blood

2010 115: 2610-2618
Prepublished online January 21, 2010;
doi:10.1182/blood-2009-06-229757

Cellular barcoding tool for clonal analysis in the hematopoietic system

Alice Gerrits, Brad Dykstra, Olga J. Kalmykowa, Karin Klauke, Evgenia Verovskaya, Mathilde J. C. Broekhuis, Gerald de Haan and Leonid V. Bystrykh

Blood (print ISSN 0006-4971, online ISSN 1528-0020), is published weekly
by the American Society of Hematology, 2021 L St, NW, Suite 900,
Washington DC 20036.

Cellular barcoding tool for clonal analysis in the hematopoietic system

Alice Gerrits,[1] Brad Dykstra,[1] Olga J. Kalmykowa,[1] Karin Klauke,[1] Evgenia Verovskaya,[1] Mathilde J. C. Broekhuis,[1] Gerald de Haan,[1] and Leonid V. Bystrykh[1]

[1]Department of Cell Biology, Section Stem Cell Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Clonal analysis is important for many areas of hematopoietic stem cell research, including in vitro cell expansion, gene therapy, and cancer progression and treatment. A common approach to measure clonality of retrovirally transduced cells is to perform integration site analysis using Southern blotting or polymerase chain reaction–based methods. Although these methods are useful in principle, they generally provide a low-resolution, biased, and incomplete assessment of clonality. To overcome those limitations, we labeled retroviral vectors with random sequence tags or "barcodes." On integration, each vector introduces a unique, identifiable, and heritable mark into the host cell genome, allowing the clonal progeny of each cell to be tracked over time. By coupling the barcoding method to a sequencing-based detection system, we could identify major and minor clones in 2 distinct cell culture systems in vitro and in a long-term transplantation setting. In addition, we demonstrate how clonal analysis can be complemented with transgene expression and integration site analysis. This cellular barcoding tool permits a simple, sensitive assessment of clonality and holds great promise for future gene therapy protocols in humans, and any other applications when clonal tracking is important. (*Blood*. 2010;115(13):2610-2618)

Introduction

Hematopoiesis involves the tightly coordinated process of blood cell production and is maintained by a small number of hematopoietic stem cells (HSCs). Resolving the exact number of HSCs that actively contributes to hematopoiesis at any given time and tracking the contribution of individual HSCs to each of the different blood cell lineages is important for a better understanding of both normal and malignant hematopoiesis. Performing such analyses in systems containing multiple HSC-derived clones has remained challenging because under normal circumstances the progeny of genetically identical HSCs are indistinguishable from each other. Therefore, analyzing the behavior of normal HSC clones has been limited to measuring their behavior on an individual basis by limiting dilution[1-3] or single purified cell[4-6] assays. However, clonal analysis has also been successful in hematologic disorders in which abnormal cells often descend from a single common ancestor and in which malignant clones could be identified by tracking unique genetic rearrangements, deletions or point mutations (reviewed in Gilliland et al[7]).

The realization that unique genetic mutations could be exploited for clonal analysis, combined with the discovery that retroviral vectors could be used to introduce new genetic material into HSCs,[8-10] led to the development of an alternative method to perform clonal analysis on hematopoietic cell populations. In recipients that received a transplant with retrovirally transduced HSCs, it was possible to distinguish between different HSC-derived clones by considering retroviral integration sites as unique clonal marks.[11,12] In this type of analysis, genomic DNA is fragmented with restrictases that cut within the vector and throughout the genome, resulting in fragments that consist of a small part of the vector and adjacent genomic DNA. Different integration sites thereby generate fragments of different lengths that can be detected by Southern blotting.[13-15] Alternatively, these fragments can be amplified with the use of polymerase chain reaction (PCR)–based approaches in which linear fragment ends are self-ligated (inverse PCR)[16] or in which primer tags are attached to the fragment ends (eg, ligation-mediated PCR and linear amplification–mediated PCR).[17,18] Although these retroviral-marking experiments did confirm the long-term self-renewal activity and multilineage differentiation potential of HSCs, the methods used have serious shortcomings and could therefore provide only a rough estimation of clonality. Southern blotting relies on restriction digestions; therefore, integrations without an appropriately positioned restriction site cannot be detected. Further, the method lacks sensitivity as well as discriminatory power to count the total number of integrated vectors. Although the PCR-based approaches are more sensitive, they still depend on restriction digestions and are prone to experimental error because of unequal PCR amplification caused by variations in template melting properties and relative differences in fragment length.

To overcome these limitations, we constructed a retroviral plasmid library that consisted of vectors containing a variable random sequence tag or "barcode." On stable chromosomal integration, this barcode introduces a unique, identifiable, and heritable mark into the genome, allowing the clonal progeny of the host cell to be tracked over time. Here, we couple the concept of cellular barcoding to a sequencing-based detection system and show the efficacy of the barcoding method to track clonal dynamics in 2 distinct cell culture systems in vitro and in a hematopoietic transplantation setting in vivo. We also show that this cellular barcoding method can be complemented with transgene expression

Submitted June 25, 2009; accepted December 30, 2009. Prepublished online as *Blood* First Edition paper, January 21, 2010; DOI 10.1182/blood-2009-06-229757.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

analysis and integration site analysis, providing additional layers of information.

Methods

Mice

C57Bl/6 (CD45.2) and DBA/2 mice were purchased from Harlan. C57Bl/6.SJL (CD45.1) mice were bred, and all animals were housed at the Central Animal Facility of the University of Groningen. All animal experiments were approved by the Groningen University Animal Care Committee.

Construction of barcoded vector libraries

The SF91 vector was kindly provided by Prof C. Baum (Hannover Medical School, Hannover, Germany), and the MIEV vector was kindly provided by Prof C. Jordan (University of Rochester, Rochester, NY). The barcode linker was created by annealing two 5′-phosphorylated primers (forward, 5′-GTACAAGTAANNATCNNGATSSAAANNGGTNNAACNNTGTAA-AACGACGGCCAGTGAG-3′; reverse, 5′-GATCCTCACTGGCCGTC-GTTTTACANNGTTNNACCNNTTTSSATCNNGATNNTTACTT-3′; Biolegio). Primers were dissolved in 0.5× ligation buffer (Fermentas) at a concentration of 100μM. After heating the mixture for 5 minutes at 95°C, the primers were allowed to anneal at gradually decreasing temperature. The annealed barcode linker was ligated into the *Bsr*GI-*Bam*HI site of the SF91 or MIEV vector (Figure 1A) at equimolar ratio. The resulting vector was transformed into 10-beta competent *Escherichia coli* cells (New England Biolabs) and grown overnight on LB plates supplemented with 50 μg/mL ampicillin (Sigma-Aldrich). Colonies were pooled by flushing plates with LB supplemented with 50 μg/mL ampicillin. After overnight culture, plasmid DNA was extracted with the use of the GenElute HP Plasmid Midiprep Kit (Sigma-Aldrich).

