HHS Public Access

Author manuscript

Nat Protoc. Author manuscript; available in PMC 2020 October 01.

Published in final edited form as:

Nat Protoc. 2020 April ; 15(4): 1436–1458. doi:10.1038/s41596-019-0290-z.

Clonal tracking using embedded viral barcoding and high-throughput sequencing

Charles Bramlett[1,2], **Du Jiang**[1,2], **Anna Nogalska**[1], **Jiya Eerdeng**[1], **Jorge Contreras**[1], **Rong Lu**[1,✉]

[1]Eli and Edythe Broad CIRM Center for Regenerative Medicine and Stem Cell Research, Keck School of Medicine, University of Southern California, Los Angeles, Los Angeles, CA, USA

Abstract

Embedded viral barcoding in combination with high-throughput sequencing is a powerful technology with which to track single-cell clones. It can provide clonal-level insights into cellular proliferation, development, differentiation, migration, and treatment efficacy. Here, we present a detailed protocol for a viral barcoding procedure that includes the creation of barcode libraries, the viral delivery of barcodes, the recovery of barcodes, and the computational analysis of barcode sequencing data. The entire procedure can be completed within a few weeks. This barcoding method requires cells to be susceptible to viral transduction. It provides high sensitivity and throughput, and enables precise quantification of cellular progeny. It is cost efficient and does not require any advanced skills. It can also be easily adapted to many types of applications, including both in vitro and in vivo experiments.

Introduction

A cell is a basic unit of biological systems. It can divide to produce progeny cells, forming a cell clone. Tracking of cell clones over time and through space can provide critical insights into cellular behavior. As genetic material is conserved during cell division, a cell can be marked and tracked when unique genetic information is inserted into its genomic DNA, a procedure called genetic barcoding. Because genetic barcodes are inherited by all progeny cells, the abundance of each barcode in a cellular population is proportional to the number of cells derived from the original barcoded cell. In conjunction with high-throughput

✉Correspondence and requests for materials should be addressed to R.L. ronglu@usc.edu.

[2]These authors contributed equally: Charles Bramlett, Du Jiang.

Author contributions

R.L. conceived and developed the protocol. C.B., D.J., J.C., and A.N. optimized the barcode extraction protocol. D.J. and J.E. improved the Python code for analyzing high-throughput sequencing data. C.B., D.J., and R.L. prepared the manuscript. J.C. and A.N. provided assistance in manuscript text preparation.

Competing interests

The authors declare no competing interests.

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-019-0290-z.

Peer review information *Nature Protocols* thanks Leila Perie and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

sequencing, genetic barcoding is a powerful technique that enables tracking of clonal behaviors in a high-throughput manner[1].

The original approach for genetic barcoding used retroviral insertion sites to mark individual cell clones and Southern blot to analyze the results[2–4]. Later, synthetic random DNA barcodes were used in conjunction with microarrays[5]. Recently, we and others developed viral genetic barcodes that mark cells using synthetic DNA segments embedded within a viral construct that can be easily quantified by high-throughput sequencing[6–10] (Fig. 1). The embedded viral barcoding technology provides high sensitivity and throughput, and enables precise quantification of cellular progeny[11–14]. The high-throughput nature of the improved technique reduces the impact of experimental noise associated with single-cell measurements by greatly increasing the number of measurements. The high sensitivity of barcode recovery provided by a single PCR step enables the identification of small changes in barcode abundance. In addition, embedded viral barcoding generates data with single-cell resolution through the use of randomized barcodes and does not involve the handling of single cells at any point. For simplicity, the term “barcoding” will refer to embedded viral barcoding throughout, unless otherwise stated.

The barcoding method has been utilized and improved by several groups[6,15–18]. However, there are no standards in the field for the generation and analysis of barcode data[6]. Here, we provide a detailed and easy-to-replicate protocol for generating and implementing genetic barcodes for cellular tracking studies. Since its first publication[1], our protocol has been substantially optimized to improve its sensitivity and detection limits[11–14]. These improvements primarily involve upgraded data analysis algorithms and experimental procedures for barcode recovery. Here, we outline the protocol in a general way so that it can be adapted to many types of applications, including both in vitro and in vivo experiments. Our protocol enables new users to easily set up barcoding at a low cost by creating their own barcode libraries and performing computational analysis in their own labs.

