Chapter 23

Barcoded Vector Libraries and Retroviral or Lentiviral Barcoding of Hematopoietic Stem Cells

Leonid V. Bystrykh, Gerald de Haan, and Evgenia Verovskaya

Abstract

Cellular barcoding is a relatively recent technique aimed at clonal analysis of a proliferating cell population of any kind. The method was shown to be particularly successful in monitoring clonal contributions of hematopoietic stem cells (HSCs). An essential step of the method is retroviral or lentiviral labeling of the hematopoietic cells. The unique feature of the method is the generation of a vector library containing specific artificial DNA tags, generally known as barcodes. The library must satisfy multiple essential requirements. Importantly, considering the number of possible variations within the barcode sequence, the actual size of the barcoded vector library, and the number of clonogenic (stem) cells in the given experiment should be in ratios far from saturation. Excessive bias in barcodes frequencies must be avoided, and the library size must be assessed prior to the sequencing analysis. The final sequencing results must undergo statistical filtering. If all requirements are met, the method ensures profound sensitivity and accuracy for monitoring of the clonal fluctuations in a wide range of biological experiments.

Key words Barcode, Vector, DNA, Random, Stem cells, Clonality

1 Introduction

The problem of monitoring clonality in hematopoiesis is descending from early theoretical and experimental studies published first in the 1960s [1–3]. With the realization of the existence of hematopoietic stem cells and the clonal nature of hematopoiesis, questions of the extent of heterogeneity of HSCs were raised. Such heterogeneity includes the variation in the level of contribution to differentiated blood cells in time, and to each particular blood cell lineage [4].

Several approaches have been suggested which allow to obtain this kind of information. This included both theoretical efforts [5, 6] and experimental assays (in vitro assays, spleen colony forming assay, limiting dilution and single cell transplantation, chemical labeling, viral integration site analysis). At present it is widely accepted that HSCs comprise a small fraction of all hematopoietic

Kevin D. Bunting and Cheng-Kui Qu (eds.), *Hematopoietic Stem Cell Protocols*, Methods in Molecular Biology, vol. 1185, DOI 10.1007/978-1-4939-1133-2_23, © Springer Science+Business Media New York 2014

cells, which are quite heterogeneous in repopulating activity, and often are highly variable with respect to lineage contribution. Thus, determining the behavior of HSCs at the single cell level has become an important goal in stem cell biology.

To improve quantitative aspects of clonal monitoring, recently others and we have suggested an upgraded version of the retroviral or lentiviral HSCs labeling technique, using a specially barcoded vector library [7–9]. This approach considerably simplified the technique of the clonal monitoring, since it only required accurate DNA isolation from any cell sample of interest, and sequencing of the barcoded region [7, 8, 10–12]. Although straightforward in principle, the quality of the library and the data (sequence) analysis are crucial factors for the adequate interpretation of the collected results. Ignoring essential steps of the protocol might jeopardize the successful clonal counting. Multiple technical nuances of the cellular barcoding are the subject of the current report.

2 Materials

2.1 DNA

2.1.1 Retroviral Vectors

MIEV (originally obtained from Prof. Craig Jordan), 633 (Open Biosystems, USA), SF91 (originally obtained from Prof. Christopher Baum), and their derivatives. Lentiviral vector pGIPZ (Open Biosystems, USA) derivatives. All vectors must be adapted for barcoding in the following way: there must be at least 2 sticky unique restriction sides close to the end of CDS of the enhanced green fluorescent protein (eGFP) or turboGFP (tGFP) transcript, usually located upstream of the 3′ long terminal repeat (LTR). This vector locus is relatively nonfunctional and sequence variation in this locus should be selectively neutral.

2.1.2 Adapters

Custom-made barcode DNA adapters (forward and reverse complement).

2.1.3 Additional Vectors

For lentiviral transductions, packaging construct (pCMV Δ8.91), glycoprotein envelop plasmid (VSV-G) (both obtained from Dr. Hein Schepers [13]).

2.2 Enzymes

Restriction endonucleases: *BsrgI, BamHI, ClaI, BstBI*, DNA ligase (Thermo Fischer Scientific, USA; New England BioLabs, USA). Note that these enzymes are all supplied with buffers and instructions for use.

2.3 Commercial Kits

Mini-prep kit (Fermentas, USA), PCR cleanup kit (Roche, Switzerland), DNA isolation kit (Sigma-Aldrich, USA) or their analogues. All kits are supplied with detailed instructions, which must be followed.

2.4 Mammalian Cells

Packaging cell lines: for retroviral transduction—Phoenix-ECO cells (originally obtained from the Nolan lab); for lentiviral transduction—293T human embryonic kidney cells. Appropriate murine cells (for hematopoietic stem cell purposes we routinely use flow cytometry-purified stem cells).