Validation of barcoded vector library

From an SF91 vector library created by combining approximately 800 bacterial clones, barcode sequences were amplified with primers directed against internal vector sequences (SF91 eGFP forward,: 5′-CTGCCCGA-CAACCACTACCTG-3′; SF91 WPRE forward, 5′-CCCTAAAAATGTA-AATGATTGCCCCACC-3′). The resulting mixture of equal-sized PCR products (ie, “crude” PCR product) was purified and sequenced by ServiceXS or StarSEQ with a primer directed against eGFP (eGFP forward, 5′-GCGATCACATGGTCCTGCTG-3′). The complexity of this SF91 library was validated by retransforming it into 10-beta competent *E coli* (New England Biolabs). Single colonies were picked and grown in LB supplemented with 50 μg/mL ampicillin and 30 μg/mL kanamycin (Sigma-Aldrich). Plasmid DNA from each bacterial clone was isolated with the use of the GeneJET Plasmid Miniprep Kit (Fermentas) and sequenced with the same primers as for the crude PCR sequencing. Pictogram (http://genes.mit.edu/pictogram.html) was used to visualize the combined monoclonal sequence traces.

Cell culture

The 32D (murine myeloid progenitor) cell line was cultured in RPMI-1640 medium with L-glutamine (PAA Laboratories) supplemented with 10% fetal bovine serum, penicillin and streptomycin (Invitrogen), and 10 ng/mL recombinant murine interleukin-3 (R&D Systems). Bone marrow (BM) cells were isolated from mice 4 days after intraperitoneal injection of 150 mg/kg 5-fluorouracil (Pharmachemie Haarlem) and cultured in StemSpan (StemCell Technologies) supplemented with 10% fetal bovine serum, 300 ng/mL polyethylene glycol–complexed recombinant rat stem cell factor (Amgen), 20 ng/mL recombinant murine interleukin-11 (R&D Systems), 1 ng/mL Flt3 ligand (Amgen), and penicillin and streptomycin.

Retroviral transduction procedure

The 32D cell line and primary BM cells were transduced as previously described.[19] Briefly, Phoenix ecotropic packaging cells were transfected with 1 μg of barcoded SF91 or MIEV vectors. Virus-containing supernatant harvested 24 and 48 hours later was used to transduce 2 × 10^5 32D cells or 7.5 × 10^5 BM cells per 3.5-cm well. Transduction efficiencies were determined by flow cytometry (FACSCalibur; Becton Dickinson) and were greater than 90% for 32D cells and 30% to 50% for BM cells. 32D cells were transduced with a barcoded SF91 vector library containing approximately 700 barcodes, BM cells for culture were transduced with a barcoded MIEV vector library containing approximately 700 barcodes, and BM cells for transplantation studies were transduced with either a low complexity (LC) or high complexity (HC) SF91 barcode library containing approximately 50 or approximately 800 barcodes, respectively.

Initiation of polyclonal and monoclonal cultures

Four days after the initial transduction, 2 × 10^4 eGFP$^+$ 32D cells were sorted with the use of a MoFlo flow cytometer (Beckman Coulter). The resulting polyclonal culture was maintained for 5 weeks; extensive barcode analysis was performed at weeks 2 and 5. Monoclonal cultures were initiated by sorting single eGFP$^+$ 32D cells into 96-well round-bottom plates and culturing those for approximately 3 additional weeks; 2 of these were selected for extensive barcode analysis. Polyclonal DBA/2 BM cultures were initiated with 8 × 10^5 transduced cells, without sorting for eGFP expression. These cultures were maintained for 5 weeks; extensive barcode analysis was performed for one selected culture at weeks 1 and 5. To generate monoclonal BM cultures, single eGFP$^+$Sca-1$^+$CD48$^-$EPCR$^+$ cells from day 7 transduction cultures of C57Bl/6 BM were sorted into 96-well round-bottom plates, and clones reaching population sizes of approximately 30 000 were selected for further analysis.[20]

Transplantation and cell purification

LC- and HC-barcoded C57Bl/6.SJL BM cells (transduction efficiencies 35% and 50%, respectively) were transplanted into lethally irradiated (9.5 Gy; IBL 637 ^{137}Cs γ-source; CIS Biointernational) C57Bl/6 mice without prior sorting for eGFP expression. Of both the LC- and the HC-barcoded cells 10^6 and 5 × 10^6 were transplanted into 2 recipients each. At 8, 17, and 33 weeks after transplantation, blood samples were taken from the retro-orbital plexus. Erythrocytes were lysed in ammonium chloride solution, and the remaining cells were stained with APC-conjugated antibody to Gr-1 (clone RB6-8C5) and PE/Cy7-conjugated antibody to CD3ϵ (clone 145-2C11; both from BD Biosciences Pharmingen). eGFP$^+$ granulocytes (Gr-1^{hi}, SSC^{hi}) and T cells (CD3$^+$, SSC^{lo}) were sorted with the use of a MoFlo flow cytometer and were stored in RNAlater (QIAGEN) for barcode analysis. One of the mice that received a transplant with 10^6 LC-barcoded cells (recipient 2) was excluded from further analysis, because the number of eGFP$^+$ cells that could be isolated was not sufficient for barcode analysis.

Barcode recovery and identification

Genomic DNA was extracted from cultured 32D and BM cells with the use of the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich). From the purified blood cells, genomic DNA was extracted with the use of the REDExtract-N-Amp Tissue PCR Kit (Sigma-Aldrich) followed by a clean-up step with the use of the Nucleospin Plasma XS kit (BioKe, Macherey-Nagel). Total RNA was isolated from cultured 32D cells with the use of the RNeasy Mini kit (QIAGEN), and standard cDNA synthesis (Invitrogen) was performed. Barcode sequences were amplified with primers directed against internal vector sequences (SF91 eGFP forward and SF91 WPRE reverse; MIEV eGFP forward, 5′-CTGCCCGACAACCAC-TACCTG-3′, and MIEV long terminal repeat reverse, 5′-CCAAACCTA-CAGGTGGGGTCTTTCATTC-3′). The mixture of equal-sized PCR products (ie, crude PCR product) was purified and sequenced as already described. To obtain individual barcodes for sequencing, the crude PCR products were subcloned into pCR4-TOPO-TA (Invitrogen) and transformed into 10-beta competent *E coli* cells (New England Biolabs).

Binomial distribution modeling

To estimate barcode frequencies we applied a binomial model. The binomial distribution determines the probability $P(k;n,p)$ of observing a particular barcode k times in n sequencing trials, given a probability p of observing this barcode in a single trial. Under the assumption that all barcodes have an equal chance of being selected, the probability p equals the inverse of the barcode complexity B (ie, $p = 1/B$) for all barcodes. Assuming independence of the individual barcode detections, the expected frequency of observing some barcode k times within n sequencing trails has been estimated by multiplying the number of barcodes B with $P(k;n,1/B)$. To estimate the barcode complexity B we minimized the mean square distance between the expected and the experimentally observed frequencies.

