HSPC enrichment via immunoabsorption of CD34+ cells, (4) \textit{ex vivo} transduction which involves culture of the CD34+ HSPC with a viral gene transfer vector and cytokines including FLT3, SCF, and thrombopoietin (TPO) to support HSPC survival with minimal cell differentiation and retention of engraftment potential, (5) conditioning of the individual to deplete remaining endogenous bone marrow cells and create “space”, (6) and finally infusion of transduced HSPC via intravenous delivery back into the individual.
The most commonly used conditioning regimens include either radiation or chemotherapy to clear away as many cells in the bone marrow as possible and make space for the new modified cells to engraft. The different forms of conditioning regimens available include high-dose (myeloablative), nonmyeloablative, and reduced-intensity conditioning (RIC). Myeloablative regimens consist of alkylating agents with or without total body irradiation (TBI) resulting in a near-total ablation of marrow hematopoiesis that lead to irreversible cytopenia in which patients require stem cell support. Nonmyeloablative agents do not require stem cell support and RIC regimens describe the regimens that fall outside the definition of myeloablative and nonmyeloablative regimens and also require HSC support. For fast and stable multi-lineage engraftment of blood cells following autologous transplantation, the optimal dose of CD34+ cells infused into a patient is > 5.0E+06 CD34+ cells/kg with the minimal dose being 2.0E+06 CD34+ cells/kg.

Retrovirus characteristics

The family of Retroviridae are subdivided into two sub-families; (1) Orthoretrovirinae which gives rise to six different Genus (alpha-retrovirus, beta-retrovirus, delta-retrovirus, epsilon-retrovirus, gamma-retrovirus, lentivirus) and (2) Spumaretrovirinae which gives rise to spumavirus genus. Common to all retroviruses is their classification as positive strand RNA viruses that reverse transcribe their RNA genome into DNA and have the unique ability to integrate into the host genome of target cells. Their ability to integrate into the genome of cells is what made these viruses appealing to be used as vector delivery systems for HSC, however, it was soon
discovered that transduction of HSCs with gamma retrovirus (γRV) is not straightforward and these viruses require target cells to be in active cell division in order to gain access to the nucleus. During cell division, the nuclear envelope disintegrates, and the virus takes advantage of this step in the cell cycle to access the genome for integration. Because HSCs are quiescent cells that infrequently undergo cell division, early HSC gene therapy trials resulted in very low transduction efficiencies when transducing with γRVs\(^{69, 70, 94-96}\). Improvements in transduction culture conditions, such as inclusion of cytokines such as thrombopoietin, stem cell factor and other factors able to maintain and cycle the most primitive HSC and use of Retronectin-coated plates (see below) resulted in success in 2 independent clinical trials carried out in the early 2000’s, targeting HSCs using γRVs for the treatment of immunodeficiencies\(^1, 2\). Unfortunately, although these gene therapy trials showed gene correction and clinical benefit, they were also linked with adverse events leading to vector genotoxicity (discussed further under gene therapy section).

Studies focused on improving viral transduction efficiencies were accelerated by the need to also find safer vector systems. One approach explored to improve the safety of γRV vectors was to remove a portion of the LTR including the enhancer/promoter region of the virus, making a self-inactivating vector (SIN) design (further discussed under vector safety). Another was to use a lentivirus (LV) instead of a γRV. Because LVs have a more favorable insertion profile than γRV and were more efficient at transducing quiescent HSC, attention focused on modifying the HIV lentivirus as a new vehicle to deliver transgenes for gene therapy, and in 1996, the first LV vector system able to stably transduce non-dividing cells in murine models was published\(^{68}\).
Although lentiviral vectors are capable of transducing quiescent cells, transduction efficiency of lentiviruses can be enhanced. One way to enhance transduction efficiency is by coating the surface of the tissue culture dishes used to culture and transduce HSPC with the compound retronectin (rFN-CH-296). Fibronectin is an extracellular matrix glycoprotein abundantly expressed in the bone marrow microenvironment that is important for HSPC adhesion via binding to integrins such as the very late antigen (VLA)-4 on HSPC. It also binds to lentiviral envelope protein and thus can facilitate the co-localization of HSPC and lentiviral particles, resulting in enhanced transduction efficiencies.\textsuperscript{97} Fibronectin consists of three functional domains: cell binding domain (C-domain), heparin-binding domain (H-domain), and CS-1 sequence. Retronectin is a recombinant chimeric compound consisting of these three functional domains. The C-domain and CS-1 sequence are important for binding to integrin receptors on target HSPC, specifically to the VLA integrins VLA-5 and VLA-4 integrin receptors respectively, and the H-domain on fibronectin is important for binding to LV particles.\textsuperscript{98-100} HSC highly express VLA-5 and VLA-4 integrin receptors and therefore bind to fibronectin via interaction with the C-domain and CS-1 sequence respectively. Another method to increase transduction efficiency include the use of protamine sulfate, which is an alkaline, polycationic, highly positively charged protein derived from fish sperm. Protamine sulfate is added during the transduction process and it is thought to counteract the repulsive electrostatic effects from the net-negative charge of the target cell and viral lipid membrane.\textsuperscript{101-103}

Lentiviruses have rapidly, almost completely, replaced \( \gamma \)RV as vectors of choice in gene therapies targeting HSC. The LV genus incudes various species including bovine
LV, equine LV, feline LV, colugo LV, weasel LV, rabbit LV, ovine/caprine LV, and primate LV. Primate LV species include the human immunodeficiency virus (HIV) and the simian immunodeficiency virus (SIV). HIV is categorized into two distinct viruses, type 1 (HIV-1) and type-2 (HIV-2), and these two viruses share similarities in their method of replication, infection, and advancement into acquired immune deficiency syndrome (AIDS). However, HIV-1 and HIV-2 differ in their viral loads, for example, HIV-2 is less pathogenic in humans and has a slower progression to AIDS, and geography, for example HIV-1 is found worldwide whereas HIV-2 is mostly confined to West Africa\textsuperscript{104}. Although HIV-2 vectors have been developed and studied in human CD34+ hematopoietic stem and progenitor cells (HSPC)\textsuperscript{105}, LV vectors used today for clinical applications in gene therapy are based on HIV-1 LV vectors.

The genome of a conventional LV is made up of two positive single stranded sense RNA’s containing a 5’ cap and a 3’ poly(A) tail that closely resembles cellular mRNA. The LV genome consists of 7-10 thousand nucleotides, is packed into a nucleocapsid core with a diameter of ~110 nm and is surrounded by a lipid membrane (Figure 2A)\textsuperscript{106}. The viral genome codes for three structural viral components, gag (group specific antigen), pol (polymerase), and env (envelope), all of which are necessary for the life cycle of the virus (Figure 2B). The gag gene codes for a polyprotein that gets cleaved by protease to make 3 core proteins; matrix, capsid, and nucleocapsid. Products of pol include protease, reverse transcriptase (RT), and integrase, and products of env include transmembrane (TM) and surface (SU) glycoproteins. The matrix protein is multifunctional with important roles in pre- and post- entry of the viral life cycle. The capsid, also referred to as the viral core, contains the pol proteins, RNA, and the
nucleocapsid which binds to the RNA molecules. Pol products such as protease, RT, and integrase are important for cleaving the viral polyproteins, transcribing the RNA into double strand DNA, and integrating the viral DNA into the cellular genome respectively. Envelope proteins, TM and SU, are essential for virus attachment and entry into target cells (Figure 2). The RNA genome is flanked by long terminal repeats (LTR) that contain untranslated regions (U3 and U5) and a short repeated (R) sequence which are important for viral replication, reverse transcription, and integration. Other elements important for viral replication include the primer binding site (PBS) located near the 5’ end of the genome, and the polypurine tract (PPT) located at the opposite end near the 3’end of the viral genome. The viral genome also codes for a psi (ψ) sequence located near the 5’ LTR important for dimerizing the genome and packaging virions during virus production. In contrast to γRV, LV are considered “complex” viruses and therefore require additional regulatory proteins such as transactivator of transcription (Tat) and regulator of expression of virion proteins (Rev) as well as accessory proteins such as virion infectivity factor (vif), viral protein R (vpr), viral protein U (vpu), and the negative factor (nef)\textsuperscript{107,108} in order to process RNA, stabilize viral structure, and other life cycle functions (Figure 2B).

Lentiviruses are desirable for gene transfer to HSC not only because they can stably integrate and express transferred genes long-term, but also for their unique ability to efficiently transduce non-dividing cells\textsuperscript{67,109}. Lentiviral vectors were especially desirable for their ability to transduce human CD34+ cells without the use of prolonged cytokine stimulation\textsuperscript{110-114} and retain normal engraftment ability long-term in immunodeficient mice\textsuperscript{112}. HSC are primarily in a quiescent state and rarely replicate,
dividing every 1-2 weeks in both murine and primate models\textsuperscript{64-66} which makes LV highly desirable for transducing HSC. Transducing HSPC using LV vectors is much simpler and faster compared to γRV-based vectors because unlike γRV-based vectors that require cells to be in a mitotic state and therefore require cells to be in culture for days with various cytokines prior to transduction, LV transduction can be performed with only a total of 24-48 hours of culture. Shortening the days of culture and ex-vivo manipulation of HSC is critical to minimize HSC differentiation and retain engraftment ability.

Replication and reverse transcription

Replication and reverse transcription of LV is a complex process that begins before the virus even enters the cell. The first step for the initiation of viral entry is the attachment of viral envelope glycoproteins to specific receptors on the membrane of the target cell. Second, the viral envelope glycoproteins fuse with the cell membrane in a way that allows internalization of the viral core into the cytoplasm of the target cell. Once inside the cytoplasm, the capsid proteins are removed and the viral core breaks down, exposing the RNA genome to host proteins that initiate reverse transcription into double stranded DNA (Figure 3). The process of reverse transcription is elaborate and has taken years to thoroughly understand the mechanism of how viral RNA genome is copied to DNA, instead of the canonical reverse normal cellular process, requiring polymerase to undergo strand-transfer twice, a process also referred to as “jumps”. Lentiviral replication is initiated with the binding of host-derived transfer RNA (tRNA) to the viral PBS site near the 5’ LTR, followed by the annealing of the virus encoded RT to the tRNA/RNA complex. The reverse transcriptase enzyme reverse transcribes the U5
and R region on the 5’ LTR and when the R region is finished being transcribed, polymerase undergoes the first “jump” in which the newly produced negative strand DNA R region from the 5’LTR complexes to the R region of the RNA on the 3’LTR via homology complementation. This DNA/RNA complex serves as a primer for RT to reverse transcribe the RNA from the 3’ LTR to the PBS site on the RNA near the 5’ LTR to synthesize a negative strand DNA. Positive strand synthesis begins with the degradation of RNA/DNA complexes facilitated by an RT enzyme, RNase H, with the exception of the PPT as it is resistant to RNase H degradation activity. The PPT then serves as a primer for RT to transcribe to the 19th position of the tRNA capturing the PBS site necessary for the second “jump” by base pairing to the PBS site on the negative strand DNA and allowing RT to transcribe all the way to the U5 region of the 3’LTR synthesizing a positive strand DNA, referred to proviral DNA. The resulting proviral DNA is a longer sequence than the initial RNA genome, containing two full-length LTRs, and is capable of integration\textsuperscript{115}. The resultant blunt-ended proviral DNA is shuttled to the nucleus with the preintegration complex (PIC) assisting in translocation through the nuclear pore\textsuperscript{116} and into the nucleus. Once in the nucleus, IN recognizes sequences at both ends of the viral DNA, termed attachment sites or “att” sites which are short inverted repeats, and removes two to three terminal nucleotides from the 3’ end through a process termed 3’-processing, to create phosphodiester reactive 3’-OH ends required for attachment to the chromosome. Integrase then guides the proviral DNA to the host genome and catalyzes its integration through a recombination process \textsuperscript{117, 118}.

Integration of LV is not a random process, instead, integration is biased towards transcriptionally active regions or certain structures within the chromatin, for example,
regions in the DNA which are bent are preferred integration sites. The LTRs play an important role in guiding retroviral integration through protein-protein interactions and this might depend on the transcription factors that bind the LTR. Host cell factors such as LEDGF/p75 have been assigned major roles in directing the target-site of viral integration\textsuperscript{119-121}. LEDGF/p75 is a host cell factor that directly binds to integrase and tethers the preintegration complex to the genome of actively transcribed genes\textsuperscript{122,123}.

Unlike γRV that target transcriptionally active promoters or regulatory regions\textsuperscript{124-126}, LV target gene dense regions \textsuperscript{124,126,127} and this difference has made LV a more favorable vector choice. Because LV integrate away from transcriptional start sites, they are thought to be less likely to upregulate genes near the integration site and reduce the chances of insertional mutagenesis. However, integration of LV into the transcriptional unit of genes come with consequences that can result in genotoxic effects. For example, LV integration into genes can induce alternative splicing resulting in the generation of aberrant transcripts and post-transcriptional deregulation\textsuperscript{9,128-130}. The SIN vector design also comes with some caveats in which removal of a portion of the U3 has been linked to lentiviral transcriptional readthrough activity that results in the generation of chimeric transcripts\textsuperscript{131}.

Chimeric HIV1-based lentiviral vector system design

The goal of LV vectors is to stably deliver genetic material into cells that are either missing or have a non-functional form of a gene. All retroviruses used for clinical purposes in gene therapy are modified to be replication-defective retroviruses (RDR), unable to produce virus following transduction and integration. A transgene replaces a
large portion of the viral genome (also discussed in Safety section). For instance, the sequences that code for the structural genes (gag, pol and env) are removed, thus the vector is unable to synthesize the necessary viral proteins for further viral replication, assembly and release once it has integrated into the target cell. This prevents productive viral infection and uncontrolled spread to non-target cells in the organism or patient. Furthermore, a series of additional modifications have been made in RDR-LVs, removing additional coding and viral gene expression control sequences to further decrease the chance of viral replication and recombination with endogenous human retroviruses.