2.5 Culture Media and Additional Chemicals

StemSpan culture medium (StemCell Technologies, Canada) supplemented with penicillin/streptomycin (P/S) and cytokines: interleukin-11 (R&D Systems, the USA), stem cell factor, Flt3 ligand (both Amgen, the USA), Dulbecco's modified Eagle's medium (DMEM) culture medium supplemented with P/S, fugene (Promega, the USA), retronectin (Takara, Japan), polybrene, 0.1 % gelatin solution (Sigma, USA).

2.6 Flow Cytometry

We perform cell sorting on MoFlo XDP and MoFlo Astrios (Beckman Coulter, the USA). Fluorofore-conjugated antibodies for cell sorting are acquired from Biolegend (the USA) and BD (the USA).

2.7 Disposables for Cell Culture and Transduction

T75 culture flasks, 6- and 12-well plates, 7.5 % BSA solution (PAA, for blocking purposes), 2 % BSA solution in PBS, 0.2 % BSA solution in PBS, 10 % FCS solution, +10 % FCS medium, Eppendorf tubes, sterile low protein binding filters (Millex HV sterile syringe filter with Durapore PVDF membrane, 0.45 μM, Millipore, the USA), 5 and 10 ml syringes, 25 ml syringes with luer lock, cryogenic vials (preferably with screw wire inside tube to prevent spills), 15 and 4.5 ml centrifuge tubes.

3 Methods

3.1 Barcode Linker

Two complementary barcode adapter sequences (forward and reverse complement) are ordered for synthesis at a company of choice (in our case BioLegio, the Netherlands). Adapters should carry the barcode sequence in its middle part. The barcode sequence consists of repeated random fragments interspersed with definite sequences. The barcode is usually flanked by 6–21 base long perfectly complementing sequences on both sides. At each end the adapter contains sticky overhangs compatible with the sticky ends of the cut vector (section below). For instance: if the vector is cut with *BsrgI-BamHI* enzymes, then the barcode should carry GTAC 5′-overhang for forward adapter and GATC for 5′-overhang of the reverse complement adapter. For this particular case the structure of the complete barcode adapter sequence will be as shown in Fig. 1a (from Gerrits et al. [8]). Note that this version preserves intact restriction sites. This original version of barcode was modified into a slightly reduced version (first NN pair was replaced by GG, Fig. 1b [8, 10]). Alternatively, the barcode shown in the figure was used in lentiviral vectors (Fig. 1c [10]).

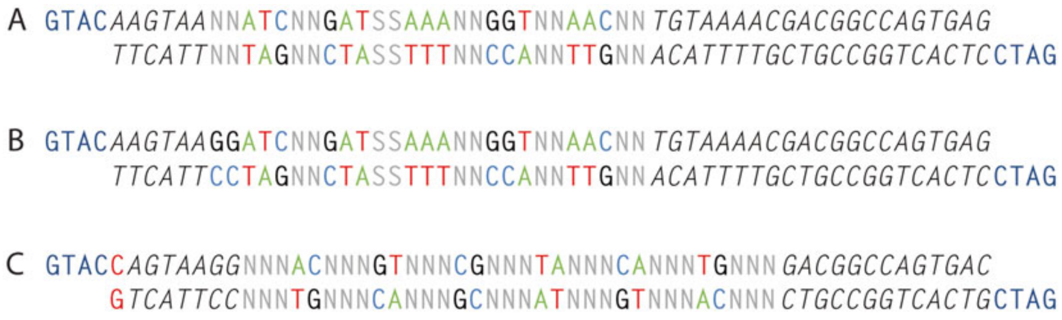
A GTACAAGTAANNATCNNGATSSAAANNGGTNNAACNN TGTAAAACGACGGCCAGTGAG
TTCATTNNTAGNNCTASSTTTNNCCANNTTGNN ACATTTTGCTGCCGGTCACTCCTAG
B GTACAAGTAAGGATCNNGATSSAAANNGGTNNAACNN TGTAAAACGACGGCCAGTGAG
TTCATTCCTAGNNCTASSTTTNNCCANNTTGNN ACATTTTGCTGCCGGTCACTCCTAG
C GTACCAGTAAGGNNNACNNNGTNNNCGNNNTANNNCANNNTGNNN GACGGCCAGTGAC
GTCATTCCNNNTGNNNCANNNGCNNNATNNNGTNNNACNNN CTGCCGGTCACTGCTAG

Fig. 1 Structure of barcode linkers. Details of linker design are described in the text

The barcoded adapters can be further modified according to the designer's preferences. Essentially, the number of N positions define the total number of possible combinations in the library, namely, $4^N \times 2^S$ (where N stands for any base, S is either G or C). Upon adapter ligation into the vector the latter version of the barcode adaptor eliminates original restriction sites (TGTACA became TGTACC), therefore once inserted the barcode cannot be removed by repeated restriction with *BsrGI-BamHI*.