Integration site analysis

To simultaneously identify barcodes and retroviral integration sites, inverse PCR was performed. DNA (0.1 μg) from monoclonal MIEV-transduced BM cultures was digested with *Tat*I (Fermentas), followed by heat inactivation. DNA fragments were then diluted to a concentration of 0.7 ng/μL and ligated overnight with 10 U of T4 DNA ligase (Fermentas). After purification of circular fragments (BioKe), a first round of PCR amplification (1-PCR) was initiated with 9 ng of DNA and PCR master mix (Fermentas), under the following conditions: 94°C for 2 minutes; 30 cycles: 94°C for 1 minute, 59°C for 1 minute, 72°C for 3.5 minutes; 72°C for 10 minutes (forward,: 5′-ACCTGTAGGTTTGGCAAGCTAGCTTAAG-3′; reverse, 5′-CAAACCTACAGGTGGGGTCTTTCATTC-3′). A second round of PCR amplification was initiated with 2 μL of 1-PCR reaction, under the following conditions: 94°C for 2 minutes; 20 cycles: 94°C for 1 minute, 58°C for 1 minute, 72°C for 3 minutes; 72°C for 10 minutes (forward, 5′-GTTTGCATCCGAATCGTGGACTC-3′; reverse, 5′-CCTCACTGGCCGTCGTTTTAC-3′). After electrophoresis, observed fragments were cut from the gel, subcloned in pCR4-TOPO-TA (Invitrogen), and sequenced. Subsequently, retroviral integration sites were identified for 2 monoclonal cultures performing a Mouse BLAT Search on the University of California Santa Cruz Genome Browser (Assembly July 2007, NCBI Build 37).[21]

Results

Construction of barcoded vector library

Random sequence tags or barcodes were generated by annealing 2 synthetic oligonucleotides consisting of sets of random nucleotides separated by fixed triplets. The resulting double-stranded linker was then ligated into different types of retroviral vectors (Figure 1A). The fixed triplets in the barcode linker served a dual purpose. First, they enabled the annealing of forward and reverse primers into the barcode linker during initial construction of the barcoded vector library. Second, the fixed triplets facilitated the analysis of sequencing results by providing an internal standard to

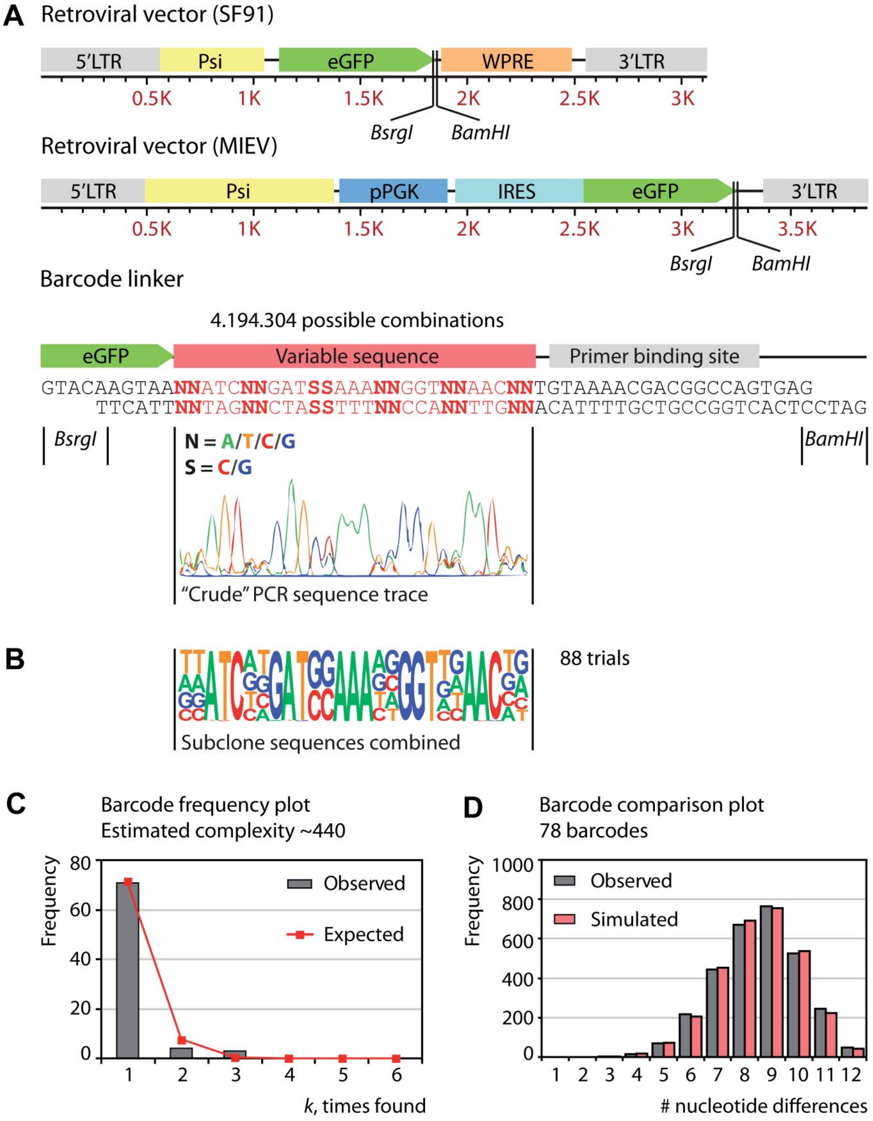

Figure 1. Construction and validation of barcoded vector library. (A) Insertion of the barcode linker into retroviral vectors SF91 and MIEV. The linker contains a variable sequence part that consists of pairs of degenerate nucleotides (N or S) flanked by fixed triplets. The crude PCR sequence trace of the resulting vector batch suggests the random insertion of nucleotides at N and S positions. (B) The HC vector library, created by combining approximately 800 bacterial clones, was retransformed into *Escherichia coli*. By combining the sequence traces of the 88 resulting clones, the crude PCR sequence trace could be reconstructed. (C) Distribution plot showing observed and expected barcode frequencies for the HC vector library; given that 88 sequencing trials were performed. Binomial distribution modeling estimates a barcode complexity of approximately 440. (D) Distribution plot showing the observed number of nucleotide differences between all 78 barcodes in the HC library by performing pairwise comparisons. Also shown is the distribution of the predicted number of differences based on the simulation of 78 random barcodes (average of 10 simulations). LTR indicates long terminal repeat; Psi, packaging signal; eGFP, enhanced green fluorescent protein; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element; pPGK, phosphoglycerate kinase promoter; IRES, internal ribosome entry site.

evaluate the quality of each sequence trace. The primer binding site was added to the barcode sequence so that barcode-positive and -negative clones could be easily distinguished with the use of PCR. Although in theory more than 4 million ($4^{10} \times 2^2$) possible linker variants could be generated, in practice the number of variants was restricted to the number of bacterial clones generated on transformation. By combining different numbers of bacterial clones, vector libraries of different complexities (ie, consisting of different numbers of barcodes) were generated. From each resulting vector library, barcode sequences were amplified with primers directed against internal vector sequences, after which the resulting mixture of equal-sized PCR products was sequenced. The resulting crude sequence traces suggested that a largely random mixture of sequence tags was generated, because N and S positions were found to be essentially equal in all 4 (A, T, C, and G) or 2 channels (C and G), respectively (Figure 1A). With the use of both restriction analysis and PCR, we could not detect unbarcoded vectors, demonstrating that the efficiency of vector barcoding was close to 100% (data not shown).