If much of the viral genome has been removed from RDR-LV vectors, how are viral particles able to transduce HSC? The proteins necessary to assemble infectious LV particles (GagPol, env, Rev-Tet) must be provided in trans, generally by delivery of genes for each on a separate plasmid via transfection of viral producer cells along with a vector plasmid, and this is further discussed under the safety section (Figure 4). Once a target cell is infected, the vector particle uses the trans-supplied proteins to undergo reverse transcription and integrate. Those viral proteins then break down and are not replaced, given lack of genes encoding them in the integrated LV vector. There is no productive HIV infection and thus no damage to the individual’s immune system.

The concept of RDR vectors date back to the early 1980’s, the first time viral sequences were removed from the vector to create the very first RDR retroviral vector \(^{132}\). Each of the LV vectors used in this study are based on RDR vectors with additional genetic modification (further discussed below) to improve transduction efficiencies in rhesus macaque HSC.
Wildtype HIV exclusively infects human immune cells, specifically CD4+ T lymphocytes, macrophages, and dendritic cells, based on viral entry via HIV env binding to CD4 and other specific receptors. However, viral vectors can be genetically modified to broaden or alter tropism (host range of the virus) via substitution of alternative env protein from a different virus, termed pseudotyping. There are several envelope proteins used to pseudotype vectors, but the vectors used in this study and almost all LV clinical trials are pseudotyped with the amphotropic vesicular stomatitis virus glycoprotein (VSV-G), meaning, the envelope protein of HIV-1 virus was replaced with the envelope protein from the vesicular stomatitis virus. Pseudotyping with VSV-G allows the vector to infect a larger range of cell types, including HSPC, and provides the additional benefit of stabilizing the vector to withstand high-speed centrifugation, because the VSV-G glycoprotein pseudotyped particles are much more sturdy than conventional HIV-1 particles\textsuperscript{133,134}.

Although HIV-1 virus is capable of infecting both human and rhesus macaque cells, infection of rhesus macaque HSPC with standard HIV-1 LV has been challenging due to innate immune factors in macaque cells that block standard HIV infection. Innate immune factors including TRIM proteins, particularly the tripartite motif-containing 5 isoform-\(\alpha\) (TRIM5\(\alpha\)) has been shown to inhibit HIV-1 infection of macaque cells. In humans, the virus escapes TRIM5\(\alpha\) attack by binding to endogenous human cyclophilin A (CypA) protein via the capsid. However, in rhesus, this pathway does not protect HIV capsid from TRIM5\(\alpha\) attack\textsuperscript{135}.

To enhance LV transduction of rhesus macaque cells, rhesus macaque studies, Uchida et. al. constructed a chimeric HIV-1 lentiviral vector (\(\chi\)HIV) exchanging the HIV-
1 capsid for the simian immunodeficiency virus (SIV) capsid. SIV capsid binding to rhesus macaque CypA evade TRIM5α degradation, and transduction efficiency in rhesus CD34+ with χHIV was significantly higher compared to transduction efficiencies using conventional HIV-1 LV\textsuperscript{136}.

**Gene therapy**

The first successful bone marrow transplant was performed 50 years ago in 1968 on a 5-month-old male patient diagnosed with X-linked severe combined immunodeficiency (SCID-X1). SCID-X1 is an immune deficiency disorder characterized by mutations in the gene encoding the interleukin-2 receptor gamma chain, a common subunit that is a required component of 6 different interleukin receptors, including interleukin-2, interleukin-4, interleukin-7, interleukin-9, interleukin-15, and interleukin-21. Mutations in the common gamma chain affect the growth and maturation of various lymphocytes, ultimately resulting in the absence of T and natural killer (NK) cells, and impairment of B cell function\textsuperscript{137,138}. The donor bone marrow cells for the first successful transplant came from the patient’s 8-year-old sister whose cells were histocompatible at the human leukocyte antigen (HLA) locus and reconstitution of both cellular and humoral immunity was achieved without graft versus host disease (GVHD)\textsuperscript{139}. Although HLA-identical donor bone marrow transplants provide a greater than 90% chance of long-term survival, unfortunately, the majority of patients lack an HLA-identical donor, making allogeneic bone marrow transplants more challenging, forcing reliance on mismatched or haplotype-matched donors, often resulting in GVHD\textsuperscript{140,141}. For this reason, gene therapy
treatments using the patient’s own HSC (autologous HSC) is an appealing approach because it provides an alternative for patients who lack HLA-identical donors.

Gene therapy describes a process introducing genetic material into cells to treat human disease. This can involve “in vivo” therapy, with direct delivery of the vector into the patient, for instance injection of a gene therapy vector into the eye to treat congenital blindness, or “ex vivo” gene therapy, with exposure of cells such as HSC to vector outside the body, followed by transplantation of treated autologous cells back into the patient. The concept of gene therapy dates back more than a half century ago when the ethical issues related to gene therapy were discussed even though the technology was not yet close to clinical application\textsuperscript{142}.

The conception and engineering of the first replication-incompetent viral vectors in the mid 1980’s set the stage for the development of gene therapies by showing efficient gene transfer into mammalian cells\textsuperscript{143} and then into murine HSC\textsuperscript{144-148}. The ability of integrated proviral forms of retroviruses to stably express genes, theoretically for the life of the organism, was the unique feature that attracted interest in these viruses and ultimately resulted in γRV becoming the first class of viruses to transfer a foreign gene into patients\textsuperscript{149}. The first patient in the very first gene transfer study used a γRV to introduce a DNA sequence for a gene coding the neomycin resistant gene. The trial involved using a γRV, moloney murine leukemia virus (MMLV), as the gene transfer vehicle into tumor-infiltrating lymphocytes (TIL) with the purpose of tracking the distribution and survival of TILs in a human immunotherapy clinical trial\textsuperscript{149}.

The detection of genetically-modified T cells in patients in this pioneering trial carried out by Stephen Rosenberg at the National Cancer Institute encouraged
approaching treatment of adenosine deaminase deficient SCID (ADA-SCID), showing encouraging data \textit{in vitro} using human ADA(-) cells\textsuperscript{150} and \textit{in vivo} in animal models\textsuperscript{151, 152}. The first human gene therapy clinical trial began using $\gamma$RV-transduced T cells expressing the ADA gene to treat 2 children who suffered from the disease\textsuperscript{94}. The decision to use T lymphocytes over bone marrow HSC for the trial was supported by a study in nonhuman primates which suggested T lymphocytes can be used as gene therapy vehicles\textsuperscript{152} as they can be easily transduced and stably express the therapeutic gene, and transduction efficiencies were better with T lymphocytes than with HSC at that time. While transduced T cells persisted over time, there was little evidence for therapeutic benefit.

Unfortunately, transduction of HSC using $\gamma$RV was proven to be a challenging task that consistently resulted in low transduction efficiencies due to the quiescent nature of HSC. Gene transfer via $\gamma$RV integration rely on the target cell to undergo mitosis for the virus to enter the cell nucleus and access to the cell genome as detailed above\textsuperscript{67, 153}. In contrast to HSC, T cells can be easily transduced with $\gamma$RV because they are not quiescent cells. Although various groups studied methods to improve transduction efficiencies, no significant breakthroughs in improvement were made until a human study was initiated that involved autologous umbilical cord CD34$^+$ cells that used a multi-cytokine stimulation cocktail to induce HSC cycling in three ADA-SCID patients. Unfortunately, even though cytokine stimulation seemed encouraging and the ADA gene was detected in peripheral blood lymphocytes of treated patients, the levels were too low to sustain clinical benefit\textsuperscript{95, 96}. Other trials, for example, in chronic granulomatous disease\textsuperscript{70} and Gaucher disease\textsuperscript{69}, followed, but showed no clinical benefit due to low
transduction efficiencies. In addition, these early trials used no conditioning prior to cell infusion, with lack of engraftment with transduced cells likely due to insufficient niche “space”.

After this initial series of disappointing HSC gene therapy trials in the early 1990s, and a serious adverse event in an in vivo gene therapy trial resulting in a patient death following administration of an adenovirus to treat a form of metabolic liver disease, the field took a step back. A 1996 NIH blue ribbon report strongly suggested that the science was not yet ready for clinical applications, and that better understanding of target cells, appropriate human diseases and improved vectors were necessary. Intense work in animal models and in the design and optimization of new transduction conditions and vectors occurred over the next decade. Encouraging in vitro as well as in vivo studies using gamma chain knockout mice and canine models demonstrated that delivery of the gamma chain gene using an integrating γRV could successfully restore gamma chain expression to normal levels even without conditioning therapy, providing support for clinical trials in the treatment of this form of SCID patients. After decades of transducing CD34+ cells with γRV and obtaining negative results due to poor HSC transduction efficiencies, finally these obstacles were overcome by advances in CD34+ culture methods. Studies from the Dunbar lab were the first to show in a relevant human model that these new transduction protocols that included a combination of cytokines and Retronectin during transduction, resulted in high-level long-term HSC gene transfer with stable and persisting marking of 10-15% in multiple lineages in rhesus macaque.
Overcoming HSC transduction efficiency obstacles resulted in the very first report demonstrating unequivocal clinical benefit following any sort of gene therapy in two trials in Europe for X-SCID\textsuperscript{2} and ADA-SCID\textsuperscript{1}. The X-SCID trials led by Cavazzana-Calvo and Alain Fischer began in and enrolled 10 infants and young boys who were infused with autologous CD34+ HSPC transduced with a γRV carrying a functional gene for the common gamma chain. Following gene therapy, 9 of the 10 patients treated showed immune reconstitution, with circulating T cells and NK cells for the first time in their lives\textsuperscript{2}. The success from this trial led to the approval of a second gene therapy clinical trial in the UK for the treatment of another 10 SCID-X1 patients\textsuperscript{162}.

Unfortunately, 3-6 years following the start of the French trial, 4 of the 9 patients treated developed leukemia due to insertional mutagenesis, with integration of the vector near the \textit{LMO2} proto-oncogene resulting in upregulation and eventual transformation to T cell acute lymphoblastic leukemia\textsuperscript{4}. Shortly after leukemias were found in the French human trial, leukemias were also found in murine animal models using the same γRV vectors from the human trials\textsuperscript{163,164}. The London human trial for the treatment of SCID-X1 used a similar vector as the French trial\textsuperscript{162,165}. Their 2 year follow-up reported on clinical improvement and although the risk of mutagenesis was a big concern because the vector used had a similar configuration to the French trial they did not observe signs of genotoxicity at that time\textsuperscript{162}. It wasn’t until their longer time point follow-up studies where they observed development of leukemia in treated patients\textsuperscript{5,165}. By 3-5 years following gene therapy treatment in both the French and English trials using a second generation γRV vector that expressed the common gamma chain from a strong viral promoter/enhancer, a total of 5 of the 20 treated patients were found to develop T cell
leukemia due to insertional mutagenesis that resulted in the activation of proto-
oncogenes, emphasizing the high risk of gene therapy using γRV at least in X-SCID 3-5, 166. Although at the beginning these gene therapy studies showed gene correction using γRV vectors1, 2, 162, 165, 167-172 at the same time these trials resulted in adverse events that were unacceptable4, 5, 173. 

Other gene therapy trials resulting in genotoxicity include those for chronic granulomatous disease (CGD)174 and Wiskott Aldrich syndrome (WAS)175. CGD is a rare inherited primary immunodeficiency that results in recurrent life-threatening infections due to a defect in the nicotinamide adenine dinucleotide (NADPH) oxidase complex176-178. The NADPH complex is necessary to generate superoxide anion to clear bacterial and fungal infections, without it, phagocytes have impaired antimicrobial activity. Most CGD cases are due to a defect in the gene encoding gp91phox and thus gene therapy targeting HSC to modify them to express a functional gp91phox gene was thought to provide a curative therapy since it had already been demonstrated to be successful in murine models179-181. A CGD human trial thus began for the treatment of two patients with a γRV vector that used the strong SFFV enhancer in the LTRs to drive the gp91phox transgene174. The trial although initially resulted in clinical benefit based on microbial killing and ability of patients to clear infections, the end result of this trial was clonal dominance and myeloid leukemia due to activation of proto-oncogenes7. WAS is another primary immunodeficiency due to mutations in the WAS gene that results in deficiency of the WAS protein. Gene therapy to correct the defect was thought to be effective and a potential therapeutic treatment. The first WAS human trial used a γRV and the transgene coding for the WAS gene was LTR-driven. Although this trial demonstrated clinical
benefit, it was unfortunately also associated with proto-oncogene activation that led to insertional mutagenesis and leukemia in 7 of the 9 evaluable patients\textsuperscript{175}.

Improving vector safety became the focus of attention. One of the first approaches to attempt to create safer vectors was a SIN vector design, which deletes a portion of the U3 in the LTR region to remove the enhancer region, given that enhancer activation of nearby proto-oncogenes was implicated in vector genotoxicity and adverse events. These vectors were harder to produce at high titer, but eventually \textit{in vitro} studies demonstrated vector efficacy\textsuperscript{182-184} and a clinical trial to treat SCID patients soon followed becoming the first SIN-\(\gamma\)RV vector used in the treatment of patients\textsuperscript{185}.

The first clinical trial conducted using a SIN-\(\gamma\)RV design used an internal human EF1a short promoter inserted within the vector to drive the IL-2 receptor gamma chain instead of the strong viral promoter/enhancer in the LTR. The trial treated 9 boys diagnosed with SCID-X1 and was run in parallel with phase 1/2 trials conducted in Paris (5 patients \# NCT01410019), London (0 patients \# NCT01175239), and the US (4 patients \# NCT01129544)\textsuperscript{185}. When comparing the patients from prior trials who were treated with the MFG-gamma chain vector to patients treated with the SIN-\(\gamma\)RV design, there was no difference in the kinetics of T cell reconstitution\textsuperscript{185} and so far there have been no genotoxic adverse events reported in these patients.

As more clinical trials are being approved involving gene therapies using LV vectors to treat various diseases, further studies are needed to assure the safety of these vectors. Gene therapies utilizing LV vectors targeting HSPC include those for the treatment of SCID-X1, ADA-SCID, Wiskott Aldrich syndrome, adrenoleukodystrophy, metachromatic leukodystrophy, human immunodeficiency, and beta-thalassemia\textsuperscript{186}. 