3.2 Barcode Annealing

Synthesized barcode oligonucleotides usually arrive as a dry pellet. This is diluted with sterile pure water to a concentration 10 μM, and incubated at 37–40 °C for 1 h in a closed tube to dissolve it completely. For annealing forward and reverse complementary adaptors are mixed at a 1:1 ratio and 3 volumes of water are added. Further, a 10 mM Tris pH 7.4 buffer with 1–5 mM $MgCl_2$ (final concentration) and 50–100 mM NaCl is added. Alternatively, a ligase (no polyethylene glycol (PEG) added) or restriction buffer (for instance *EcoRI* buffer) can be added from 10× stock at dilution 20-fold. The buffer composition is approximate, the pH might vary in a range 7.3–7.8. Similarly, $MgCl_2$ (or other salts of Mg) should be present in sufficient concentrations, which can vary from 1 to 10 mM. Excessive amounts of ethylenediaminetetraacetic acid (EDTA) must be avoided since DNA is naturally a Mg^{2+} salt, and EDTA will chelate Mg^{2+}. PEG must be avoided because it will disturb DNA melting and force annealing at higher temperatures.

The mixture is placed in a closed tube on a dry thermostat with a sufficiently massive metal rack (to preserve heat for a few hours). The tube is heated to 90 °C, the thermostat is then turned down to the room temperature to allow for *passive* cooling of the mixture during 2–3 h. The mixture can be directly used for ligation to the vector of interest; alternatively aliquots of annealed adapter are stored in the freezer at −20 °C for later use.

3.3 Vector Preparation

Before barcoding the researcher must decide which vector to use and which locus of the vector is most suitable for barcoding. Our preferred locus is a region upstream of 3′LTR (usually flanked by eGFP/tGFP from other side). This site is nonfunctional, therefore integration of the barcode should not affect vector properties. The locus of interest must contain two unique sticky-end forming restriction sites. If the locus of interest does not reveal a suitable pair of restriction sites, they must be made by site-directed mutagenesis and/or custom made adaptors (in case only one available site is present, more sites can be added via adapter ligation).

Vector DNA is usually prepared from *E. coli* stocks, namely, single colony seed into 2 ml Luria broth (LB) medium supplemented with proper antibiotics, incubated with rotation overnight at 37 °C. Antibiotic concentrations can be found for instance online at the Roche LAB FAQs book. Plasmid DNA is isolated using available mini-prep or midi-prep kits (Thermo Scientific mini-prep kit, protocol according to the commercial provider of the kit) and stored in at −20 °C in small aliquots.

Routinely we use a pair of sticky end restriction endonucleases which cut the vector near the end of eGFP CDS upstream of 3′LTR. The MIEV vector is usually cut with *BsrgI* and *BamHI* [8]. Restriction is performed in a buffer compatible with both enzymes. Both NEB and Thermo Scientific offer a range of common buffers (NEB Double Digest Finder, Thermo DoubleDigest pages). A similar scheme is adapted for the properly modified SF91 vector [8], as well as for the pGIPZ modified vector [10]. Usually 20–40 μl of the standard mini-prep (0.2–0.4 μg/μl, 60–80 μl total) is taken for the restriction. Details of the restriction protocol (buffer, duration) must conform to recommendations of the enzymes provider, and can be found in the accompanying instructions and on the companies' Web sites (NEB or Thermo Scientific). After restriction is completed (usually 6–8 h is enough), the reaction is terminated by heat-inactivation at 80 °C for 20 min and chilled on ice. The cut vector can be either used directly or cleaned from reaction components using PCR cleanup kit (according to the recommendations of the company). In practice, 10 μl of the mini-prep (0.2–0.4 μg/μl) is sufficient to proceed with construction of the vector library. The rest of the prep is stored for later use.