Validation of barcoded vector library

To validate the complexity of the prepared barcode libraries, we transformed a HC vector batch (created by combining ~ 800 bacterial clones) back into *E coli* and performed monoclonal sequence analysis. Of 88 sequencing trials, 78 represented unique barcodes. By combining the sequence traces of all 88 trials the crude PCR sequence trace could be reconstructed (Figure 1B), confirming that the insertion of nucleotides at each variable position was close to random. We next applied a binomial model to approximate the total complexity of the library (for details see "Binomial distribution modeling") and estimated it to be approximately 440 (Figure 1C). It should be noted that this prediction of the barcode complexity is only a rough estimate, because the number of sequencing trials was much lower than the complexity of the barcoded library. The accuracy of this prediction becomes progressively higher as the number of sequencing trials approaches or exceeds the number of barcodes in the library.

We also assessed the number of nucleotide differences between the 78 unique barcodes identified in the vector library and found that most barcodes differed in 9 of 12 variable nucleotides. When this was compared with data generated by simulating the generation of 78 random barcodes, the frequency distributions were almost identical (Figure 1D). Together, these data confirmed that the number of bacterial clones that are combined to generate vector libraries can be used as a rough estimate of the complexity of these libraries and that the generated vector libraries are indeed composed of vectors marked with randomly generated barcodes.

Clonal analysis of polyclonal 32D culture

As a first validation of the cellular barcoding method, we performed clonal analysis on an immortalized cell line that is known to be highly susceptible for retroviral transduction, with high transduction efficiencies and multiple retroviral integrations per cell. We transduced 32D cells with a barcoded vector library and initiated a polyclonal culture. After 2 weeks of culture, the crude PCR sequence trace showed a complex barcode signature that was comparable to the sequence trace of the barcoded vector library. Subclone sequencing confirmed the high degree of polyclonality in this culture, because 77 unique barcodes were identified of 105 sequencing trials. After 3 additional weeks of culture with semiweekly passaging, 68 unique barcodes were identified of 95 sequencing trials (Figure 2A). By applying a binomial model, we were able to estimate that the complexity of this culture decreased slightly from approximately 210 at week 2 to approximately 160 at week 5 (Figure 2B). Nevertheless, as would be expected for an immortalized cell line, the barcoded culture largely retained its polyclonal nature over time.

Enumeration of integration sites in 32D monoclones

We also initiated monoclonal cultures with single transduced 32D cells. Genomic DNA was isolated from 2 selected monoclonal cultures, and barcode sequences were amplified. The crude PCR sequence traces of these 2 cultures showed dissimilar levels of complexity, although both were noticeably simpler than the polyclonal culture (Figure 2C). Subclone sequencing was then performed to identify the underlying barcodes. The monoclonal culture with the simpler crude PCR signature (monoclone I) was found to contain 2 different barcodes, whereas the more complex (monoclone II) was found to contain 9 different barcodes (Figure 2D). Because these cultures were initiated with single cells, each identified barcode represents a unique retroviral insert. It could be determined with confidence that the barcodes detected represented the actual number of integrations, because the probabilities of having missed a third integration site in monoclone I or a tenth integration in monoclone II were only .035 and .022, respectively. Compared with traditional methods for integration site analysis, the barcoding method can determine the actual number of integrations in monoclonal populations with greater confidence. Overall, the crude PCR signatures proved to be a good representation of the actual underlying barcodes, because by combining the individual barcode sequence traces the crude PCR sequence trace could be reconstructed (Figure 2D).

Transgene expression analysis in 32D monoclones

Because the barcode tag is positioned immediately downstream of the eGFP coding sequence, each eGFP mRNA molecule contains the same barcode as the DNA from which it was transcribed. As such, the presence of barcodes at the RNA level permits the direct analysis of clone-dependent and/or integration site–dependent eGFP expression. To demonstrate this, the proportion of eGFP expression corresponding to each of the integrated vectors in the 2 monoclones was measured by isolating RNA and amplifying the expressed barcode sequences. This transgene expression analysis showed that in both monoclones eGFP molecules could be detected that were driven from each of the different integration sites (Figure 2E). Although 1 of 9 barcodes from monoclone II (marked by an asterisk) appeared to be overrepresented on the RNA level, this finding did not reach statistical significance ($P = .066$), in part because the number of subclones sequenced was relatively small compared with the number of integrations. Nevertheless, if a sufficient number of subclones are sequenced, the barcoding method will be able to identify clones and/or integrations that dominate expression within mixed cell populations.

Clonal analysis of primary BM cell culture

To further validate the cellular barcoding method, we performed clonal analysis on primary BM cultures that are less susceptible to retroviral transduction and would probably exhibit decreased clonal complexity over time. BM cells were transduced with a barcoded vector library, and the barcode compositions of the cultures were measured at 2 different points in time. One culture was selected for barcode analysis, and its clonal complexity was