28
Safety of integrating vectors

Efforts to improve safety

Initial efforts in gene therapy were done using γRV and one of the critical and first safety features introduced was the conversion of replication competent retroviruses (RCR) into replication defective retroviruses (RDR). The RDR modification maintained the ability of the virus to infect target cells but took away its ability to replicate and still allowed for high-titers of virus particles to be produced using stable packaging cell lines (or also termed helper cell lines, as detailed above)\(^{187, 188}\). The purpose for the RDR vector design is the removal of the essential viral proteins from the viral vector to prevent the virus from replicating on its own. The essential viral proteins deleted from the viral genome include gag, pol, and envelope, without these proteins, the virus is unable to produce new virus particles. Theoretically because RDR vectors cannot replicate, RDR would prevent the uncontrolled spread of the virus within the treated host as well as the spread of infection between treated patients and healthy individuals. The goal of RDR vectors is for each viral particle to only be capable of a single integration event, without leading to productive viral infection, and three different vector designs were ultimately made; first, second, and third generation retroviral vectors with each generation having improved vector characteristics.

First generation γRV RDR vector design was based on using packaging cell lines that contained the viral genome encoding for the essential viral proteins necessary for the viral life cycle with the psi packaging sequence deleted to prevent viral proteins from being encapsulated into virions\(^{132, 189, 190}\). To these packaging cell lines, the vector plasmid containing the transgene of interest was introduced, usually by transient
transfection, to produce viral particles containing only the vector DNA encoding the transgene without any other viral proteins encapsulated in the virion. Unfortunately, these first generation RDR design was unsuccessful in preventing eventual production of replication-competent virus from producer cell lines, due to recombination between the helper genome and sequences from the vector genome, resulting in the production of RCR viral particles\textsuperscript{191,192} impossible to use clinically due to the safety risks of RCR viruses.

A second generation $\gamma$RV vector packaging cell system was then constructed which still included the deletion of the psi sequence, but also had additional mutations to decrease the risk of recombination with helper genomes and production of RCR\textsuperscript{188}. Finally, a third generation $\gamma$RV packaging cell system was developed to further increase the safety and reduce the possibility of recombination that would result in RCR. It involved splitting the expression cassettes encoding the essential viral proteins gag, pol and env onto at least two different plasmids, one encoding the ecotropic gag/pol and the other encoding the amphotropic envelope\textsuperscript{187,193,194}. Both plasmids had the psi sequence deleted to prevent encapsulation into virions. Because ecotropic viruses cannot infect human cells, their use was thought to be safer and because amphotropic viruses are capable of infecting most mammalian cells, this was a good characteristic for pseudotyping the virus.

Over 20 years of vector research conducted world-wide to design safer versions of integrating vectors have passed shedding light on the various hypothesis of how $\gamma$RV resulted in insertional mutagenesis. After revealing that genotoxic effects of $\gamma$RV-based vectors were a result of their insertion profile targeting transcriptional start sites\textsuperscript{195-198}
one of the initial methods to improve the safety of vectors was to replace MLV based γRV vectors with LV vectors\textsuperscript{68}. This was based on two main reasons, one, LV tend to have a more favorable integration profile than γRV and two, LV are capable of transducing non-dividing cells, a big attraction for trying to provide long-term expression in nonproliferating cells such as HSC. In terms of integration profile of these vectors, γRV have a preference to target transcriptional start sites, and LV tend to favor the transcriptional unit itself resulting in integration within genes\textsuperscript{197-200}.

Three different LV systems were made after moving away from γRV; first, second, and third generation, with each generation increasing in their safety features. All three generation vectors vary in terms of their sequence modifications, but what they all have in common are sequences in the transfer vector which are absolutely necessary for making effective viral particles and those include sequences that code for the RNA packaging signal, the major splice donor, and the REV response element (RRE)\textsuperscript{201-204}, these sequences are always part of the viral vector that get incorporated in the genome after integration. Many deletions of viral sequences in the vector have been made, but some of those sequences deleted from the transfer vector are important for virus production and are therefore must be provided in trans and these include the essential proteins (gag, pol, env) and regulatory proteins (tat, rev). These genes delivered in trans do not get integrated into the genome because they are not part of the transfer vector and the plasmids carrying the genes lack a packaging signal.

The first generation LV vector design was based on a three expression plasmid system that included (1) the packaging plasmid containing the essential viral proteins gag, pol, rev, and tat (2) a plasmid carrying the envelope gene coding for a pseudotyped
envelope protein that gives the vector a wider tropism, for example, the virus is not restricted to CD4 expressing cells and therefore increases the range of cells it can infect, and (3) the transfer vector that contains the transgene of interest. The psi sequence is absent from both helper plasmids and only the transfer plasmid containing the transgene contains the psi sequence so at the end of the viral life cycle, only the vector viral RNA is packaged into the virus and not RNA produced from the helper plasmids. After further studies to try to improve the HIV-1 vector system, a second-generation LV vector was made that included additional modifications to the vector. The additional modifications of the second generation LV vector design were thought to further increase the safety of the vector and consisted of deleting the accessory genes that coded for vpr, vif, vpu, and nef after it was found that they were not required in the viral replication process using immortalized cell lines. Current third generation vector system is based on a four-plasmid expression system that includes (1) the packaging plasmid containing gag and pol (2) regulatory plasmid containing only rev (3) plasmid coding for the envelope gene, and (4) the vector plasmid containing the transgene of interest (Figure 4). Additional modifications to the third generation LV transfer vector systems include deletions in the 3’ LTR to remove the enhancer region from the LTR and result in a SIN vector design and also the removal of the regulatory gene tat. Because the design of the SIN-LV vector is devoid of its enhancer region, internal promoters are added to the vector to drive the expression of the gene of interest.

To put the deletions made in LV vectors into context, the complete HIV-1 genome is made up of 9.7kb, and after removing the essential viral components, including the enhancer region from the vector, the resulting vector sequences encode for
1.7kb (17.5%) of the complete HIV genome. These deletions not only make the vector safer, but also allows the ability for larger sized transgenes, for example, transgenes of approximately 8.8kb in size can be packaged\textsuperscript{207, 209}. In addition, removal of the majority of HIV vector sequences decreases the chance of recombination with wildtype HIV vector in patients, should the patient happen to later get infected with HIV. More recently, a fourth-generation vector system has been proposed containing additional deletions to further reduce the amount of the HIV genome that gets transferred to patient cells\textsuperscript{210}. The additional deletions in this new vector design results in just 441bp (4.8%) of the complete genome in the transduced product and has shown to be effective in a hemophilia B murine model \textsuperscript{210}.

Although multiple modifications have been made to LV vectors, the risk of genotoxicity from the integration into the genome still remain. It is established that the genotoxic risk of LV vectors that result in gene disruption/inactivation is much lower compared to the risk of gene activation by LTR-driven γRV (see below). However, to further improve the safety of LV, additional modifications to the vector were made. One such modification targets viral proteins such as integrase where mutations have been introduced to create integrase defective lentiviruses (IDL)\textsuperscript{211} and avoid integration altogether. However, IDL vectors are not of any utility for transduction of cell targets such as HSC that must undergo cell division as part of their life cycle, since nonintegrated virus gets diluted with each cell division step. Fusion proteins added to vectors to enable site-specific integrations\textsuperscript{212-215} are other modification made to allow for more desirable integration profile, for example targeting integration at safe harbor locus
where the risk of affecting the host genome is minimized, however these approaches to date are inefficient.

The most significant modification to improve the safety of integrating lentiviruses is still the deletion of the enhancer region in the U3 portion of the LTR. These enhancer deprived vectors were termed SIN. Because LTRs were thought to upregulate nearby or adjacent genes from their site of integration, it was hypothesized that by removing the LTR enhancers and replacing it with an internal promoter located within the vector, the chances of nearby genes being upregulated would be reduced. This first SIN configurations were designed based on MLV γRV vectors and these studies demonstrated the difficulties of making deletions within the LTR\textsuperscript{216-219}. Soon after LV were found to be a better vector choice, the first LV-SIN configuration was made by the Naldini and Trono groups as an efficient in vivo gene delivery system\textsuperscript{207}. This SIN-LV design was hypothesized to have a safer profile because the internal promoter would be distant enough from genes reducing the risk of insertion genotoxicities.

Following the first publication of SIN LV vectors, comparison of different SIN configurations was studied particularly comparing γRV- and LV-based vectors. One of the first studies compared SIN LV with the relatively weak internal promoter human phosphoglycerate kinase promoter (PGK) to LTR-driven γRV consisting of the strong enhancer promoters in tumor-prone mice. In this study they found that LTR-driven γRV accelerated tumor formation in contrast to SIN LV with weak promoters\textsuperscript{220}. Other studies have also shown that SIN γRV profile is less genotoxic compared to its LTR-driven counterpart and SIN LV is even less genotoxic than SIN γRV.
In vitro studies using SIN γRV containing the very strong internal spleen focus-forming virus (SFFV) promoters driving reporter genes showed that SIN configuration was less likely to immortalize murine HSPCs compared to LTR-driven vectors\textsuperscript{17}. Furthermore, in vitro assay using SIN γRV containing the intron-less human elongation factor 1 alpha (EFS) promoter showed it was not able to induce immortalization and was also found to be significantly less mutagenic compared to identical SIN vectors that had internal SFFV promoters. These in vitro studies suggest that the combination of using a SIN design where the vectors are deprived of their LTR enhancers as well as using less potent promoters such as endogenous promoters (e.g. EFS) together could be a safer vector design to lower the risk of genotoxicity from integrating vectors.

Other methods attempted to reduce the risk of genotoxicity include the addition of insulators in the LTRs of viral vectors. Insulator core elements are sequences of DNA that are thought to have enhancer-blocking activity when inserted between an enhancer and nearby genes, and were initially thought to be useful in preventing activation of nearby genes. It was thought that duplicating insulator core elements might be more effective than having a single insulator element\textsuperscript{221}, but tandem repeats of insulator elements has been associated with lower vector titers\textsuperscript{184} and studies have shown that the addition of the 250bp chicken HS4 locus insulator does not add additional safety features as no significant differences were observed with or without insulator elements. Tissue specific promoters/enhancers have opened additional features to reduce the risk of genotoxicity\textsuperscript{184}. However, despite all the efforts to improve the safety profile of lentiviruses, there still remains concern for their genotoxicity. Strong viral promoter/enhancers are still utilized in LV vectors, including those in some clinical trials,
when strong constitutive expression of a transgene is required to treat serious diseases, such as adrenoleukodystrophy. The risk of LV genotoxicity was felt to be low enough to take this risk.

*Risks associated with integrating vectors*

Insertional mutagenesis is defined as the deregulation of genes, in particular, proto-oncogenes, that are located either near or at the site of viral integration and results in genotoxicity, or a disease phenotype linked to dysregulation of a gene or genetic element by the integration event. Thus far, trials using SIN-γRV or SIN-LV vectors have not yet resulted in clinically-apparent genotoxicity. It is encouraging that even the trials that used SIN-γRV vectors coding for the gamma chain have not yet shown adverse events in patients at the last time follow-up, tests for RCR are negative, and SIN-γRV efficacy has shown to be similar to γRV MFG-gamma chain vectors. Although only time will tell, these results are already reassuring because genotoxicity in patients from trials that used the γRV MFG vectors developed leukemia within 2-5.5 year post gene therapy.

Other major concerns from the integration of these vectors, in addition to the risk of activating nearby genes, particularly proto-oncogenes, is the risk of inactivating tumor suppressor genes. Prior to SIN vectors, activation of proto-oncogenes, also termed “promoter insertion” was known to be the most common method of insertional mutagenesis. This occurred as a result from LTR-driven vectors with strong enhancers/promoters that interacted via long distance interactions and triggered distant enhancers of endogenous genes. With the combination of the SIN vector design
where the vector is devoid of the LTR enhancer regions and the use of moderately active internal cellular promoters, the risk of genotoxicity due to promoter insertion has been significantly reduced. Thus far, no human genotoxicities using SIN vectors, whether SIN- γRV or SIN-LV have been reported. However, the risk of mutagenesis due to insertion gene inactivation by SIN-LV still exists. Studies using tumor prone murine models have shown that SIN-LV with a moderate PGK promoter inactivates Pten and Rasa1 tumor suppressor genes which coincides with their prior finding that these same vectors compared to mock treated animals, showed an accelerated tumor onset and although it was only a small increase, the difference was significant\textsuperscript{14}.

Therefore, it was startling and concerning when clonal expansion was noted for the first time in an LV clinical trial in patient with beta-thalassemia\textsuperscript{9}. This patient never developed leukemia nor abnormal hematopoiesis, but a single clone massively expanded and was found to have aberrant splicing between vector and endogenous mRNA sequences, resulting in overexpression of the nearby HMGA2 gene. In murine models, overexpression of this gene results in clonal expansion\textsuperscript{225}.

Aberrant splicing results in chimeric transcripts which are mRNAs that consist of both cellular and vector sequences. Chimeric transcripts are a result of read-through transcription in which transcription from the vector itself doesn’t stop and instead, continues to transcribe the flanking cellular genes resulting in a chimeric transcript. Alternatively, a cellular gene may continue transcribing through the vector sequence and also result in a chimeric transcript with cellular and vector sequences\textsuperscript{9, 220, 226}. Studies in murine models using lentiviral vectors have shown aberrant transcripts as a result of LV integrations. For example, an insertion site obtained from a myeloid tumor in a tumor-
prone mouse model, the LV vector with the strong enhancer SFFV in the LTR integrated within intron 11 of Braf, resulting in a chimeric LV-Braf transcript that gave rise to a truncated Braf protein that lacked its regulatory domain\textsuperscript{15}. In a recent human CAR-T immunotherapy trial, the LV integrated in intron 9 of the TET2 gene and resulted in chimeric RNAs and a partial loss of TET2 function\textsuperscript{227}. TET2 deficient T cells had enhanced survival, but did not transform to leukemia. Leaky polyadenylation signals from the 3’LTR have also been suggested to be the cause for chimeric transcripts. One of the ways to address leaky vectors and lower the frequencies of read-through transcription is by including a strong polyadenylation signals in the LTR of the vector\textsuperscript{228}. Unfortunately, it seems LV vectors are more prone to read-through transcription since they tend to target the actual transcriptional unit of genes. Although there have been no clinically-serious events linked to LV genotoxicity to date, longer-term follow-up analysis will be critical for assessing genotoxicity of these integrated vectors, particularly before utilization in non-lethal diseases.