3.4 Ligation

When barcode adapters and vectors have been prepared (barcodes annealed, vector cut with appropriate restriction enzymes generating mutually compatible sticky ends) the ligation can be carried out. Routinely, freshly cut vector is incubated at 65–85 °C for 20 min to inactivate restriction enzymes, diluted three times with sterile water, and barcode adapters are added to the linearized vector. Ligation buffer (usually 10× concentrated) and T4 DNA ligase is added to the mixture in amounts recommended by

Table 1
Calculation of linker to vector ratio

Component	Length, bp	μg/μl	Final 1:1 ratio	1:10 ratio
Plasmid DNA	9,000	0.1		
Linker	60	0.7		
Dilution	150	7	1,050	105

the company protocol. In case it is a regular DNA ligase, the initial temperature is set to 19 °C for 30 min, then gradually decreased to 17 °C and kept overnight. Fast ligase protocol can be performed as well, temperature conditions are defined by the enzyme provider. Essential for this step is mixing the adapters and vectors at a correct ratio. Unlike regular cloning purposes aimed on obtaining only few vector constructs of the desired configuration, vector barcoding is aimed on maximal efficiency of the barcode integration into the vector (i.e., preventing the formation of non-barcoded vectors, *see* **Note 1**). For this purpose an approximately 10–50-fold excess of the barcode adapter over the vector concentration is essential. In practice, the calculation can be done as follows. First, consider the length of the barcode and the vector. Second, consider concentrations of both (vector concentration can be measured, barcode adapters arrive with exact description how much μg is synthesized). From these two measurements we fill out the table above and find a final volumetric mixing ratio of the vector and barcoded linker. To illustrate, in our case we had the numbers specified in Table 1. This means that for a ratio 1:10 (vector to barcode) we take 20 μl of vector and 0.2 μl of a barcode adapter. Note that further excess of the barcode concentration (above indicated range) will inhibit efficiency of the ligation. It will lead to the drastic decrease of the successfully ligated product. In addition, the risk of a barcode concatemer forming will increase. Additionally, a titration of the vector to barcode ratio can be performed to ensure optimal conditions for ligation.

3.5 Post Ligation Handling

After ligation is completed, the sample is usually heat-inactivated at 80 °C for 20 min. Further it is cleaned up (we use Roche PCR cleanup kit). This is not a strictly obligatory step, yet cleaned DNA usually gives better efficiency for *E. coli* transformation. If the adapter is made in a way that eliminates the original restriction site (like lentiviral adapter above for *BsrGI* site), a control cut with corresponding enzyme (*BsrGI*) is recommended. This will further minimize the risk of the concomitant presence of unbarcoded vectors in the barcoded library.

3.6 E.coli Transformation

Competent *E. coli* cells (NEB, Life Sciences, or analogue) are transformed with a ligated vector mixture according to the protocol provided by the company. It is usually 5 μl of the vector DNA (0.05–0.1 μg/μl) per 50 μl of the *E. coli* cell suspension. At the end of the protocol the suspension is diluted to 300 μl by SOC medium. Cells are plated in variable dilutions (usually up to 50 μl per plate, totally 6 or more plates per one ligated vector prep) to ensure single colonies growth on petri plates. Note that barcodes consisting of completely variable uninterrupted N-mers (which are used by some groups) cannot be used in this protocol; such barcodes have a high risk of strand misalignment, which will be eventually modified by the DNA-repair system of *E. coli*.

3.7 Analysis of the Vector Barcoding Efficiency and a Library Size

At the optimal conditions all colonies on the plates should be well separated from each other, so one can count them reliably. This number is usually under 100–200 colonies per plate (at the vector concentration mentioned in the previous chapter). At this stage it is important to take 20–40 colonies from the plates, perform vector isolation (mini-prep from Thermo Scientific, Qiagen, or any other company) and test for the presence of barcodes. Insertion of the barcode in the vector causes a slight but detectable increase in the size of eGFP-LTR fragment of the vector. Using appropriate restriction enzymes around the barcoded site or PCR with barcode flanking primers will allow to detect the frequency of barcoded vector preps and to count the efficiency of barcode integration. Note that both unbarcoded vectors and aberrant barcodes will be detected at some frequency. It is inevitable that a fraction of the vectors will be not properly barcoded. It is important to ensure that the total percentage of such aberrant vectors is kept low (usually under 5 %). The quality of barcoding will determine the further strategy of preparing vector prep library. If the efficiency is high, colonies can be pooled in 5–20 colonies per flask and cultured as one midi- or maxi-prep. As a rule 1–2 ml of LB medium is used in liquid cultures per single colony in the prep. Culturing in pools saves labor and timing costs. Yet, the quality of such prep can be reduced by differences in barcode concentrations and by presence of non-barcoded vector. Additionally, the whole prep can be used only once, because preparing an *E. coli* stock from such a prep is not an option. Alternatively, each colony can be cultured separately and mini-prep is performed with each cultured clone individually. This approach requires considerable labor and time efforts. However, it provides the best quality result at the end since each clone is confirmed and stored separately, and all aberrant clones can be discarded. In this case *E. coli* stocks must be made for each barcoded vector. This will allow to re-culture and regenerate the very same barcode library whenever it is needed, so the library become endlessly reusable (*see* **Note 5**).