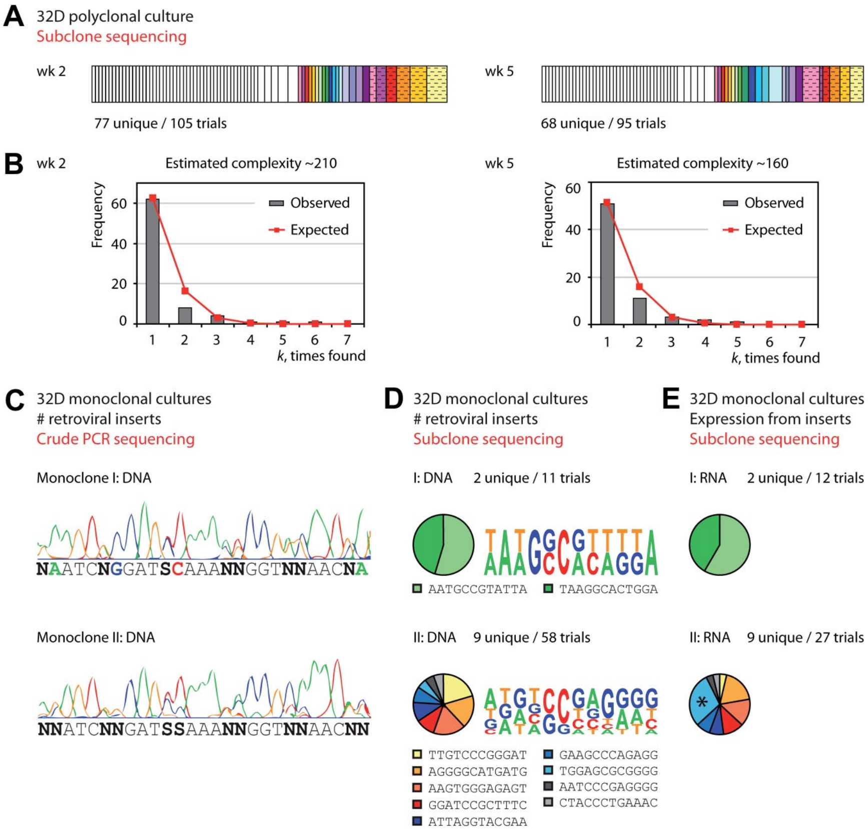

Figure 2. Barcode analysis of polyclonal and monoclonal 32D cell cultures. (A) Tracking the complexity of a polyclonal 32D cell culture over time. Subclone sequencing showed 77 unique barcodes of 105 sequencing trials after 2 weeks of culture, and 68 of 95 after 5 weeks of culture. White bars represent barcodes that were identified at 1 time point only, and colored/patterned bars represent barcodes that were found at both time points. (B) Distribution plots showing the observed and expected barcode frequencies for the 2 time points. Binomial distribution modeling predicts that the barcode complexity in this culture remained complex over time. (C) Crude PCR sequence traces of 2 monoclonal cultures provide an estimate of the total number of retroviral integrations per cell. The sequence trace for monoclone I shows that all variable positions in the barcode are restricted to 1 or 2 nucleotides, suggesting the presence of only 2 integration sites. The sequence trace for monoclone II is more complex, but 1 nucleotide is missing at several of the variable positions. (D) Subclone sequencing identified the unique barcodes that underlie the crude PCR sequence traces. Different barcodes are represented by different colors. Monoclone I contains 2 integrated retroviral vectors per cell, whereas monoclone II contains 9 integrated retroviral vectors per cell. Also shown are the actual barcode sequences identified for both monoclones. The sequence consensus pictograms show how the crude PCR sequence traces in panel B can be reconstructed by combining all identified barcodes in silico. (E) Gene expression analysis identified barcodes that were present at the RNA level. One of 9 barcodes in monoclone II (marked with an asterisk) appears to be overrepresented on the RNA level, but this finding does not reach statistical significance. Note that identical barcodes in panels D and E of this figure are represented with matching colors.

measured over time (Figure 3A). After 1 week of culture, the estimated barcode complexity was approximately 110, and the observed and expected barcode frequency distributions were almost identical (Figure 3B left). In contrast, after 4 additional weeks of culture one barcode was clearly found to dominate in the culture, being found in 36 of 50 sequencing trials. When this obviously overrepresented barcode was removed from the observed frequencies, the estimated complexity of the remaining culture was approximately 20, assuming that the remaining barcodes were equally distributed (Figure 3B right).

Integration site identification in BM monoclones

To show that clonal analysis can be complemented with integration site analysis, we purified single MIEV-transduced BM progenitor cells by fluorescence-activated cell sorting and initiated monoclonal cultures. Genomic DNA was isolated from several of these cultures, and inverse PCR was performed to simultaneously identify integration sites and corresponding barcodes of the inserted retroviral vector (2 examples are shown in Figure 3C).

Clonal analysis of repopulation dynamics in vivo

Having established the feasibility of using cellular barcoding as a tool for clonal analysis in vitro, we next used this method to track hematopoietic repopulation dynamics in vivo after transplantation of barcoded HSCs. First, to track general trends in hematopoietic reconstitution, we transduced BM cells with a relatively small number of barcodes (LC, ~ 50) and transplanted 5 million cells. Second, to directly track the clonal behavior of individual HSCs, we transduced BM cells with a relatively high number of barcodes (HC, ~ 800) and transplanted only 1 million cells. At 8, 17, and 33 weeks after transplantation, eGFP$^+$ granulocytes and T cells were sorted from the peripheral blood, and barcode analysis was performed by sequencing 19 to 30 subclones from each blood cell type at each time point (Figure 4).

In both mice, the barcode signature was dominated by relatively few barcodes at all time points, suggesting that within the eGFP$^+$ population only a few stem/progenitor cells were the main contributors of granulocytes and T cells. It should be noted that the LC-barcoded mouse probably contains multiple HSC-derived clones marked with the same barcode, and so it is likely that the actual number of clones in this mouse exceeds the number of clones in the HC-barcoded mouse.

In both recipients, the barcode signatures of the T cells isolated at 8 weeks after transplantation differed substantially from the signatures of T cells and granulocytes at all other time points. This might be explained by the long life span of T cells relative to granulocytes. T-lineage cells produced in the first few weeks by transduced multipotent or lineage-restricted progenitors might still result in T cells to be present in the blood at 8 weeks after transplantation. In contrast, most or all multipotent or myeloid-restricted progenitors targeted during the transduction procedure would not continue producing peripheral blood granulocytes for 8 weeks after transplantation. In addition, some barcodes were present in granulocytes at early time points, but appeared in T cells only at later time points. This might also reflect the difference in life span of both cell types, in combination with the difference in time required for these cells to undergo maturation and migrate to the peripheral blood.

Whereas some barcodes marked both granulocytes and T cells, others were identified in only 1 of the 2 cell types. It is possible that these barcodes may have distinguished multipotent HSCs that contributed to both myelopoiesis and lymphopoiesis from those that were myeloid-biased or lymphoid-biased in their lineage contributions.