\textit{Methodologies for assessing vector genotoxicity and gaps}

Various pre-clinical methods are available for testing the safety of gene therapy vectors, but they all have limitations. Most assays to assess genotoxicity are based on in vitro assays or murine models, though previous detection methods for vector genotoxicity using murine models has been hindered not only due to the requirement of large numbers of mice, but also in the requirement to develop severe disease in experimental animals in order to track clonal dominance. However, one particular murine model that is frequently used is the tumor prone murine model in which mice develop tumors at a
predictable time of onset. One favored tumor-prone murine strain is the Cdkn2a/- mouse model which is deficient in two proteins, p16\(^{\text{ink4a}}\) and p19\(^{\text{Arf}}\). Protein p16\(^{\text{ink4a}}\) is a kinase inhibitor (inhibits CDK4-6/D kinases that mediate Rb1 phosphorylation) and the protein p19\(^{\text{Arf}}\) that activates the tumor suppressor p53 via a mechanism that involves binding to Mdm2, a p53 negative regulator, to protect p53 from degradation. Deficiency of p16\(^{\text{ink4a}}\) and p19\(^{\text{Arf}}\) in these knock out mice result in early onset of various types of tumors and a major benefit is tumor onset can be predicted, and the impact of an intervention on tumor latency, such as vector transduction of HSC, can be measured. These models have been used primarily to study insertional mutagenesis by linking genes to tumors, but unfortunately the genes linked to tumors in these models do not match the responsible loci in human trials decreasing their preclinical utility. This is not surprising given the marked differences in the rates and types of hematologic tumors encountered in mice versus humans, and the fact that humans do not have mutations in the predisposing genes mutated in the tumor-prone murine models.

Various in vitro assays have been used assess genotoxicity, and although they have been helpful in identifying major differences in genotoxic risk, all involve murine cells, and none are representative of the complex in vivo cellular environment and heterogeneous HSPC target cells. The most commonly-used is an in vitro immortalization assay exposing murine HPSC to high concentrations of vector, then plating cells at close to limit dilution and looking for immortalized clones. This method is reasonably rapid, but only semi-quantitative, and it captures myeloid but not lymphoid transformation, and identifies only insertions activating MDS1 and EVI1 complex (MECOM) pathways. Therefore, a more appropriate animal model to test new vector
designs before moving into human trials is highly desirable. For example, in the case of CGD, gene correction was obtained in murine models \textsuperscript{179-181} but unfortunately in humans correction was suboptimal and genotoxicity with eventual leukemia was observed\textsuperscript{69, 70}. Although improvements with the viral vectors have been made, for example, the issues of low transduction efficiencies obtained with $\gamma$RV have been addressed by switching to the LV vector system and the safety of these vectors has been improved, but a model to predict vector safety in humans has not been developed and safety concerns with integrating vectors remain. An animal model that is more predictive of human outcomes, such as a non-human primate, is attractive to develop for prediction of genotoxicity.
Figures

Figure 1. HSC hierarchical model.
Hematopoietic stem cells (HSC) are located at the very top of the hierarchy. HSC are capable of self-renewal and are responsible for giving rise to all cell types in the blood. HSC differentiate through a series of progenitor steps before committing to a specific cell lineage and maturing into a functional immune cell. The various branching from
multipotent progenitors is continuing to unravel. Shown is a simplistic hierarchical model that combines the conventional hierarchical model with the more current models. MPP, multipotent progenitor; CMP, common myeloid progenitor; NKP, natural killer cell progenitor; LP, lymphocyte progenitor; MEP, megakaryocyte/erythrocyte progenitor; GMP, granulocyte/macrophage progenitor.
Figure 2. Schematic of the structure of a retrovirus particle.
Retrovirus particle (A) and retrovirus genome (B) demonstrating the 3 structural genes (1) gag (group specific antigen) which codes for matrix, capsid, and nucleocapsid, (2) pol (polymerase) which codes for protease, reverse transcriptase, and integrase (3) env (envelope) which codes for surface (gp120) and transmembrane (gp41) proteins. Complex retroviruses such as Lentiviruses code for additional accessory proteins (vif, vpr, vpu, nef, tat, and rev). Modified figure from Sakuma et. al., 2012.
**Figure 3. Lentiviral entry into a target cell.**

Viral entry begins with virus attachment via its viral envelope glycoproteins to specific receptors on the membrane of the target cell followed by fusion with the cell membrane that internalizes the viral core into the cytoplasm of the target cell. Once inside the cytoplasm, the capsid proteins are removed, and the viral core breaks down exposing the RNA genome to host proteins that initiate reverse transcription of the viral RNA genome into double stranded DNA that gets integrated in the host genome. Modified figure from Campbell EM and Hope TJ, Nature Reviews Microbiology 2015.
Figure 4. Schematic of lentiviral third generation vector.

Replication-defective lentiviruses lack structural proteins in their genome. To make viral particles, the structural proteins, gag-pol, envelop (VSV-G), and regulatory proteins (Rev-Tat) must be added in trans as separate plasmids. Third generation vector design involves a 4-plasmid system that includes the three LV structural genes, and the vector plasmid containing the gene of interest.
Chapter 2. Materials and Methods

Construction and production of barcoded lentivirus vectors

High diversity barcoded libraries used as a starting point for this study consisted of a 35 base pair barcode, a 6 base pair library ID, and PCR primer binding sites that flank the library ID and barcode\textsuperscript{26, 27}. Barcoded libraries were originally made in a protein and nucleic acid facility (Stanford University School of Medicine) where oligos were synthesized (ABI 3900 DNA/RNA synthesizer, Fremont CA) and cloned into the lentivirus backbone pCDH (System Biosciences, Palo Alto, CA) downstream of a marker copGFP gene\textsuperscript{26} and utilized in first murine and then rhesus macaque animal models to track clonal output from hundreds-thousands of individual HSPC\textsuperscript{26, 27}. These original vectors drove marker gene expression from a murine stem cell virus (MSCV) promoter positioned within the LTR of the provirus. The vector libraries were documented to be of sufficient diversity to transduce standard doses of murine and rhesus engrafting cells\textsuperscript{26, 27, 231}.

For the current experiments, we generated a set of new lentiviral vectors containing these highly diverse libraries, moving the barcode cassette consisting of the primer binding sites for PCR amplification flanking the library ID and barcodes into the U3 region of the 3’LTR resulting in a copy of the barcode at both ends of the LTRs in the integrated provirus. We chose three different barcode libraries with three distinct 6 bp library IDs preceding the highly diverse barcodes and extracted the barcodes via restriction enzyme digestion using EcoRI (ThermoFisher Scientific, Waltham MA) and BamHI (ThermoFisher Scientific, Waltham MA) located outside the primer sequences that flank the barcode region. The restriction digest reaction for each of the barcoded
libraries were loaded onto a 2% agarose gel to separate the barcode via electrophoresis, barcode was excised from the agarose gel and gel purified (MinElute Gel Extraction Kit, Qiagen, Valencia CA). Three different lentiviral vector plasmids, each containing a different enhancer/promoter were also subjected to restriction enzyme digestion with BamHI and EcoRI, restriction sites located in the U3 region of the 3’LTR of the vectors. Digested lentiviral vectors were loaded onto a 1% agarose gel, vector backbone was excised from the agarose gel and gel purified. The barcode cassettes were subcloned into the U3 region 40 base pairs from the 5’ end of the 3’LTR in each of three lentiviral vector plasmids (Figure 5). The three vector plasmids used were as follows: vector with internal EF1-α promoter with library ID TCAAGT using vector plasmid pCDH-EF1-MCS-T2A-copGFP (System Biosciences, Palo Alto, CA, #CD526A-1), vector with LTR MSCV promoter library ID GATCTG using vector plasmid pCDH-MCS-T2A-copGFP-MSCV (System Biosciences, #CD523A-1), and vector with LTR SFFV promoter library ID GTAGCC using vector plasmid LV.SF.LTR (obtained from Eugenio Montini, Ospedale San Raffaele Srl15).

Viral particles were produced by calcium chloride cotransfection of 293T cells, an adherent human embryonic kidney cell line, cultured in T162 cm² culture flasks (Corning, NY). The 4-plasmid transfection mixture consisted of 50 ug barcoded lentiviral vector plasmids expressing the green fluorescent protein (GFP), 10 ug HIV-1/Rev-tat, 10 ug pCAGGS-VSV-G, and 30 ug χHIV (HIV-1 Gag/Pol-plus-sCA) in calcium chloride (Quality Biological, Gaithersburg MD) and HBS. The χHIV codes for the gag-pol gene with the additional modification of substituting the simian immunodeficiency virus (SIV) capsid (sCA) in the HIV-1 vector to overcome a block to
HIV transduction of rhesus macaque target cells\textsuperscript{136}. The plasmid mixture was incubated at room temperature for 30 minutes to allow for precipitates to form prior to the 6-hour transfection at 37°C, 5% CO\textsubscript{2} at which point the transfection media was carefully removed and fresh Iscove’s Modified Dulbecco’s Medium (IMDM) (Life Technologies Corporation, Grand Island NY) supplemented with 10% fetal bovine serum (FBS) (Sigma, St. Louis, MO) was added. The transfection media was removed because calcium chloride is toxic to the cells and we aimed to strike a balance between keeping the transfection media in place long enough for the plasmids to enter the cells, but not long enough to decrease viability of the cells. Virus particles were harvested at 24- and 48-hours post-transfection by carefully collecting the media from the transfected cells, replacing it with fresh media, and storing the 24-hour harvest at 4°C overnight until the 48 hour harvest the next day was collected in which both virus collections were combined. The harvesting step is done carefully as transfected 293T cells loosely attach to the flask and can easily detach upon agitation; the more cells are attached to the flask, the more virus is obtained resulting in higher virus titers. Virus particles were concentrated by high-speed ultracentrifugation (18000xg, 4°C, 3 hours) of the combined 24- and 48-hour viral harvest. Viral pellets from each of the lentiviral vectors were resuspended in a total of 1mL X-vivo media by gentle shaking at 4°C or on ice.

Viral titers for each lentiviral vectors’ preparation was measured by transducing HeLa cells, an adherent permissive human cervical cancer cell line commonly used for titering virus. Hela cells were seeded at 8.0E+04 cells/well in a 12-well plate in Dulbecco’s Modified Eagle Medium (DMEM) (Life Technologies Corporation, Grand Island NY) supplemented with 10% FBS, Penicillin/Streptomycin/Glutamine (PSG) (Life
Technologies Corporation, Grand Island NY), and cultured for 24 hours at 37°C. After 24-hour culture, the cells were transduced using serial dilutions of virus preparations in a total volume of 0.5 mL DMEM in the presence of 4 ug/mL protamine sulfate (Sigma, St. Louis, M). Ten serial dilutions were carried out for each vector (Figure 6) to increase the accuracy for titer calculations. Fourteen to sixteen hours after transduction, media was removed by careful aspiration and replaced by fresh DMEM/10% FBS media and seventy-two hours after transduction, the frequency of GFP-expressing cells was measured by flow cytometry using the BD LSR Fortessa cell analyzer (BD Biosciences, San Jose CA). The aim is to produce high-titer virus of 1.0E+08 virus particles per mL to avoid having to use too high volume of virus when transducing CD34+ HSPC. When virus is concentrated using ultracentrifugation, in addition to virus particles, proteins, molecules, and factors from the FBS included in the media are also concentrated. High-titer virus allows for smaller volumes of virus to be used during transduction of HSPC and this minimizes inclusion of additional factors concentrated in the virus pellet that are toxic to the cells. To calculate viral titers, an average cell count is obtained on the day of transduction using 3-5 wells of HeLa cells that were plated specifically for cell counting. To account for autofluorescence when evaluating GFP by flow cytometry, 2-3 wells of HeLa cells were kept as untransduced to serve as negative controls for GFP expression (Table 1 and Figure 6). Titer calculation is based on multiplying the cell number obtained on the day of transduction with the percent GFP from the flow cytometry analysis. This is then divided by the volume of virus in microliters from the specific serial dilution and the final number is multiplied by 1000 to convert to milliliters (Figure 7, Table 1). Two rhesus macaques were used for this study, their animal identification
are ZJ41 and ZJ48. Viral titers for the vectors used to transduce CD34+ HSPC from ZJ41 were 2.5E+08, 1.9E+08, and 8.1E+07 virus particles per mL for EF1a, MSCV, and SFFV respectively. Viral titers of virus used to transduce CD34+ HSPC from ZJ48 were 2.6E+08, 1.3E+08, and 2.9E+08 virus particles per mL for EF1a, MSCV, and SFFV respectively (Table 2).