Table 2
Calculating the risk of identical barcodes and barcodes differing by 1 nucleotide based on the barcode structure and library size

Type	Barcode	Max complexity	Library size	Probability to pick the same barcode	Probability for barcodes to differ by 1 nt
1	NN…NN…SS…NN…NN…NN	4194304	100 500	0.000091 0.00008	0.001 0.004
2	NN…SS…NN…NN…NN	262144	100 500	0.001 0.0026	0.009 0.052
3	NNN..NNN..NNN..NNN.. NNN..NNN..NNN	4.398E+12	500 2,000	<0.0005 <0.0005	<0.0005 <0.0005

At the end, all isolated barcoded vector preps are mixed at equal amounts (equal μg DNA per each barcoded vector), which essentially becomes a barcoded vector library of known size/complexity. Importantly, the library size is counted at this stage for the first time (*see* **Note 4**). It is based on the actual number of colonies collected from the plates. Note that the library size should be far below the total number of possible combinations of the barcode sequence ($Library\ size << 4^N \times 2^S$). This important condition is required for two reasons:

1. To avoid the risk of obtaining the same barcode more than once (this will eventually decrease the real size of the library).
2. To make sure all barcodes are sequence-wise significantly different from each other [14], therefore the chance of having two barcodes differ in only one nucleotide will be negligible. This is important for sequence data filtering.

More exact solutions can be found using random simulations, Poisson and binomial distribution models. For detailed analysis (if not available in the lab) we recommend to refer to specialist in bioinformatics (*see* **Note 3**).

For the three types of barcodes we routinely calculate the risk of obtaining the same barcode twice or receiving two barcodes that differ by one nucleotide (with a minimal distance ($D_{\min}$) of 1). The calculations are shown in Table 2.

The statistics are obtained from a random simulation script in Python (taking into account the complexity of the barcode and the library size). Each parameter was tested in 50–500 simulations.

One can see from the table that barcode type 2 can be used for medium size libraries. If the size of a barcode type 2 library is increased to 500 barcodes, the library reaches its maximal size as the odds of including barcodes that are identical becomes critical.

Barcode type 1 suits well for a 500 barcodes library, while barcode type 3 can be used for much bigger libraries, which can count into the thousands.

Low saturation of the library is necessary for library validation by deep sequencing; however, it is not entirely critical in experiments with relatively low number of clones. If the number of clonally expanding cells in the biological experiment is 50–100 times below the size of the library, then problems of low barcode sequence distance or high redundancy in the library will be compensated by low sampling frequency in the real experiment. However, if the expected number of clonally expanding cells is close to or higher than the library size, then the size of the barcode library is critical. In both cases, it is beneficial to measure the number of clonal cells by at least two independent methods, one of those being barcoding.

3.8 Creating Viral Particles and Transduction of Hematopoietic Stem Cells

For retroviral vectors an ecotropic packaging cell line is routinely used. Excellent description of protocols is available at the Nolan lab page (http://www.stanford.edu/group/nolan/protocols/pro_helper_dep.html). Some variations in the method can be dictated by specifics of the cell line used. Our routine protocols have been described recently [8, 10] both for retroviral and lentiviral [13] barcoded vector libraries.

1. Day 1: Cell plating
   - Plate Phoenix-ECO cells: 2.5 × 10^5 cells per 1 well of 6-well plate in 2 ml DMEM + 10 % FCS. Incubate overnight at 37 °C, 5 % CO_2 to allow cells to attach.
2. Day 2: Transfection
   - Prepare transfection mix. Amounts needed for one well:

DMEM (no serum)	ad	100 μl
Fugene		3 μl
Barcoded vector construct		1 μg

   - Calculate the volume of the medium to be added, pipette in microcentrifuge tube.
   - Add dropwise fugene (do not touch walls of the tube), then vector preparation.
   - For mock transduction, add no vector.
   - Tick gently, no vortexing, allow complex formation for 15 min at room temperature.
   - Add 100 μl transfection mix dropwise to Phoenix-ECO cells, swirl really gently, incubate at least for 12 h (optimally 24 h) at 37 °C, 5 % CO_2.

3. Cell sorting

   - Collect the bone marrow cells and FACS-sort desired populations (in our case lineage negative Sca1$^+$ c-Kit$^+$ CD150$^+$ CD48$^-$ cells (LSK150$^+$48$^-$) and control LSK CD150$^-$ CD48$^-$ cells).
   - Collect cells directly from cell sorter into microcentrifuge tubes coated with 7.5 % BSA to avoid cell losses during transfer of cells to the plate.
   - Plate 10,000–20,000 sorted cells into 1 ml StemSpan + cytokines in 1 well of 12-well plate (block wells with 7.5 % BSA, remove it before plating).