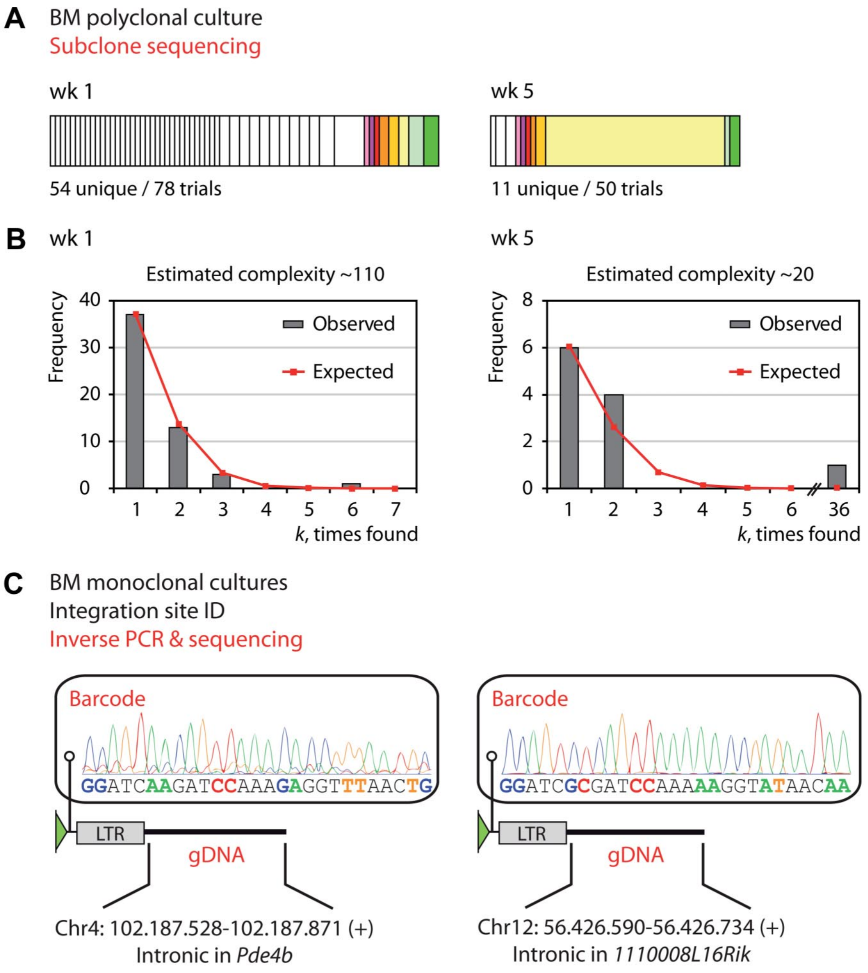

Figure 3. Barcode analysis of primary bone marrow cell cultures. (A) Tracking the clonal complexity of a primary BM culture over time. Subclone sequencing shows 54 unique barcodes of 78 sequencing trials after 1 week of culture, and only 11 of 50 after 5 weeks of culture. White bars represent barcodes that were identified at 1 time point only, and colored bars represent barcodes that were found at both time points. (B) Distribution plots showing the observed and expected barcode frequencies for the 2 time points. Binomial distribution modeling predicts that the barcode complexity in the culture markedly decreased over time. Note that the model for week 5 excludes the outlier that was found 36 times and is therefore calculated for 14 trials only. (C) Inverse PCR identifies the integration site and the corresponding barcode for 2 monoclonal cultures initiated with single transduced BM progenitor cells. LTR, long terminal repeat.

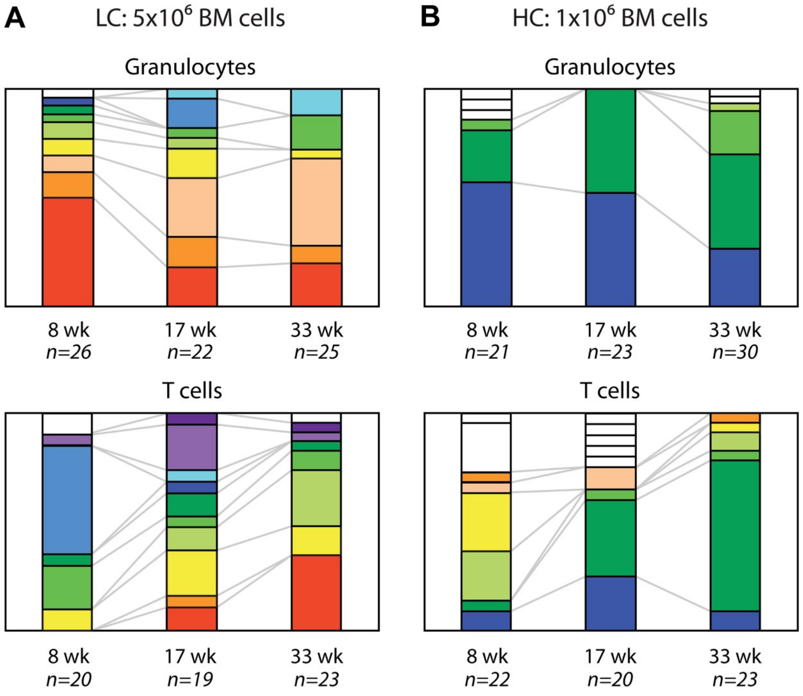

Figure 4. Clonal dynamics of in vivo hematopoiesis. Clonal composition of peripheral blood granulocytes and T cells over time in one recipient that received a transplant with 5 × 10^6 LC-barcoded BM cells (A) and a second recipient that received a transplant with 10^6 HC-barcoded cells (B). White bars represent barcodes that were found in only 1 cell type and time point, and colored bars represent barcodes that were found in both cell types and/or multiple time points.

Combined, these data suggest that cellular barcoding can be used for the simultaneous tracking of clones contributing to hematopoiesis. Depending on the complexity of the vector library and the number of subclones sequenced, the resolution of this technique could be increased to track minor clones with more confidence.

Clonal analysis of long-term HSC activity in vivo

Next, we used the barcoding method to measure the relative activity of individual HSCs in a polyclonal BM transplantation setting. As a surrogate measurement for stem cell activity, we analyzed the barcodes present in granulocytes, which have a short life span and must therefore be constantly replenished by the currently active HSC clone(s). eGFP$^+$ granulocytes were isolated at 33 weeks after transplantation from the peripheral blood of 3 recipients of LC-barcoded cells and 4 recipients of HC-barcoded cells, and 33 to 54 barcodes were sequenced from each sample (Figure 5A). In all recipients at least 1 major clone could be identified that was represented at a level higher than would be expected by chance because of random sampling error from an equally distributed population. Conversely, at least 1 and up to 8 barcodes were detected only once. This suggests that long-term hematopoiesis is for the most part maintained by a few HSC-derived clones, but that underneath these major clones a whole spectrum of minor clones may exist. Indeed, the fact that several barcodes were only seen once suggests that the number of identified minor clones is probably an underestimation of the actual number of minor clones in the population. By increasing the depth of sequencing a more accurate snapshot of the HSC compartment can be taken.

As would be expected, a much higher proportion of barcodes was found to be shared between the LC-barcoded mice (5 of 15 barcodes shared between 3 mice) than the HC-barcoded mice (1 of 37 barcodes shared between 4 mice; Figure 5B). This indicates that redundantly marked HSCs were indeed transplanted into the LC-barcoded mice, whereas the complexity of the HC library was sufficient to mark almost all HSCs uniquely.

Discussion

In this study, we demonstrate that cellular barcoding can be used as a powerful tool for clonal analysis in the hematopoietic system. This method overcomes many of the limitations of the Southern blot– and PCR-based techniques that were used for clonal analyses in the past, because the sequencing-based barcode detection system does not depend on restriction digestion or probe hybridization, and it is less prone to unequal PCR amplification. Therefore, we can identify with greater confidence both major and minor clones within complex cell populations in vitro and in vivo.