In order to ensure that each target transduced HSPC has a high likelihood of containing a unique barcode, it is critical to monitor library barcode diversity in each vector preparation prior to use. In order to assess diversity, each vector preparation was used to transduce K562 cells, a non-adherent human chronic myelogenous leukemia cell line. We chose to use K562 cells because of their non-adherent properties, just like our target cells, HSPCs and also for their ease of handling in which avoids the additional step of trypsinizing. K562 cells were transduced separately with each of the 3 different barcoded lentiviral libraries. Twenty-four hours following transduction, transduced K562 cells were pelleted by centrifugation, and DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Valencia CA). Low-cycle PCR using cycling parameters of 98°C (30 seconds) as the initial denaturing step followed by 28 cycles of 98°C (10 seconds) for denaturation, 70°C (30 seconds) for annealing, and 72°C (30 seconds) for extension and then a final extension at the end of the 28th cycle at 72°C for 10 minutes was performed to retrieve the barcode. The PCR reaction was loaded onto a 2.5% agarose gel to separate the barcode DNA fragments and then the barcode DNA fragments were gel purified (MinElute Gel Extraction Kit, Qiagen, Valencia CA) for high-throughput illumina sequencing (further explained under the barcode retrieval section). Primer sequences used for high-throughput sequencing are listed on Table 3.
Retention of sufficient barcode diversity within each barcoded vector library was confirmed via Monte Carlo simulations that are based on running trials using custom Python code to calculate the number of engrafting cells that could be transduced with 95% certainty that each unique barcode is only present in a single transduced cell 95% of the time, code available at (https://github.com/dunbarlabNIH)\textsuperscript{27,231}. A Monte Carlo simulation with a null hypothesis that >95% of barcodes representing single cells using the K562 barcode retrieval data is shown in Figure 8. Barcode retrieval data from K562 cells transduced separately with each of the three lentiviral barcode libraries were used to compile a list consisting of each retrieved barcode, with each instance of retrieval of an individual barcode included. A target cell population size is chosen for a simulation trial (e.g. 1500 cells) and for each trial, the program randomly selected 1500 barcodes from the weighted retrieved barcode list. If 1500 barcodes chosen in that round are unique 95% of the time or greater, this round is counted as a success, otherwise the round is counted as a failure. 100 rounds are performed per trial. Each trial is then plotted as percent failure (1 - fraction success), shown as a black dot on the graphs (Figure 8). Vector libraries were only utilized for transplantation experiments if sufficiently diverse. We target to transduce $1.0\times10^7$ rhesus CD34+ cells with each vector library. The frequency of engrafting cells has been previously estimated to be 5-10 HSC per 100,000 CD34+ cells, therefore, a total of 500 – 1,000 engrafting cells must be transduced with each preparation.
Transduction and autologous transplantation of rhesus macaque CD34+ HSPC

The NHLBI Animal Care and Use Committee approved all animal studies. Rhesus macaque HSPC were mobilized into the peripheral blood from the bone marrow by treatment with 4 days of granulocyte colony-stimulating factor (G-CSF) 10 µg/kg/day (Amgen, Thousand Oaks, CA) and AMD3100 (plerixafor) 1mg/kg (Sigma, St. Louis, MO) on the morning of the 5th day, 3-4 hours prior to apheresis232. Collection of mobilized peripheral blood stem cells (PBSCs) from rhesus macaque was done via apheresis using the Fenwal CS-3000 Plus cell separator (Baxter Healthcare)233. The apheresis procedure is a multistep process with the animals first being sedated with 15 mg/kg ketamine (Fort Dodge Animal Health, Fort Dodge IA) followed by isoflurane (Abbott Laboratories, Lake Bluff, IL) for general anesthesia to allow the insertion of a 18-20 gauge angiocatheter (connected to the Fenwal CS-3000) to be inserted into a saphenous or cephalic vein. Throughout the apheresis procedure 15U/kg heparin (American Pharmaceutical Partners, Schaumburg IL) was administered via IV to maintain anticoagulation. Apheresis was continued for 2-3 hours. The apheresis collection bag contained mononuclear-cell enriched blood cells, including peripheral blood stem cells (PBSC) in a volume 37mL. This apheresis product was run through the Sysmex XT-2000i complete blood cell count (CBC) analyzer (Sysmex America, Inc., Lincolnshire, IL) which identifies the different blood cells according to their cellular characteristics. The CBC was obtained before administration of cytokines (pre-cytokines), pre-, mid-, and end-apheresis and of the final apheresis product (Table 4, Table 5). Based on the CBC counts, the total number of white blood cells (WBC) in 37mL of the apheresis product was 1.42E+10 for ZJ41 and 1.35E+10 WBC for ZJ48.
Likewise, the total number of PBMNC in the final apheresis product can be calculated from the CBC and yielded 5.28E+09 for ZJ41 and 2.87E+09 for ZJ48 PBMNC (Table 4, Table 5). The apheresis product was further processed by gently layering diluted PBSCs (1:3 in PBS) on top of Ficoll Paque plus (GE Healthcare, Marlborough MA) with caution not to disturb the interface that forms, and centrifuged at 1600rpm, room temperature, 30 minutes with the centrifuge break off to avoid disturbance of the density gradient layers. The mononuclear layer was collected from the Ficoll interface via gentle pipetting, treated with Ammonium-Chloride-Potassium (ACK) lysis buffer (Quality Biological, Gaithersburg MD), to lyse anucleated red blood cells, washed with PBS, and cell counts were obtained using the Vi-CELL XR cell viability analyzer (Beckman Coulter Life Sciences, Indianapolis, IN). PBMNC cell counts resulted in a total of 1.6E+10 for ZJ41 1.2E+10 for ZJ48. PBMNC were then enriched for CD34+ HSPC via immunoselection using a CD34+ antibody clone 12.8 (Dr. Robert Andrews, Fred Hutchinson Cancer Research Center, Seattle WA), and MACS micro-beads/separation columns (Miltenyi Biotec, Auburn CA). Immunomagnetic enrichment has been the method of choice, not only is it gentle on the cells, but it is also a quick method to sort rare cells, like CD34 HSPC, in large samples. FACS on the other hand, although provides high purity of sorted cells, it is slow as each cell is processed one at a time, as opposed to immunomagnetic enrichment when the whole sample passes through a column at once allowing for quick enrichment of bulk sample with high purities. Time-is-of-essence for CD34+ enrichment to minimize ex vivo manipulation and lower the incident of CD34 cell differentiation. CD34 enrichment resulted in 6.2E+07 total CD34+ cells for ZJ41 with a 96.3% purity and 2.8E+07 total CD34+ cells for ZJ48 with 86.3% purity.
The purity of CD34+ HSPC was obtained using the Cytomics FC 500 flow cytometer (Beckman Coulter Life Sciences, Indianapolis, IN) on a small sample of the CD34+ enriched cell product using a CD34 antibody conjugated to PE (Figure 9, Table 6).

To transduce HSPCs with the different vectors, CD34+ cells from each collection were split into three equal fractions, and following overnight culture at a density of 1.0E+06 cells/mL on RetroNectin-coated plates (Takara, #T100B, Mountain View CA) in X-VIVO 10 media (Lonza, Rockland ME) supplemented with 100ng/ml recombinant human fms-related tyrosine kinase 3 (flt3) ligand, stem cell factor (SCF) and thrombopoietin (Miltenyi Biotec, Auburn CA), each fraction was transduced with one of the three vectors at a multiplicity of infection (MOI) of 25 in the presence of 4 ug/ml protamine sulfate (Sigma, St. Louis, MO). Multiplicity of infection is defined by the ratio of the number of virus particles per target cell, for example, MOI 25 is equivalent to using 25.0E+06 virus for every 1.0E+06 cells. Our laboratory has found that transduction of HSC with an MOI 25 results in approximately 30% transduction efficiency, optimal for favoring each HSC to be transduced with no greater than a single vector copy number or barcode. Twenty-four hours later, all three fractions of transduced cells were collected from the plates, pooled, and reinfused into the autologous rhesus macaque.

Each animal received total body irradiation at a dose of 500 cGy/day for two days prior to the day of cell reinfusion. Aliquots of transduced CD34+ cells were removed prior to pooling and reinfusion and cultured for an additional 48 hours prior to flow cytometric analysis of GFP expression allowing maximal expression of the GFP marker.
gene from integrated proviral forms of the vectors. Cell aliquots from the end of transduction were also removed for DNA extraction and barcode recovery.

Cell processing and flow cytometric sorting

Peripheral blood was separated into mononuclear cells and granulocytes via centrifugation over lymphocyte separation medium (MP Biomedicals, Solon OH), which is a mixture of ficoll and sodium diatrizoate with a specific density of 1.077-1.080 g/mL. Cells were then incubated with ACK lysis buffer (Quality Biological, Gaithersburg MD) to remove red blood cells, and labeled with fluorophore-conjugated monoclonal antibodies against CD3, CD20, CD14, CD56, CD16 and CD33 (antibody clone and source information summarized in Table 6), allowing sorting of individual granulocyte, monocyte, T cell and B lineages on a FACSARiaII flow cytometer (BD Biosciences, San Jose CA). The gating strategy to sort for specific cells is based on first separating the cell populations based on the forward and side scatter measurements determined by the optical detectors on the flow cytometer. Forward scatter is a measurement based on size, it measures the diameter of the cell, therefore, the higher the forward scatter intensity, the larger the diameter of the cell. Because immune cells are of different sizes, the forward scatter is a useful strategy for identifying immune cells based on size, for example, lymphocytes are smaller than monocytes and therefore can be separated via the forward scatter. The side scatter measures the internal complexity of the cell, such as the granularity which includes both granules and the nucleus of cells. The side scatter is especially useful in discriminating between monocytes and granulocytes since granulocytes have many granules (higher complexity) and display a higher side scatter.
measurement. The second set of gates selects for single cell populations and is based on the forward side scatter height and area, it excludes any non-single cells as a result of either clumping or cellular debris and is also the last gate before gating on the specific fluorophores. To sort for the granulocyte population, following the singlet gate, cells were gated on CD33+ expression, in which GFP expression could be further evaluated (Figure 10). To sort for the different lymphocyte populations, the first gate was plotted against CD3 and CD20, T cells were selected based on CD3+CD20- expression and B cells on CD20+CD3- expression. GFP levels were evaluated for each cell type (Figure 10). Furthermore, monocytes and NK cell populations were identified based on CD3-CD20- expression. Monocytes were selected by their CD14 expression and natural killer (NK) from the CD3-CD20-CD14- population selected via their CD56+CD16+ expression. The identification of NK cells based solely on CD56+ and CD16+ is a very simplistic method to identify NK cells and for a more accurate NK identification, more complex cellular labeling for NK cells is required. However, for the purpose of this study, simple NK cell labeling was sufficient (Figure 11). DNA was extracted from each of the sorted cell pellets using the DNeasy Blood and Tissue Kit (Qiagen, Valencia CA) for barcode retrieval.

Barcode retrieval

500ng DNA samples extracted from each purified cell type were amplified via 28 cycles of PCR using primers flanking the library ID and barcode. Primers (Table 3) added an 8pb sample ID to facilitate multiplex sequencing. Samples were run on a 2.5% agarose gel, and the resulting 157bp amplified fragment was gel purified (MinElute Gel
Extraction Kit, Qiagen, Valencia CA). Equal amounts of product were pooled and prepared for high-throughput sequencing, allowing pooling of 12-24 multiplexed samples per lane in the sequencing flow cell of the Illumina HiSeq 2000 sequencer.

Barcode retrieval and data analysis

After de-multiplexing, 4 million reads (selected randomly to avoid lane position bias) sequencing output from each sample were analyzed using custom python code. Reads were extracted via the presence of one of the three library IDs, reads with abundance lower than 100 reads were discarded (a step shown previously not to change results, excluding reads almost certainly resulting from PCR and sequencing artifacts, and speeding processing time from days to hours for each file), reads with 0-2 mismatches or indels were counted as identical and were combined to determine the read number for each library ID and then barcode for each sample. No indels were allowed within the library ID. The fractional contribution of each barcode to a sample was then calculated from the ratio of the individual barcode’s read number divided by the total number of valid barcode reads in the sample. A read number threshold of 2000 was applied to exclude “false” barcodes resulting from sequencing artifacts or presence of false library IDs in the genome. A master list of barcodes was then generated for each animal and each vector, consisting of any barcode with a read abundance of over 2000 in any sample for that animal and vector. Once a barcode was included on the master list, all analyses included the read numbers for that barcode, even if the read number in other samples fell below 2000. The threshold of 2000 was chosen conservatively as shown in Figure 12 and has been previously validated for this rhesus macaque model. Figure 13
shows the impact of applying lower and higher thresholds, which impact on enumeration of overall clone numbers, but do not impact on tracking of high contributing or potentially expanding clones. Distinguishing very low contributing clones from “false” barcodes or recurrent sequencing artifacts is impossible, but real clones contributing at a very low level in every sample are unlikely to be biologically important.

Using our methodology on primate samples with this level of marking we have previously shown that lower thresholds result in counting of “false” barcodes and sampling issues, and higher thresholds result in exclusion of valid and reproducibly retrieved clones

Stacked area plots were made using R studio (Integrated Development for RStudio, Inc. (2015), Boston MA URL http://www.rstudio.com/) and custom R code. Heatmaps were produced using custom R code. All custom Python and R code can be accessed at: github.com/truittll/genotoxicity. Prism 6.0 (GraphPad software, La Jolla CA) was used to make dot plots diagrams.