4. Medium change

   - Carefully remove medium of Phoenix-ECO cells by suctioning.
   - Replace medium with 1.5 ml medium of eventual target cells (Stemspan, no serum should be added to prevent inactivation of the virus, no cytokines to avoid stress to Phoenix-ECO cells), incubate O/N 37 °C, 5 % CO_2.

5. Day 3: Coat wells with retronectin

   - Dilute retronectin stock (as received from a company) 40 times. This solution can be reused 4–5 times.
   - Add 1 ml retronectin solution to desired number of wells of 12-well plate, gently rotate to cover the whole bottom of the plate.
   - Leave at room temperature for 2 h.
   - Remove retronectin solution by suctioning.
   - Block retronectin by adding 1 ml of 2 % BSA solution in PBS. Leave for 30 min at room temperature.
   - Remove BSA solution (optional—wash with 0.2 % BSA solution in PBS before cell plating).

6. Virus harvest:

   - Very carefully collect supernatant from Phoenix-ECO cells using 5 or 10 ml syringe.
   - Pass through filter (Millex HV, low protein binding) to remove residual Phoenix-ECO cells.
   - Add 1 ml of PBS solution to Phoenix-ECO cells.

7. Cell transduction:

   - Collect LSK48$^-$150$^+$ and control (LSK CD48$^-$ CD150$^-$) cells that have been prestimulated in StemSpan + IL11+ SCF + Flt3L for 22–24 h (14 ml screw cap tubes).

Wash every well at least two times with 1 ml of PBS + 0.02 % BSA. Check under the microscope that all the cells were collected. Repeat the washing step if necessary.

- Spin cells down for 5 min, discard the supernatant.
- Add:
  - Viral supernatant in desired volume.
  - Polybrene solution.
  - Cytokine solution.
- Transfer cells to retronectin-coated wells, centrifuge for 15–25 min at 22 °C.
- Incubate O/N at 37 °C 5 % CO_2.

  Check fluorescence of Phoenix-ECO cells to control efficiency of transfection.

8. Day 4: Control whether transduction was successful
   - Collect transduced and mock-transduced control (LSK48$^-$150$^+$) cells in 14 ml tubes.
   - Add 10 % FCS up to 14 ml to wash (and partly inactivate) the virus and spin down.
   - Remove supernatant by suction.
   - Resuspend in PBS + 0.02 % or propidium iodide solution.
   - Bring on ice to FACS-analyzer and check fluorescence in GFP-channel. %GFP in control cells indicates whether transduction was successful. It is usually lower than in LSK CD150$^+$ CD48$^-$ cells, and additionally not all cells express GFP at this time point, so the value underestimates the real transduction frequency and has to be used only for indication.
   - If transduction was successful, collect LSK CD150$^+$ CD48$^-$ cells (as described above). Count the number of cells in a small aliquot using hemocytometer.
   - Leave a small aliquot in culture with StemSpan + 10 % FCS + cytokines to check final transduction efficiency.
9. Day 6–7: Check final transduction efficiency
   - Collect remaining cells, analyze by FACS.

3.9 Transduction Protocol with Lentiviral Vector Library

1. Day 1: Cell plating
   - Coat T75 flasks for 2 h with 0.1 % gelatin (4 °C). Use 3 flasks per transduction.
   - Plate 293T cells: 5 (1.5–5) × 10^6 cells per flask in 10 ml DMEM + 10 % FCS. Incubate O/N at 37 °C, 5 % CO_2.

2. Day 2: Fugene transfection
   - Prepare transfection mix. Amounts needed for one T75 flask:
     – Tube 1:

DMEM–FCS	100 µl
Packaging construct (pCMV Δ8.91)	3 µg
Glycoprotein envelop plasmid (VSV-G)	0.7 µg
Vector construct (pGIPZ derivative)	3 µg

     – Tube 2:

DMEM–FCS	400 µl
Fugene	21 µl

     (Add Fugene in medium, not against the edges of the tube).
   - Add content of tube 1 to tube 2, flick gentle and allow complex formation for 20 min at RT.
   - Add 500 µl transfection mix dropwise to 293T cells, swirl real gentle, incubate O/N 37 °C, 5 % CO_2.
3. Day 3: Medium change
   - Carefully remove medium of 293T cells by suctioning.
   - Replace medium with 5.0 ml medium of eventual target cells (e.g., HPGM, Stemspan, etc), incubate O/N 37 °C, 5 % CO_2.
4. Day 4: Virus harvest
   - Very carefully collect 5 ml medium from 293T cells into a 12 ml tube.
   - If there are a lot of floating cells, consider centrifugation first.
   - Pass over two filters (Millex HV, low protein binding) to remove residual 293T cells. Nb: first time, the filter might clog and snap, since a lot of 293T cells are also in this medium, hence the second filter step with new filter.
   - Freeze virus-containing supernatant in aliquots of 500 µl in cryotubes in −80 °C and use when necessary.