The data presented here provide important insight into the spectrum of active stem cell clones in a transplantation setting. In particular, our observations suggest that hematopoiesis at any given time appears to be dominated by a few HSC-derived clones, with additional contributions from a number of minor clones. Although the exact number of minor clones that are contributing at any given time remains undefined because of a lack of sequencing depth, it is still interesting to speculate about

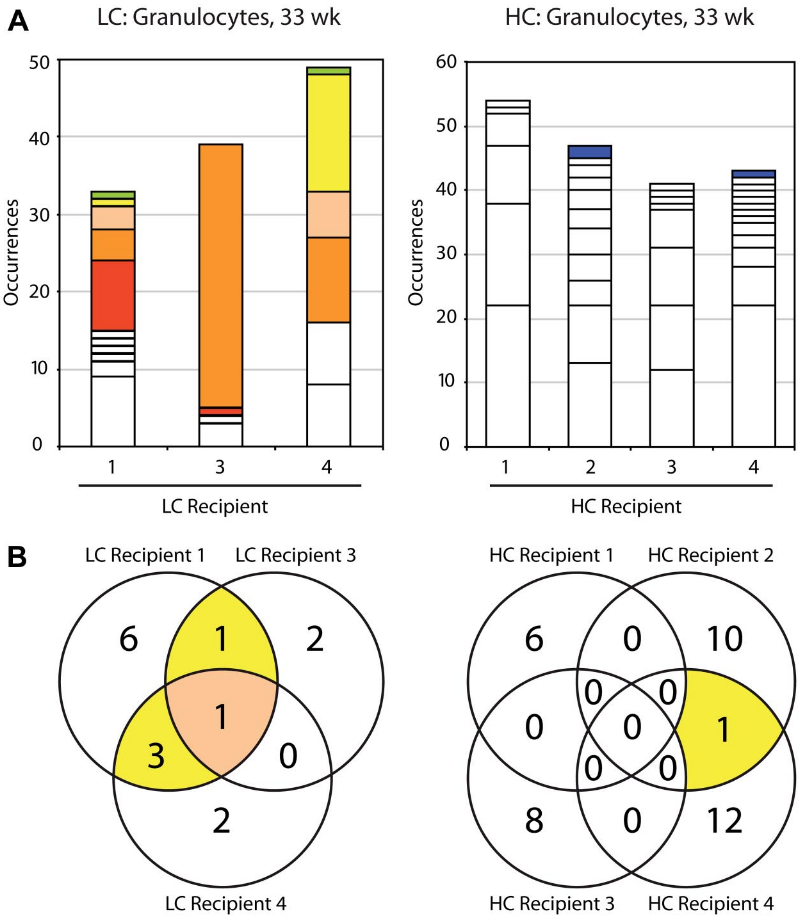

Figure 5. Barcode profiles of long-term HSC activity. (A) Clonal composition of granulocytes at 33 weeks after transplantation in 3 recipients of LC-barcoded cells and 4 recipients of HC-barcoded cells. White bars represent barcodes that were identified in 1 recipient only, and colored bars represent barcodes that were shared between recipients. (B) Venn diagrams showing the degree of redundant barcoding between multiple recipients receiving the same starting cells. Recipients 1 and 2 received a transplant with 1 million cells, recipients 3 and 4 received a transplant with 5 million cells. Note that a portion of the barcodes shown for LC recipient 4 and HC recipient 1 are also presented in Figure 4.

their biologic identity. Although their behavior could be due to simple stochastic variation, they might represent one or more functionally distinct HSC subtypes. A further implication is that current functional definitions for HSCs may need to be reconsidered, given that the output of some minor clones might not exceed the commonly used thresholds for repopulation.

Cellular barcoding can also be applied to help resolve other outstanding questions in HSC biology. For example, the concept of lineage-biased HSCs has been suggested,[6,22] but it is still controversial. By tracking barcodes in various cell types, lineage-biased clones could be directly confirmed in a competitive polyclonal setting. Similarly, the concept of dormant or hibernating HSCs has been suggested by several groups.[23-25] Cellular barcoding has the potential to measure the differentiated cell output from individual HSCs at different points in time, thereby enabling the identification of dormant or hibernating HSCs. It would be interesting to continue tracking these same clones after subjecting them to serial transplantation or other hematopoietic insult, to determine whether their properties can be reset or modified. In addition, different subpopulations of the primitive hematopoietic hierarchy could be purified by fluorescence-activated cell sorting from a recipient that received a transplant with barcoded HSCs, and the clonal relationships between the cell compartments could be analyzed. Compared with previous studies that used single-cell transplants, the amount of information that could be gleaned from one transplant recipient would be considerable. As well, such data would be useful to corroborate or further inform long-standing theoretical concepts and quantitative models on the clonal dynamics of HSCs.[26,27] Collectively, these applications of the cellular barcoding method have the potential to greatly increase our understanding of clonal dynamics in a BM transplantation setting.

In 2 recent reports, cellular barcoding was coupled to a high-throughput microarray-based detection system to study lineage relationships between T-cell subsets[28] and T-cell recruitment and expansion in response to infection.[29] Although this approach does allow clonal analysis to be performed on the population level, it requires a labor-intensive selection of barcodes and custom-made microarrays. In contrast, the approach described in this study can be more easily implemented. At the same time, a limitation of the approach presented here is that the resolution by which minor clones in a complex mixture can be identified is restricted by the number of subclones sequenced. To overcome this limitation, we are in the process of coupling the method to a high-throughput sequencing-based detection system, which will permit the simultaneous measurement of all uniquely marked clones at the population level.

Although retroviral vectors have already been used for decades as tools to introduce new genetic material into HSCs,[8-10] it has become clear only recently that insertional mutagenesis from the retroviral integrations themselves can trigger clonal expansion of HSCs.[18,30] Particularly in this context it is of great added value to barcode retroviral vectors in such a way that clonal analysis can be complemented with transgene expression and integration site analysis. To streamline this process in future experiments, we have now designed a next-generation vector in which the barcode is placed closer to the long terminal repeat. By first enriching for retroviral inserts, followed by high-throughput sequencing, we anticipate that integration sites, and their corresponding barcodes can be recovered simultaneously. At the same time, barcoded cDNA from the same cell populations could be sequenced. By normalizing the barcodes found in the cDNA to those found in the DNA, expression could be determined for each integration site in the entire population. Such multilevel analysis would be impossible using unbarcoded vectors.

Recently, a successful HSC gene therapy trial involving 2 patients with X-linked adrenoleukodystrophy was reported,[31] which has renewed interest and optimism in the gene therapy field.[32,33] It is exciting to consider how future applications of clinical gene therapy protocols would benefit from the use of barcoded vectors because this would allow a simple and sensitive detection of clonal dominance. For this purpose, the complexity of the library need not be particularly high, because the events triggering clonal dominance are rare. For example, a barcode-to-HSC ratio of 10:1 would already result in an average of approximately 90% uniquely marked cells (assuming equal representation of barcodes and disregarding the possibility of multiple integrations per HSC). Increasing this ratio to 20:1 or 40:1 would only result in an additional 5% or 7% uniquely marked cells, respectively. In the adrenoleukodystrophy gene therapy trial, patients were transplanted with approximately 250 transduced HSCs (assuming 25 kg of body mass, HSC frequency of 1 in 10^5 transfused cells, and 17% gene transfer efficiency into HSCs). A barcode library with a complexity of approximately 2500, which could be generated within a week in a regular laboratory, would uniquely mark approximately 90% of the transduced HSCs in these patients, with the remaining HSCs sharing a barcode with only 1 or 2 others. This would be easily sufficient to detect clonal dominance, even in its early stages. Because of the simple and flexible nature of the barcoding method, it is applicable to other viral (or even nonviral) gene delivery approaches and to systems other than hematopoietic. Therefore, we advocate the use of barcoded vectors in all future clinical gene therapy protocols.