K562 validation of quantitative simultaneous barcode retrieval from 3 lentiviral libraries

K562 cells were transduced separately with each of the 3 different vectors at a multiplicity of infection of 0.5 in the presence of 4 ug/ml protamine sulfate (Sigma, St. Louis, MO) in RPMI 1640 medium (Life Technologies Corporation, Grand Island NY) supplemented with 0.5 mg/mL penicillin-streptomycin-glutamine (Life Technologies Corporation, Grand Island NY) and 10% fetal bovine serum (Sigma, St. Louis, MO). Cells were cultured at 37°C in the presence of 5% CO₂ followed by single cell sorting on
GFP expressing cells to obtain single cell clones. DNA was extracted from the different K562 clones and subjected to Southern blotting\textsuperscript{27} (Figure 14) for identification of clones containing a single vector copy. Two enzymes were used to digest the DNA in preparation for Southern blot, one, a single cutter previously used in our lab and known to work well, high-fidelity SphiI (New England Biolabs, Ipswich, MA) and two, a high-fidelity SalI (New England Biolabs, Ipswich, MA) absent in our vectors, but known to be a frequent cutter. Probe used for hybridization targeted a 360bp fragment of the WPRE site on the vectors. Because our laboratory does not have access to a radioactive facility, DNA samples along with the probe were sent out to Lofstrand Labs for Southern blot generation (Lofstrand Labs Limited, Gaithersburg MD). A total of 15ug DNA of each sample (11 EF1-\(\alpha\) clones, 8 MSCV clones, 9 SFFV clones, and 1 negative control) and 1.5ug of probe were used for Southern blot. Clones containing single vector copies based on the presence of a single band on Southern blot were identified (Figure 14). Single copy clones, highlighted for each vector in Figure 14, were mixed at different ratios of cell numbers and then diluted into untransduced K562 cells to create mixtures at a range of vector ratios and overall marking levels, corresponding to the overall levels of marking and ratios between barcode libraries (vectors) obtained \textit{in vivo} in ZJ48 and ZJ41.
Figures and Tables

- Restriction digest
  - EcoRI and BamHI
- Load onto agarose gel
- Gel extract barcode band
- Purify barcode

Ligate purified barcode to Lentiviral vector

MSCV
7156 bp

SFFV
7856 bp

EF1a
7386 bp
Figure 5. Schematic of the construction of lentiviral vector libraries.
High diversity barcoded libraries consisting of a 35 base pair barcode, a specific 6 base pair library ID, and PCR primer binding sites that flank the library ID and barcode. For each library, the barcode was obtained by restriction enzyme digestion using EcoRI and BamHI located outside the primer sequences that flank the barcode region. Restriction digest reaction for each barcoded library was loaded onto an agarose gel, followed by gel extraction and purification of barcodes. These cassettes were then subcloned into the U3 region of the 3’LTR of three different lentiviral vector plasmids that were also subjected to restriction enzyme digestion with BamHI and EcoRI. The purified barcodes were ligated to the LTR region of the lentiviral vectors through complementary overhang base pairing.
Figure 6. Virus serial dilutions.
Virus dilutions are made with media containing DMEM+10%FBS+1%PSG+Protamine sulfate. The first well on the 12-well tissue culture plate contains 1.5mL media and wells 2-10 contain 600 uL of media. Twelve microliters of virus are added to the first well, mixed, and then 600uL is transferred to the second well. The second well is mixed and then 600uL is transferred to the third well. This procedure is repeated up until the tenth well. Two to three wells are kept as untransduced controls to account for background noise/autofluorescence.
Figure 7. Virus titer calculation.
The calculation for virus titers is based on the average cell number obtained on the day of transduction, the percent GFP after subtraction of background obtained from untransduced cells, and the volume of virus used from the serial dilutions. Typically, 2-3 values from the serial dilutions are averaged to obtain a more accurate value.
Figure 8. Barcoded library vector diversity.
Top panel: Samples of K562 cells were transduced separately with each of the 3 different barcoded lentiviral libraries. 24 hours following transduction, DNA was extracted for barcode PCR and high-throughput sequencing analysis. Bottom panel: Monte Carlo simulation of the null hypothesis that >95% of barcodes represent single cells using the K562 barcode retrieval data. Monte Carlo simulations were performed using custom Python code. Barcode retrieval data from K562 cells transduced separately with each of the three lentiviral barcode libraries were used to compile a list consisting of each retrieved barcode, with each instance of retrieval of an individual barcode included. Each trial is plotted as percent failure (1 – fraction success), shown as a black dot on the graphs. Simulation trials each consisting of 100 rounds are repeated for a range of possible cell numbers, and each trial’s (1-fraction success) is plotted. A best fit curve is fit to the model using the formula displayed on the graphs. The dashed line shows the 95% cut-off for confidence that 95% of cells receive a unique barcode. The p value represents with 95% confidence that 95% of the cells or greater have a unique barcode.
Figure 9. HSPC purity.
Purity analysis of CD34+ enriched cells for ZJ41 (A) and ZJ48 (B). A small aliquot of the CD34+ enriched product was labeled using a CD34 antibody conjugated to PE and analyzed via flow cytometry. ZJ41 CD34+ cells obtained 96.3% purity and ZJ48 obtained 86.3% purity.
Figure 10. Cell sorting strategy for T, B, and granulocyte cell populations. Peripheral blood mononuclear cells (PBMNC) and granulocytes (A) were sorted based on forward and side scatter followed by the singlet cell populations and then their specific fluorophore expression. Granulocytes are gated based on CD33 expression, T cells on
CD3 expression, and B cells gated based on CD20 expression. GFP expression from T, B, and granulocytes was evaluated from each cell type.
Figure 11. Cell sorting strategy for monocyte and NK cell populations. Following the lymphocyte/monocyte gate and singlet gate, monocytes were further gated from the CD3-CD20- cell population based on CD14+ expression. NK cells are gated based on CD3-CD20-CD14- population based on CD16+ and CD56+ expression. GFP expression from monocytes and NK cells was then evaluated for each cell type.
**Figure 12. Threshold comparisons.**
The number of barcodes detected based on different thresholds applied are shown. Each line represents a different library. An optimal threshold is when the threshold can be changed, but the number of barcodes is not significantly changed. Threshold is selected based on the location where the plot begins to plateau (knee/elbow of the graph). The goal is to choose a balance between a small threshold that includes real barcodes with limited background/noise and a large threshold that will be stringent enough without excluding too many small contributing clones. Based on these plots, a threshold of 4000 (ZJ41) and 3000 (ZJ48) is where the number of barcodes are more stable, so a 2000 threshold will include a lot of small insignificant clones yet even including these clones, we observe no change in our data and all clones are stable. Because our lab has found that a threshold of 2000 is applicable to our barcoding data, we chose to use a threshold of 2000 for consistency.
Figure 13. Impact of threshold variation on clone numbers and top clone contribution plots.
The number of barcodes detected based on two different thresholds of 1000 and 4000 reads were applied and plotted (compared to Figures 4 and 5 using standard threshold of 2000). (A) Number of unique barcodes at each time point and the cumulative numbers of
barcodes for each vector (EF1-α, MSCV, SFFV) retrieved from all lineages (granulocytes, B cells and T cells) in ZJ41 and ZJ48. Thresholds of 1000 and 4000 were applied to each master clone list and unique and cumulative barcodes were plotted over time. To include a barcode on the master list it must have contributed 1000 reads (for a 1000 threshold) or 4000 reads (for a 4000 threshold) to at least one time point. Once established on the master list, these barcodes could then be counted throughout other time points as long as they appear at least once, even if contributing less than 1000 or 4000 reads (at their respective thresholds) at later time points. The same barcode found contributing to more than one lineage was counted only once.  

(B) Stacked area plots using master clone lists generated using two different thresholds of 1000 and 4000 reads are shown. Plots delineate the percent contribution of the top 20 contributing clones in each lineage for each vector (EF1-α, MSCV, SFFV) at the time of longest follow-up (38 months for ZJ41 and 33 months for ZJ48), tracked over all time points. Each of the top 20 clones is shown as a separate shaded region in the stack. Contributions from non-top 20 clones are shown as the remaining single solid color.
Figure 14. Southern blot for single clone identification.
Southern blot of K562 clones transduced separately with each of the vectors to identify vectors containing a single vector copy. Two restriction enzymes were used, high-fidelity SphI and high-fidelity SalI. Probe used for hybridization targeted a 360bp fragment of the WPRE site on each of the vectors. Sample #6 from the 11 EF1-α clones (black box), sample #17 from the 8 MSCV clones (pink box), and sample #24 from the 9 SFFV clones (blue box) were chosen for spike in experiments (Chapter 4).
Table 1. Viral titers for MSCV vector library used in ZJ48. Viral titers are based on serial dilutions of virus and GFP expression.

<table>
<thead>
<tr>
<th>Well No.</th>
<th>Virus volume (uL) equivalent per well</th>
<th>GFP %</th>
<th>Virus/mL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.0</td>
<td>91.48</td>
<td>2.78E+07</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>88.98</td>
<td>5.41E+07</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>81.68</td>
<td>9.93E+07</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>54.88</td>
<td>1.33E+08</td>
</tr>
<tr>
<td>5*</td>
<td>0.25</td>
<td>39.18</td>
<td>1.91E+08</td>
</tr>
<tr>
<td>6*</td>
<td>0.125</td>
<td>19.28</td>
<td>1.88E+08</td>
</tr>
<tr>
<td>7</td>
<td>0.0625</td>
<td>4.96</td>
<td>9.66E+07</td>
</tr>
<tr>
<td>8</td>
<td>0.03125</td>
<td>2.47</td>
<td>9.63E+07</td>
</tr>
<tr>
<td>9</td>
<td>0.015625</td>
<td>0.748</td>
<td>5.82E+07</td>
</tr>
<tr>
<td>10</td>
<td>0.0078125</td>
<td>0.351</td>
<td>5.46E+07</td>
</tr>
<tr>
<td>Background</td>
<td>---</td>
<td>0.016</td>
<td>---</td>
</tr>
</tbody>
</table>

Asterisked (*) wells were used to average the virus/mL values to obtain the final titer. GFP values shown already account for autofluorescence from untransduced control.

Table 2. Summary of virus titrations of EF1a, MSCV, and SFFV for animals ZJ41 and ZJ48.

<table>
<thead>
<tr>
<th>Vector</th>
<th>ZJ41 Virus/mL</th>
<th>ZJ48 Virus/mL</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF1a</td>
<td>2.5E+08</td>
<td>2.6E+08</td>
</tr>
<tr>
<td>MSCV</td>
<td>1.9E+08</td>
<td>1.3E+08</td>
</tr>
<tr>
<td>SFFV</td>
<td>8.1E+07</td>
<td>2.9E+08</td>
</tr>
</tbody>
</table>
Table 3. Primer sequences used for multiplexing illumina high-throughput sequencing.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univ rev</td>
<td>CAAGCAGAAGACGGCATACGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-01</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-02</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-03</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-04</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-05</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-06</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-07</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-08</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-09</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-10</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-11</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-12</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-13</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-14</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-15</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-16</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-17</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-18</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-19</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-20</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-21</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-22</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-23</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-24</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-25</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-26</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-27</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-28</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
<tr>
<td>Fwd i5-29</td>
<td>AATGATACGGCCACCACCGAGATCTACGTGATAACGGCATACGAGCTCTTCCGATCT</td>
</tr>
</tbody>
</table>

Highlighted in red is the 8bp i5 index
Table 4. ZJ41 CBC results for the various steps during apheresis and the final apheresis product.

<table>
<thead>
<tr>
<th></th>
<th>WBC (cells/μL)</th>
<th>% lymph CBC analyzer</th>
<th>Theoretical PBMNC</th>
<th>Total PBMNC</th>
<th>Total CD34+ (after selection)</th>
<th>% CD34+ purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre cytokines</td>
<td>8.42E+03</td>
<td>24.2</td>
<td>7.54E+07</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Pre apheresis</td>
<td>8.99E+04</td>
<td>18.6</td>
<td>6.19E+08</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Mid apheresis</td>
<td>8.99E+04</td>
<td>10.5</td>
<td>3.49E+08</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>End apheresis</td>
<td>9.54E+04</td>
<td>7.5</td>
<td>2.65E+08</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Final product</td>
<td>3.85E+05</td>
<td>37.1</td>
<td>5.28E+09</td>
<td>1.60E+10</td>
<td>6.20E+07</td>
<td>96.3</td>
</tr>
</tbody>
</table>

Table 5. ZJ48 CBC results for the various steps during apheresis and the final apheresis product.

<table>
<thead>
<tr>
<th></th>
<th>WBC (cells/μL)</th>
<th>% lymph CBC analyzer</th>
<th>Theoretical PBMNC</th>
<th>Total PBMNC</th>
<th>Total CD34+ (after selection)</th>
<th>% CD34+ purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre cytokines</td>
<td>8.46E+03</td>
<td>26.4</td>
<td>8.26E+07</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Pre apheresis</td>
<td>7.62E+04</td>
<td>15.8</td>
<td>4.46E+08</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Mid apheresis</td>
<td>6.54E+04</td>
<td>11.4</td>
<td>2.76E+08</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>End apheresis</td>
<td>6.47E+04</td>
<td>9.9</td>
<td>2.37E+08</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Final product</td>
<td>3.64E+05</td>
<td>21.3</td>
<td>2.87E+09</td>
<td>1.20E+10</td>
<td>2.80E+07</td>
<td>86.3</td>
</tr>
</tbody>
</table>

Table 6. Fluorochrome-conjugated antibodies.

<table>
<thead>
<tr>
<th>Antigen</th>
<th>Antibody clone</th>
<th>Conjugate fluorophore</th>
<th>Company/catalog number</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD3</td>
<td>SP34-2</td>
<td>APC-Cy7</td>
<td>BD Pharmingen (557757)</td>
</tr>
<tr>
<td>CD20</td>
<td>2H7</td>
<td>PE-Cy5</td>
<td>BD Pharmingen (555624)</td>
</tr>
<tr>
<td>CD14</td>
<td>TUK4</td>
<td>Pac blue</td>
<td>Invitrogen (MHCD1428)</td>
</tr>
<tr>
<td>CD16</td>
<td>3G8</td>
<td>APC</td>
<td>Biolegend (302012)</td>
</tr>
<tr>
<td>CD56</td>
<td>B159</td>
<td>PE</td>
<td>BD Pharmingen (555516)</td>
</tr>
<tr>
<td>CD33</td>
<td>AC104.3E3</td>
<td>PE</td>
<td>Miltenyi (130-091-732)</td>
</tr>
<tr>
<td>CD34</td>
<td>563</td>
<td>PE</td>
<td>BD Pharmingen (550761)</td>
</tr>
</tbody>
</table>
Chapter 3. Results I: Development of Barcoded Vectors

Development of an in vivo platform to study relative vector toxicity

We generated three high diversity barcode libraries containing one of three 6 base pair library ID’s followed by 35 random base pair barcodes and flanked by sequences complementary to Illumina sequencing primers, and inserted these cassettes into the U3 region of three different lentiviral vector plasmids (Figure 15)\(^{26,27}\). Each vector also contains an internal green fluorescent protein (GFP) marker to allow the evaluation of transduction efficiencies and tracking. During the lifecycle of the lentivirus, specifically during reverse transcription, the virus copies the 3’LTR U3 region over to the 5’LTR resulting in the integrated provirus to contain the barcode in both LTRs (Figure 15). The vectors were chosen to reflect a predicted range of genotoxicities. The elongation factor-1 alpha (EF1-\(\alpha\)) vector contains an internal relatively weak EF1-\(\alpha\) promoter, and no additional promoter/enhancer within the LTR, and would be predicted to be the least genotoxic. The murine stem cell virus (MSCV) vector contains the medium-strength MSCV promoter/enhancer cloned into the U3 region of the LTR. The spleen focus-forming virus (SFFV) barcoded lentiviral vector contains the very strong SFFV promoter/enhancer within the U3 region of the LTR, predicted to be the most genotoxic LV vector design based on *in vitro* immortalization and murine tumor-prone mouse assays\(^{15,184}\.\)

We confirmed that all three vectors contained the expected library ID and retained the ability to transduce and integrate in K562 cells despite insertion of the barcode cassettes into the LTR (Figure 16). Promoter strength in the three vectors was assessed in transduced K562 populations by measuring GFP expression via flow cytometry. As
expected, cells transduced with the SFFV vector had the brightest GFP expression, followed by the MSCV vector and then the EF1-α vector with stepwise dimmer GFP expression (Figure 16A). The larger the separation between GFP+ and GFP- populations, the higher the separation index (SI) value and the stronger the promoter/enhancer. As expected, SFFV-transduced K562 cells had the highest SI followed by MSCV and then EF1-α with the lowest SI value (Figure 16B).