3.10 Critical Parameters

Number of target cells. To address the research question appropriately, it is important to approximate expected number of barcoded target cells in the population to be studied. For our experiments with HSCs, we performed limited dilution transplantations to estimate the stem cell frequency (*see* **Note 6**). Ex vivo culture and gene transfer can functionally influence transduced cells, and these factors have to be considered in experimental planning.

Barcode per cell ratio. A viral transduction event is a random Poisson process [15]. For any given transduction efficiency one can assess probabilities of single, double, triple etc. viral transduction per cell. While at low transduction efficiencies only one barcode is incorporated in most cells, number of vectors per cell might dramatically increase when reaching transduction efficiencies higher than 75 %. Researchers must be able to assess those probabilities experimentally and theoretically.

Limitations of blood sampling. For studying kinetics of clonal contribution to hematopoiesis, blood samples can be repeatedly taken from the same mouse. Since the blood volume should not exceed ~100 μl within intervals of 1–2 months, it creates some limitation for the resolution and statistical power of the experiment. If cells of different lineages are sorted from the blood sample the number of sorted GFP^+ cells defines the upper limit of the maximum number of barcodes that can be found in a sample. The sensitivity of barcode detection should be tested in samples containing known number of cells and barcodes. Work in clean environment free of high copy DNA handling is critical to prevent cross-contamination in the process of DNA extraction and preparation for PCR. It is a good practice to test all samples in duplicate to allow testing for consistency of reads if problems are detected. For DNA isolation standard commercial kits are routinely used. In our lab we combine RedExtract (Sigma) and Blood and tissue DNA isolation kit (Macherey-Nagel, Germany), which allow for extraction of DNA from cells in a range 1,000–1,000,000. For highly proliferative cells available in small numbers (HSCs), it can be recommended to perform monoclonal expansion in vitro to allow for robust barcode detection [10].

3.11 Barcode Sequencing and Primary Data Handling

3.11.1 Primary Processing

Genomic DNA is subjected to PCR with specifically tagged (indexed) primers, allowing for multiplexing of multiple samples in a single sequencing run. This obviously greatly reduces the cost of sample analysis. Details of primer tag design we use have been previously described [10, 14]. Unlike barcode sequences, primer tags for multiplexing are custom made 8–9 nt long DNA oligonucleotides added to the 5′ end of the PCR primers sequences. Importantly, tags must differ from each other by at least three bases, in other words they must conform to the minimal distance, $D_{\min} > 2$, to resist at least 1 sequencing error. Various barcoded DNA samples can be pooled together and sequenced in one lane. We use Illumima HiSeq machine for barcode detection. In this configuration 200–300 samples per run can be routinely sequenced and generate sufficient number of reads per sample.

When sequencing is completed a few essentials steps must be done.

1. Quality controls. In addition to machine filtering, we routinely discard all sequence reads with low quality reads (or N) if these occurs within the expected barcode region.
2. Data compression. All identical reads are compressed as one single sequence, and the frequency value (number of reads a particular sequence occurs) is kept next to the sequence.
3. Data demultiplexing. Experimental samples are separated on the basis of the exact match to the sample multiplexing tag and PCR primer fragment, 13 nt total length. Note that at this stage more sophisticated algorithms can be applied, allowing for instance single error correction and therefore improving recovery of reads.

3.11.2 Filtering of Sequencing Noise

When all steps mentioned above are completed a barcode sequence is routinely identified using matrix-scoring method (for our scripts in Python and Perl we used MOODS package [16]). Note that limiting the analysis to only exact matches to the barcode backbone sequence, contains the risk of missing slightly aberrant barcoded structures which occasionally occur (*see* **Note 2**). The initial set of reads is further reduced to the set of unique barcode sequences (with all frequencies counted and stored), ignoring all other errors/mutations outside of the barcode region. The first and the most powerful step of sequencing noise filtering is applying the rule that in a barcode data set, sorted in descending order by frequency, barcodes that occur less frequently and that have a $D_{\min} = 1$ with respect to the most abundant barcodes, must result from PCR and/or reading errors, and are removed. Failure to remove these spurious barcodes results in greatly inflated clonal counts. Note that this rule can be applied if barcode library size and clonality of the analyzed population make a probability of any pair of barcodes in the population on such distance negligible (see chapter barcode library construction above). Authors must be able to assess those risks or consult a statistician.