Acknowledgments

We thank Selvi Durmus, Martha Ritsema, Ellen Weersing, and Jaring Schreuder for technical assistance; Geert Mesander and Henk Moes for assistance in cell sorting; and Verena Zuber and Tilo Buschmann for statistical advice.

This work was supported by the Dutch Cancer Society (RUG2007-3729); the Netherlands Genomics Initiative (Horizon, 050-71-055); the Netherlands Organization for Scientific Research (Top Talent, 021-002-084 to E.V.; VENI 916-86-009 to B.D.; and VICI 918-76-601 to G.d.H.); and by the European Community (Marie Curie RTN EUrythron, MRTN-CT-2004-005499; and Euro-System, 200720).

Authorship

Contribution: A.G. and B.D. analyzed data and wrote the manuscript; and L.V.B. conceptualized and designed the study with contributions from B.D. and G.d.H., performed the experiments with O.J.K., K.K., A.G., E.V., and M.J.C.B., analyzed data, and, with G.d.H., critically reviewed the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Gerald de Haan, Department of Cell Biology, Section Stem Cell Biology, University Medical Center Groningen, University of Groningen, Groningen 9700 AD, The Netherlands; e-mail: g.de.haan@med.umcg.nl.

References

1. Smith LG, Weissman IL, Heimfeld S. Clonal analysis of hematopoietic stem-cell differentiation in vivo. *Proc Natl Acad Sci U S A.* 1991;88(7):2788-2792.
2. Zhong RK, Astle CM, Harrison DE. Distinct developmental patterns of short-term and long-term functioning lymphoid and myeloid precursors defined by competitive limiting dilution analysis in vivo. *J Immunol.* 1996;157(1):138-145.
3. Cho RH, Muller-Sieburg CE. High frequency of long-term culture-initiating cells retain in vivo repopulation and self-renewal capacity. *Exp Hematol.* 2000;28(9):1080-1086.
4. Takano H, Ema H, Sudo K, Nakauchi H. Asymmetric division and lineage commitment at the level of hematopoietic stem cells: inference from differentiation in daughter cell and granddaughter cell pairs. *J Exp Med.* 2004;199(3):295-302.
5. Ema H, Sudo K, Seita J, et al. Quantification of self-renewal capacity in single hematopoietic stem cells from normal and Lnk-deficient mice. *Dev Cell.* 2005;8(6):907-914.
6. Dykstra B, Kent D, Bowie M, et al. Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell.* 2007;1(2):218-229.
7. Gilliland DG, Blanchard KL, Bunn HF. Clonality in acquired hematologic disorders. *Annu Rev Med.* 1991;42491-506.
8. Joyner A, Keller G, Phillips RA, Bernstein A. Retrovirus transfer of a bacterial gene into mouse haematopoietic progenitor cells. *Nature.* 1983;305(5934):556-558.
9. Williams DA, Lemischka IR, Nathan DG, Mulligan RC. Introduction of new genetic material into pluripotent haematopoietic stem cells of the mouse. *Nature.* 1984;310(5977):476-480.
10. Miller AD, Eckner RJ, Jolly DJ, Friedmann T, Verma IM. Expression of a retrovirus encoding human HPRT in mice. *Science.* 1984;225(4662):630-632.
11. Capel B, Hawley RG, Mintz B. Long- and short-lived murine hematopoietic stem cell clones individually identified with retroviral integration markers. *Blood.* 1990;75(12):2267-2270.
12. Jordan CT, Lemischka IR. Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes Dev.* 1990;4(2):220-232.
13. Dick JE, Magli MC, Huszar D, Phillips RA, Bernstein A. Introduction of a selectable gene into primitive stem cells capable of long-term reconstitution of the hemopoietic system of W/Wv mice. *Cell.* 1985;42(1):71-79.
14. Keller G, Paige C, Gilboa E, Wagner EF. Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. *Nature.* 1985;318(6042):149-154.
15. Lemischka IR, Raulet DH, Mulligan RC. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell.* 1986;45(6):917-927.
16. Nolta JA, Dao MA, Wells S, Smogorzewska EM, Kohn DB. Transduction of pluripotent human hematopoietic stem cells demonstrated by clonal analysis after engraftment in immune-deficient mice. *Proc Natl Acad Sci U S A.* 1996;93(6):2414-2419.
17. Schmidt M, Hoffmann G, Wissler M, et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. *Hum Gene Ther.* 2001;12(7):743-749.
18. Kustikova O, Fehse B, Modlich U, et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science.* 2005;308(5725):1171-1174.
19. Kamminga LM, Bystrykh LV, De Boer A, et al. The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood.* 2006;107(5):2170-2179.
20. Dykstra B, Schreuder J, Bystrykh L, De Haan G. Flow cytometric purification of primitive hematopoietic progenitors from bone marrow transduction cultures permits clonal analysis of retroviral integrations. Paper presented at Annual Meeting of the Society for Hematology and Stem Cells (ISEH), July 9-12, 2008. Boston, MA.
21. Kent, J. Mouse BLAT Search. http://genome.ucsc.edu/cgi-bin/hgBlat. Accessed June 9, 2009.
22. Muller-Sieburg CE, Cho RH, Thoman M, Adkins B, Sieburg HB. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood.* 2002;100(4):1302-1309.
23. Yamazaki S, Iwama A, Takayanagi S, et al. Cytokine signals modulated via lipid rafts mimic niche signals and induce hibernation in hematopoietic stem cells. *EMBO J.* 2006;25(15):3515-3523.
24. Wilson A, Laurenti E, Oser G, et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell.* 2008;135(6):1118-1129.
25. Foudi A, Hochedlinger K, Van Buren D, et al. Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nat Biotechnol.* 2009;27(1):84-90.
26. Roeder I, Horn K, Sieburg HB, et al. Characterization and quantification of clonal heterogeneity among hematopoietic stem cells: a model-based approach. *Blood.* 2008;112(13):4874-4883.
27. Roeder I, Loeffler M. A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity. *Exp Hematol.* 2002;30(8):853-861.
28. Schepers K, Swart E, Van Heijst JW, et al. Dissecting T cell lineage relationships by cellular barcoding. *J Exp Med.* 2008;205(10):2309-2318.
29. Van Heijst JW, Gerlach C, Swart E, et al. Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient. *Science.* 2009;325(5945):1265-1269.
30. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science.* 2003;302(5644):415-419.
31. Cartier N, Hacein-Bey-Abina S, Bartholomae CC, et al. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science.* 2009;326(5954):818-823.
32. Naldini L. Medicine. A comeback for gene therapy. *Science.* 2009;326(5954):805-806.
33. Gene therapy deserves a fresh chance [editorial]. *Nature.* 2009;461(7268):1173.