Barcode library diversity for each vector construct was assessed via transduction of K562 cells, barcode retrieval via PCR, and Illumina sequencing. Monte Carlo simulations were performed on K562 cells transduced with each vector preparation to ensure each of the three barcoded libraries were diverse enough to result in a 95% chance that each barcode is present in a single engrafting HSPC (Chapter 2 Figure 8). These simulations indicated that the EF1-α, MSCV, and SFFV lentiviral libraries could transduce 1,547, 1,986, and 1,943 long-term engrafting HSPC respectively with >95% confidence (i.e. p<0.05) that >95% of individual barcodes would mark a single cell. Based on frequencies of long-term engrafting macaque HSPC within CD34+ populations estimated via integration site retrieval, flow cytometric phenotyping for primitive markers, or immunodeficient mouse engraftment\textsuperscript{23, 236, 237}, the diversities of the three libraries were sufficient to transduce the rhesus macaque CD34+ HSPC grafts utilized in our experimental design.

Peripheral blood CD34+ HSPC collected following G-SCF/plerixafor mobilization from two rhesus macaques were split into three equal fractions, and each fraction was transduced independently with one of the three lentiviral vectors (Figure 17A). A transduction efficiency of no greater than 30% was targeted in order to favor
transduction of individual HSPC with no more than a single vector copy, containing a unique barcode. The three transduced cell fractions were then combined and reinfused simultaneously into the autologous rhesus macaque recipient, following delivery of 1000 rads ablative total body irradiation to the animal (Figure 17A).

Information regarding CD34+ cell collection, transduction and transplantation for both animals are summarized in Table 7. Following infusion of the transduced CD34+ cells, the animals recovered blood counts promptly. GFP expression in peripheral blood granulocytes, monocytes, B cells, and T cells was assessed over time by flow cytometry, and both animals showed relatively low but stable GFP percentages following engraftment (Figure 17B) well within the target range calculated by Poisson distribution analysis predicting the fraction of HSPC containing a single barcode (Figure 18).
Figures and Tables

A. EF1α

Three high diversity barcode libraries were cloned upstream of the U3 region of the long terminal repeat (LTR) in three lentiviral vectors. The structure of the proviral integrated form is shown schematically for each vector: (A) Internal human elongation factor 1-α (EF1-α) promoter driving GFP expression, (B) Murine stem cell virus (MSCV) promoter/enhancer in the LTR driving GFP expression, and (C) Spleen focus-forming virus (SFFV) viral promoter/enhancer driving GFP expression.

Figure 15. Design of barcoded vectors.
Figure 16. Analysis of comparative promoter strength of each lentiviral vector. (A) Flow cytometric analysis of K562 cells transduced with the different vectors at the same multiplicity of infection demonstrate varying intensities of GFP expression within the GFP+ cell population, with SFFV>MSCV>EF1α. (B) Separation Index (SI) was evaluated to compare the distance between positive and negative GFP populations for each vector. The higher the SI value, the larger the separation between the peaks and the stronger the expression from each vector’s promoter/enhancer.
Figure 17. Experimental design and GFP marking post-transplantation.
(A) Experimental design. Rhesus macaque peripheral blood stem and progenitor cells were mobilized into the blood with G-CSF and plerixafor and collected via apheresis. CD34+ HSPC were enriched via immunoabsorption. CD34+ cells were split into 3 equal aliquots and each fraction was transduced with one of three LV vectors as shown. Following transduction, the cells were collected, combined and reinfused into the autologous macaque after the macaque completed total body irradiation (TBI) 1000 rads.
(B) Percentage of GFP+ cells in peripheral blood granulocytes, B cells and T cells tracked over time to longest follow-up post-transplantation in animals ZJ41 (38 months) and ZJ48 (33 months).
Figure 18. Predicted single copy transduction efficiency of CD34+ HSPC.

Transduced CD34+ HSPC from ZJ48 with the three different vectors were assessed for their transduction efficiencies using flow cytometric analysis. Using theoretical and measured transduction efficiencies based on GFP expression from each of the vectors (20.3% SFFV, 6.82% MSCV, 7.95% EF1-α), a Poisson distribution of barcode events was used to calculate the percent of cells that received exactly one barcode. The plot demonstrates that for each of the vectors EF1-α (grey), MSCV (pink), and SFFV (blue), of the cells containing at least one barcode, more than 90% of those cells would be predicted to receive a single barcode.
Table 7. Macaque information.

<table>
<thead>
<tr>
<th>Animal ID</th>
<th>ZJ41</th>
<th>ZJ48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>5.82</td>
<td>4.86</td>
</tr>
<tr>
<td>Mobilization cytokines</td>
<td>G-CSF, AMD3100</td>
<td>G-CSF, AMD3100</td>
</tr>
<tr>
<td>Date of transplant</td>
<td>2/11/2015</td>
<td>7/15/2015</td>
</tr>
<tr>
<td>Transduction efficiency</td>
<td>N/A</td>
<td>SFFV (20.3%) MSCV (6.82%) EF1α (7.95%)</td>
</tr>
<tr>
<td>CD34+ cells transduced</td>
<td>3.60E+07</td>
<td>2.82E+07</td>
</tr>
<tr>
<td>CD34+ cells infused</td>
<td>6.00E+07</td>
<td>2.84E+07</td>
</tr>
<tr>
<td>Age day of transplant</td>
<td>43m 13d</td>
<td>48m 3d</td>
</tr>
<tr>
<td>Infused cells/kg</td>
<td>1.03E+07</td>
<td>5.84E+06</td>
</tr>
<tr>
<td>SFFV virus titer</td>
<td>8.10E+07</td>
<td>2.90E+08</td>
</tr>
<tr>
<td>MSCV virus titer</td>
<td>1.90E+08</td>
<td>1.30E+08</td>
</tr>
<tr>
<td>EF1α virus titer</td>
<td>2.50E+08</td>
<td>2.60E+08</td>
</tr>
</tbody>
</table>
Chapter 4. Results II: In vivo Data from Autologous Transplantation

Post-transplantation analysis of vector clonal contributions over time

To assess clonal patterns in vivo at time points from one month through 33 and 38 months post-transplantation, blood was obtained and granulocytes, T cells and B cells were lineage-purified by flow cytometry, followed by DNA extraction, low cycle PCR with primers bracketing the library IDs and barcodes, high-throughput Illumina sequencing, and data processing to retrieve and quantitate the read fraction for each individual barcode linked to one of the three vector-associated library IDs. To be included on a master clone list for tracking and analysis, a barcode had to be retrieved at a read count of at least 2000 at one or more timepoints, as detailed in the Methods section, in order to exclude potential false barcodes and to account for sampling constraints.

The overall level of contribution for each of the three vectors in a sample was assessed via normalization of the read number of each of the three library IDs to the total overall reads containing a valid library ID in that sample. We observed that the relative contributions of each of the three vectors to ongoing hematopoiesis was not uniform in either animal. In ZJ41, the SFFV vector accounted for over 90% of the reads in almost every sample in all three hematopoietic lineages, with far lower contributions from cells containing the MSCV or EF1-α vectors. In contrast, in animal ZJ48, the MSCV and EF1-α contributions were much higher and quite similar, with much lower contributions from SFFV (Figure 19A). We have no clear explanation for the observed differences in transduction efficiency of engrafting cells between the different vectors within each animal, but given that each vector performed well in at least one animal, the variability
does not reflect vector design, but instead likely resulted from some unknown characteristic of the actual viral preparation used for transplantation or variability in handling of the cell fractions during transduction with each vector. To rule out any technical variability or bias regarding barcode retrieval from each of the three vectors, a range of mixtures of single-copy barcoded K562-transduced cell clones for each of the three vectors were conducted (Figure 20). For example, spike-in dilutions of cells were mixed based on cell counts at different ratios and untransduced K562 cells were included in the mix as a diluent to resemble marking levels from each of the animals, 4.41% and 1.59%, for ZJ48 and ZJ41 respectively. The mixtures consisted of the following ratios for EF1-α, MSCV, SFFV; mix 1 (33%:33%:33%) with 4.41% GFP marking, mix 2 (38%:60%:3%) with 4.41% GFP marking, mix 3 (33%:33%:33%) with 1.59% GFP marking, and mix 4 (3%:0.7%:97%) with 1.59% GFP marking as shown in Figure 20. Barcode retrieval quantitation was closely correlated with the expected barcode frequency based on Pearson correlation coefficient of 0.9961, 0.999, and 0.9994 for EF1-α, MSCV, and SFFV respectively, as our lab has previously reported for other vector designs27.

Clonal diversity and clonal dynamics over time

Quantitative tracking of clonal contributions from each unique barcode allows tracking of output from large numbers of individual HSPC and allows for assessment of clonal dominance linked to LV vector design. Both animals had individual barcode contributions assessed for each vector at each time point through 33-38 months. The number of unique barcodes obtained from granulocytes, B, and/or T cells at each time
point for the given vectors were counted (Figure 19B). At month one, a total of over 2000 unique barcoded contributing HSPC clones, derived from all three vectors, were detected in both ZJ41 or ZJ48. By two-three months, the number of barcoded contributing clones detected decreased, as we have previously reported, as contributions from waves of abundant short-term repopulating HSPC declined and were replaced by contributions from less abundant intermediate and long-term repopulating HSPC\textsuperscript{27,231}. Thereafter clone numbers contributing at later time points remained relatively stable, with some late gradual decline in both animals likely due to loss of intermediate-contributing clones. The cumulative number of barcodes retrieved plateaued from all three vectors in both animals by 6-12 months post-transplantation, indicating clone retrieval is efficient, and that few “dormant” HSPC were present that began contributing only at later timepoints (Figure 19C). As expected, application of different thresholds for inclusion of clones in the enumeration of clone numbers altered the numbers of clones retrieved, but did not change the overall patterns obtained over time, the critical issue for genotoxicity analysis (Chapter 2 Figure 13).

Although each animal had differences in the absolute number of clones tracked between the three vectors, similar total numbers of clones were tracked and assessed for clonal expansion: 7314 for SFFV, 5488 for MSCV and 4450 for EF1-\(\alpha\) when using a threshold of 2000 reads at one or more time points to be included in the master clone list. While obtaining similar marking levels for all three vectors within an animal would have been ideal, it is not crucial since the barcode permits tracking of contributions from each individual HSPC clone independently.
As an overall population measure of \textit{in vivo} clonal diversity, both the Gini-Simpson’s and Shannon’s diversity indexes were calculated. Both Simpson and Shannon indexes measure diversity by taking into account the richness (the number of barcodes present) and the evenness of the measured barcodes. However, the Gini-Simpson diversity index provides the option to calculate diversity solely on the evenness of the barcodes measured, independent of sample richness. Because vector marking was not equally balanced between the 3 different vectors in both animals, the Gini-Simpson diversity index provides a more applicable representation in the context of our data. As shown in Figure 19D, the Simpson diversity index for the vectors with reasonable HSPC engraftment in ZJ41 (SFFV) and ZJ48 (MSCV and EF1-α) maintained a high diversity index for all three lineages. For comparison, the more commonly-used Shannon diversity index was also calculated and demonstrated similar results (Figure 21) as the Gini-Simpson, where all three libraries with reasonable HSPC engraftment maintained a high-diversity index over time, with no evidence for progression to oligoclonality with any of the three vectors.

**Clonal stability of vectors by their top-contributing clones**

To identify clonal expansion, we evaluated the top-contributing clones for each of the vectors in each lineage, focusing on the largest clones at the time point of longest follow-up as representing potentially expanding clones, and tracking each of these clones over all time points. Due to short half-life and lack of peripheral expansion, granulocytes best represent ongoing hematopoiesis from HSPC in the context of this study, however we also performed these analyses on B cells and T cells. Figure 22 shows the
contributions mapped back over time from the top 20 clones detected in each lineage for each vector at the latest time point of 33 and 38 months respectively in the two animals. As shown in Figure 13 in chapter 2, these plots are virtually identical when applying higher or lower thresholds for inclusion of clones, indicating that setting of thresholds only impacts clones contributing very low levels at all time points. Long-term engrafting clones, as expected from our prior study of the dynamics of HSPC clonal reconstitution\textsuperscript{231}, first appeared and began replacing contributions from shorter-term progenitors between 2 and 12 months. The contributions from these long-term clones increased sharply at the time of appearance as expected, and then began to plateau. There was some relative increase in contributions for all these large long-term clones over time as short and intermediate-term progenitors eventually disappeared (see Figure 19B), a process that appeared slower in ZJ41. Stabilization of lymphoid production from long-term HSPC, particularly of T cells, was delayed compared to granulocytes, as previously reported\textsuperscript{27, 239}, likely due to thymic damage from irradiation. As seen in the stacked area plots for all three vectors, the relative contributions from these clones to granulocytes, T cells and B cells were then overall relatively stable for the duration of follow-up. No individual clone(s) showed a pattern of disproportionate outgrowth expected for genotoxic expansions, in contrast to the overall slow increases in relative contributions from long-term engrafting cells as compared to intermediate or short-term HSPC observed. No clone size was larger than 4.84%.