Some additional noise filters to consider are:

Number of reads per cell. Suppose 30,000 barcoded cells were sorted, from these cells DNA was extracted and PCR of the barcode sequence performed. Subsequent sequencing returned 100,000 reads per whole set. These reads will be unequally distributed among different unique barcodes. Barcodes with frequencies less than 1/30,000 of total reads, in this case 100,000/30,000 = 3.3, will be discarded since their relative representation is less than one barcode per cell, per definition indicating presence of technical noise in the sample.

Biological significance. Suppose in a time series a barcode has been detected which never reached a frequency above 0.5 % of the population. Even though this barcode might be real, its contribution to overall blood regeneration remains of low biological significance (keeping in mind the routine standards for acceptable limits for HCS contribution). On this basis this barcode could be rejected as it non-robustly contributes to blood cell formation.

Consistency of barcode spectra in time series (*outliers*). Our experience shows that the great majority of clones can be consistently detected in a time series, which allows concluding that overall blood contribution is stable, although gradually drifting in a time [8, 10]. If this assumption is accepted, incidental outliers in time series are suspect to result from technical errors or contaminations. It is advised that samples organized as time series are checked for consistency and all outliers should be taken with caution. Performing barcode analysis in duplicate helps resolving this issue.

4 Notes

1. Synthesis of the barcoded vector via a ligation protocol suffers from several imperfections, such as the occasional truncation of the backbone, concatenation of barcode sequences (when used in excess), and some fraction of vectors remaining unbarcoded. Thorough experimental optimization of the protocol should keep those side products at minimum.
2. Retrieving barcoded sequences must allow for detection of those aberrant vectors, and they must be considered as real and included in the analysis.
3. Increasing the barcode library size is a labor intensive process. Therefore, researchers must consult specialist in statistics on a reasonable library size for planned experiments.
4. Library size cannot be determined solely on the basis of sequencing data, because sequencing alone does not discriminate between real barcodes and false barcodes which are generated due to PCR/sequencing/reading errors.
5. Bias in barcoded vector representation inevitably reduces effective library size, and therefore must be kept at minimum.
6. It is desirable to have a parallel, independent assessment of the initial frequency of the barcoded clonogenic cells. Counting clones using barcodes is probabilistic by nature. It relies on specific probabilities of D_{min} values in the barcoded vector libraries; also it is used in combination with technical thresholds and distinct levels of biological significance. This creates certain risks of obtaining false positive and false negative types of errors. Those risks must be approached using proper statistical analysis.

Acknowledgements

This study was supported by the Netherlands Organization for Scientific Research (TopTalent grant to E.V.), and the Netherlands Institute for Regenerative Medicine (NIRM).

References

1. Till JE, McCulloch EA (1961) A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiat Res 175: 145–149
2. Becker AJ, McCulloch EA, Till JE (1963) Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. Nature 197: 452–454
3. Kay HE (1965) How many cell-generations? Lancet 2:418–419
4. Copley MR, Beer PA, Eaves CJ (2012) Hematopoietic stem cell heterogeneity takes center stage. Cell Stem Cell 10:690–697
5. Harrison DE, Astle CM, Lerner C (1988) Number and continuous proliferative pattern of transplanted primitive immunohematopoietic stem cells. Proc Natl Acad Sci U S A 85: 822–826
6. Abkowitz JL, Linenberger ML, Newton MA et al (1990) Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. Proc Natl Acad Sci U S A 87:9062–9066
7. Lu R, Neff NF, Quake SR et al (2011) Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat Biotechnol 29: 928–933
8. Gerrits A, Dykstra B, Kalmykowa OJ et al (2010) Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood 115:2610–2618
9. Schepers K, Swart E, van Heijst JW et al (2008) Dissecting T cell lineage relationships by cellular barcoding. J Exp Med 205:2309–2318
10. Verovskaya E, Broekhuis MJ, Zwart E et al (2013) Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. Blood 122:523–532
11. Grosselin J, Sii-Felice K, Payen E et al (2013) Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. Stem Cells 31:2162–2171
12. Naik SH, Perie L, Swart E et al (2013) Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature 496:229–232
13. Schepers H, van Gosliga D, Wierenga AT et al (2007) STAT5 is required for long-term maintenance of normal and leukemic human stem/progenitor cells. Blood 110:2880–2888
14. Bystrykh LV (2012) Generalized DNA barcode design based on Hamming codes. PLoS One 7:e36852
15. Fehse B, Kustikova OS, Bubenheim M et al (2004) Pois(s)on – it's a question of dose. Gene Ther 11:879–881
16. Korhonen J, Martinmaki P, Pizzi C et al (2009) MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics 25: 3181–3182