To assess comparative lineage contributions for these larger clones, we prepared heatmaps showing the contributions of each individual clone to each lineage over time for SFFV in ZJ41 and MSCV and EF1-\(\alpha\) in ZJ48. These heatmaps show that the 20
largest clones in each lineage were primarily balanced multipotent clones (Figure 23), and contributed at relatively similar levels to all lineages, in contrast to clones emerging and becoming malignant in the γRV clinical trials. In conclusion, we observed no patterns of clonal expansion of HSPC clones transduced with any of the three vectors consistent with genotoxicity.
Figures

A. ZJ41

Granulocytes
B cells
T cells

ZJ48

B. ZJ41

ZJ48

C. ZJ41

ZJ48

D.
Figure 19. Vector contributions and clonal diversity of EF1-α, MSCV, and SFFV barcoded vectors.

(A) The percent contribution of each vector (EF1-α, MSCV, SFFV) to total transduced hematopoiesis in granulocytes, B cells and T cells purified from the blood of animals ZJ41 and ZJ48 over time post-transplantation. The read number of each vector’s library ID in a sample over the total read numbers for all library IDs in the sample X 100 is plotted. (B) Number of unique barcodes for each vector (EF1-α, MSCV, SFFV) retrieved from all lineages (granulocytes, B cells and T cells) at a single time point for each vector in ZJ41 and ZJ48. A threshold of 2000 reads was applied to establish a master list of barcodes for each vector within each animal: to include a barcode on the master list it must have contributed 2000 reads at least one time point. Once established on the master list, this barcode would be counted even at time points where it contributed less than 2000 reads. The same barcode found contributing to more than one lineage was counted only once. (C) Cumulative number of unique barcodes retrieved for each vector (EF1-α, MSCV, SFFV) from all three lineages combined (granulocytes, B cells, and T cells) over time in ZJ41 and ZJ48. (D) Simpson’s diversity index for all lineages combined (granulocytes, B cells, and T cells) combined in ZJ41 and ZJ48.
<table>
<thead>
<tr>
<th>Mixture</th>
<th>Vector</th>
<th>Spike 1</th>
<th>Spike 2</th>
<th>Spike 3</th>
<th>Spike 4</th>
<th>Percent GFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mix 1</td>
<td>EF1α</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>4.41%</td>
</tr>
<tr>
<td></td>
<td>MSCV</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>4.41%</td>
</tr>
<tr>
<td></td>
<td>SFFV</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1.56%</td>
</tr>
<tr>
<td>Mix 2</td>
<td>EF1α</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>4.41%</td>
</tr>
<tr>
<td></td>
<td>MSCV</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>1.56%</td>
</tr>
<tr>
<td></td>
<td>SFFV</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1.56%</td>
</tr>
<tr>
<td>Mix 3</td>
<td>EF1α</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>1.56%</td>
</tr>
<tr>
<td></td>
<td>MSCV</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>1.56%</td>
</tr>
<tr>
<td></td>
<td>SFFV</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1.56%</td>
</tr>
<tr>
<td>Mix 4</td>
<td>EF1α</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1.56%</td>
</tr>
<tr>
<td></td>
<td>MSCV</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>1.56%</td>
</tr>
<tr>
<td></td>
<td>SFFV</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1.56%</td>
</tr>
</tbody>
</table>
Figure 20. Validation of barcode retrieval to assess clonal contributions.

Single-copy K562 cell clones from each of the three vectors (EF1-α, MSCV, SFFV) were mixed at different ratios based on cell number. Each of the mixtures were diluted with untransduced K562 cells to resemble marking levels from each of the animals, 4.41% and 1.59%, for ZJ48 and ZJ41 respectively. The mixtures represent n=4 for each of the mixtures and consisted of the following ratios for EF1-α, MSCV, SFFV; mix 1 (33%:33%:33%) with 4.41% GFP marking, mix 2 (38%:60%:3%) with 4.41% GFP marking, mix 3 (33%:33%:33%) with 1.59% GFP marking, and mix 4 (3%:0.7%:97%) with 1.59% GFP marking. DNA was extracted from each of the mixtures followed by low cycle PCR with primers bracketing the library IDs and barcodes for high-throughput Illumina sequencing, and data processing for retrieval and quantitation of the read fraction of each barcode linked to the vector-associated library ID. Pearson’s correlation was obtained between barcode retrieval quantitation and expected barcode frequency for each vector (r = 0.9961, 0.999, and 0.9994 for EF1-α, MSCV, and SFFV respectively). Each mix is represented by a color; mix1 (red), mix2 (blue), mix3 (green), and mix4 (black).
Figure 21. Diversity of barcoded vectors in hematopoiesis.
Shannon diversity index for each vector (EF1-α, MSCV, SFFV) over time for the combined lineages (granulocytes, B, T cells) in sJ41 and ZJ48.
Figure 22. Clonal stability of top contributing clones.
Stacked area plots delineating the percent contribution of the top 20 contributing clones in each lineage for each vector at the time of longest follow-up (38 months for ZJ41 and 33 months for ZJ48), tracked over all time points. Each of the top 20 clones is shown as a separate shaded region in the stack. Contributions from non-top 20 clones are shown as the remaining single solid color.
Figure 23. Heatmaps showing lineage contributions of top contributing clones.
Each row in the heatmap corresponds to a barcode and each column to a sample. The barcodes are ordered by unsupervised hierarchical clustering of the Euclidean distance between barcodes’ log fractional abundances in the samples. The color gradient depicts the log fractional contribution of individual barcodes to each sample. The contributions from the 20 most abundant barcodes identified in samples from every lineage and time point are plotted over all lineages and time points. The top 20 most abundant barcodes in an individual sample are designed with asterisks, and each column thus contains 20 asterisks. Each row (barcode) contains a minimum of one asterisk, because a barcode must have been a top 20 contributor to at least one sample to be included in the heatmap. (A) Top 20 EF1-α clones at each time point in each lineage mapped for animal ZJ48. (B) Top 20 MSCV clones at each time point in each lineage mapped for animal ZJ48. (C) Top 20 SFFV clones at each time point in each lineage mapped for animal ZJ41. Short-term lineage-restricted progenitors contributing at months 1-3 can be seen being replaced by long-term multilineage progenitors.
Chapter 5: Discussion

I developed a highly quantitative and sensitive technology to track the output of thousands of individual transduced HSPC in the rhesus macaque model, and applied this approach to assess potential genotoxicity of clinically-relevant LV vectors with a range of predicted genotoxicity, based on murine and in vitro models. Inclusion of an internal EF1-α promoter was predicted to have the least genotoxic effect, followed by an MSCV promoter/enhancer located within the lentiviral LTR, and finally a very strong SFFV promoter/enhancer within the LTR. In the animals included in this study, I did not achieve balanced initial levels of engrafted HSPC transduced with each of the three vectors in the individual animals, due to unknown factors resulting in variability during ex vivo transduction or transplantation, however, a total of at least 4450 clones were tracked for each of the three vectors between the two animals. Clonal diversity and stability were observed in the absence of any ongoing clonal expansions for 33- and 38-months follow-up to date in the two animals.

The utility of this type of preclinical in vivo NHP model for insertion-related clonal expansions depends on whether such expansions predict and generally precede later overt clinical transformation to MDS or leukemia. In trials for both chronic granulomatous disease (CGD) and Wiskott-Aldrich Syndrome (WAS) utilizing γ-retroviral vectors with very strong viral promoter/enhancers, myeloid transformations were preceded by a steady increase in contributions from clones harboring insertions activating MECOM or other proto-oncogenes, preceding overt MDS or AML by a year or more. Clinical transformation was accompanied by acquisition of “second hits” such as gross chromosomal losses or translocations. Acute T cell leukemias occurring in
γ-retroviral clinical trials have been less clearly preceded by long periods of detectable clonal expansions. The initial report from the French X-SCID trial detected an expanding clone two years before leukemia in one patient, however the British X-SCID and German WAS trials reported much shorter windows of detection prior to overt T-ALL. These and other studies using VIS retrieval analyses for detection of dominant clones sampled patients infrequently, demonstrated variable sensitivity levels, and were hampered by the need for complex statistical methodologies to estimate clonal contribution over time, thus clonal expansion prior to overt leukemia may have been missed. Our goal was to apply highly sensitive and quantitative barcoding based on the hypothesis that clonal expansion could be detected earlier and more reproducibly using this methodology.

Prior studies of NHP transplanted with HSPC transduced with integrating retroviral vectors and followed long-term suggested that these models closely mimic genotoxicity observed in human clinical trials. Analyses of VIS from human trials and non-human primate experiments have demonstrated close similarities in integration patterns between species for both RV and LV vectors. Both humans and rhesus macaques transplanted with γRV-transduced HSPC have developed clonal expansions linked to insertions in or near the MECOM (MDS1/EVI1) gene complex, primarily in myeloid cells. NHP transplanted with γRV-transduced CD34+ cells have developed clonal dominance proceeding to overt MDS/leukemia due to insertional mutagenesis. Although overt leukemia with RV transduction of HSPC in large animal models have been infrequent events, most studies involved follow-ups shorter than relevant for clinical trials, relatively low transduction efficiency (median of 5% gene
marking), and did not clearly provide data demonstrating whether overt malignancy was preceded in time by a long period of detectable clonal expansion. Our group previously demonstrated effective LV gene transfer to repopulating non-human primate HSPC and long-term polyclonal hematopoiesis, without evidence of marked clonal expansion, even utilizing a vector with an extremely strong internal viral promoter/enhancer\textsuperscript{246}.

Other assays used to assess genotoxicity \textit{in vitro} or in murine models have various strengths and limitations as summarized in the Introduction. However, these approaches suggested that the SFFV promoter/enhancer within a $\gamma$-retroviral backbone is extremely genotoxic compared to other RV constructs, correctly predicting the extremely high rate of leukemogenesis in the WAS trial utilizing this vector design as compared to other RV clinical gene therapy trials using weaker viral promoters/enhancers\textsuperscript{17, 175}. The initial study comparing RV vs LV with an internal strong promoter in a murine tumor-prone model demonstrated no leukemia induction over background for the LV, but a follow-up study did find detectable tumor acceleration when the strong SFFV viral promoter was included in the LV LTR, albeit 10 fold less than with the same SFFV design in an RV, informing our choice of this design as the potentially most genotoxic LV vector in the current NHP study\textsuperscript{15, 220}. \textit{In vitro} immortalization assays comparing different strengths and locations of promoter/enhancers also demonstrated genotoxicity of the SFFV cassette within an LV LTR or internally, compared to almost undetectable genotoxicity with internal weaker promoters such as PGK\textsuperscript{184}. A recent paper from Aranyossy and coworkers applied a competitive barcoding approach in a murine HSPC transplantation model, very similar to the approach in our rhesus macaque model, to study patterns of reconstitution and clonal dynamics with various RV and LV constructs\textsuperscript{247}. Unfortunately,
in this model, intermouse variation was very high, and long-term engraftment was generated from a very small number of clones in most animals, precluding genotoxicity comparisons.

Clinical trials utilizing LV transduction of HSPC in a total of over 30 patients to treat WAS\textsuperscript{167, 242, 248} metachromatic leukodystrophy (MLD)\textsuperscript{249}, adrenoleukodystrophy (ALD)\textsuperscript{8, 250}, and sickle cell anemia (SCA)\textsuperscript{251} have not demonstrated any consistent or persistent clonal expansions following retrieval of many hundreds of thousands of individual VIS, but tracking has been limited to relatively few time points, and has utilized semi-quantitative VIS retrieval. These trials used vectors with internal endogenous promoters or promoter/enhancers (WAS, SCA), weak internal constitutive promoters such as EF1-\(\alpha\) or PGK (MLD), or a strong internal retroviral promoter (ALD). However, a single thalassemia patient receiving HSPC transduced with a LV vector driving a hemoglobin gene from a strong erythroid-specific promoter/enhancer has shown prolonged but eventually transient clonal expansion of cells, linked to an LV insertion resulting in aberrant splicing and overexpression of the gene HMGA2\textsuperscript{9}. LV insertions have been shown to result in aberrant splicing and chimeric transcripts\textsuperscript{9, 15, 252}.

In sum, these clinical trials, and our similar findings in the NHP model using a more quantitative tracking methodology are encouraging in terms of the relative lack of genotoxicity of LV vectors, even those containing strong promoter/enhancer elements such as an LTR-inserted SFFV cassette. However, the risks associated with any integrating vector, including LV, are finite, and genotoxic risks must be considered in the context of disease severity and alternative therapies. All these trials also reported very encouraging clinical results regarding disease amelioration, suggesting that LV HSPC
gene therapy approaches will be applied to larger numbers of patients and disease indications in the near future, making safety predictions even more important. While our model did not detect differential genotoxic risk between the LV vectors utilized in this study, I believe the approach could be useful in the future to test novel vector platforms or pretransplant cell expansion approaches. Sensitivity might be enhanced, or clonal expansions accelerated by application of proliferative stress post-transplant, for instance with busulfan, which our lab has previously shown can induce clonal dominance 22.
References


Coelen, RJ, Jose, DG, and May, JT (1983). The effect of hexadimethrine bromide (polybrene) on the infection of the primate retroviruses SSV 1/SSAV 1 and BaEV. *Archives of virology* 75: 307-311.


115. Pedersen Finn, S, Pyrz, M, and Duch, M. Retroviral Replication.


154. H. Orkin, S, and G. Motulsky, A.