Applications of the method

Barcoding can be applied to any cells that are susceptible to lentivirus infection[17–20]. It generates clonal behavior information that is important for many fields of research. For example, it can identify the cell of origin during development and track the differentiation patterns of stem cells. Using this approach, we have identified a distinct lineage origin for natural killer cells in a rhesus macaque transplantation model[13]. The high-throughput nature of this technology enables comparison of many individual cells simultaneously and provides a direct assay of cellular heterogeneity. For example, we have used barcoding to show how hematopoietic stem cells heterogeneously differentiate after transplantation in mice[11,12,14].

Barcoding can also be used to study diseases, particularly those that originate from rare cells such as cancer[19,21,22]. For example, barcoding can help reveal the cellular origins of cancer genesis, relapse, and metastasis. It can also reveal the heterogeneous responses of cancer cells to treatment. These studies require ex vivo barcoding of candidate cells, typically, samples from patients or animal models. Tracking can then be performed in vitro or in animal models.

Barcoding has also been applied to facilitate gene and drug screens. For example, barcoding has been used in CRISPR screens, in which the guide RNAs (gRNAs) serve as genetic barcodes[23,24]. Although these studies typically do not require single-cell resolution, genetic barcodes still provide high throughput and tremendous cost savings. Moreover, barcoding can provide further insights regarding cellular heterogeneity for these screens.

Comparisons with other methods

The conventional strategy for studying clonal behaviors is simply to track a single cell at a time, for example, in single-cell transplantation and single-cell culture[25–27]. Although these approaches do not require viral transduction, they are labor intensive and cost prohibitive for most applications. To increase throughput and reduce cost, fluorescent proteins, either singly or in combination, have been used to mark clonal identities[28–31]. However, the number of fluorescent colors is small, limiting the number of cells that can be confidently tracked at the clonal level. By contrast, the synthetic DNA segments that we use have virtually unlimited variations, enabling thousands of clones to be tracked with a quantifiably high degree of accuracy in the clonal-level labeling. It is also cost effective because our viral library designs enable multiple samples to be sequenced together.

Other techniques have tried to overcome the limited number of fluorescent colors by using viral insertion sites paired with linear amplification-mediated PCR (LAM-PCR)[32–35] and quantitative shearing linear amplification PCR[36]. These methods are still in use for human studies, particularly those that involve gene therapy. However, the difficulty of recovering genomic insertion sites precludes obtaining the quantitative data needed for most clonal tracking questions. By contrast, the synthetic DNA segments used in barcoding are easy to recover and produce highly quantitative results. The design of our genetic barcodes enables their recovery using a single PCR step, during which primers that are needed for downstream sequencing are simultaneously incorporated. This simple and elegant approach greatly reduces experimental noise during barcode recovery and produces quantitative and reproducible data. In our day-to-day experiments, replicate samples are highly consistent (Fig. 2), and barcode quantification is directly proportional to donor cell doses[12], demonstrating the high fidelity of our quantitative measurements when applied to in vivo experiments.

The barcoding procedure for in vivo studies requires cell transduction and transplantation. Attempts to eliminate the transduction and transplantation steps have been tried in several approaches in which cells are engineered with transposon, Polylox or CRISPR–Cas9 technologies. Transposon-based methods temporarily express a transposase to activate the mobilization of a DNA segment, called a transposon, which is randomly inserted into the genome to label individual cells[37]. Similar to the viral insertion site detection technique, this approach suffers from poor quantification because of technical difficulties in recovering the genomic insertion sites.

The Polylox-based system uses a series of unique loxP sites embedded in the genome that are excised randomly upon exposure to *Cre* recombinase[38]. This approach has been commonly used to generate fluorescent protein combinations[30]. The CRISPR–Cas9-based system edits genomic DNA, or synthetic DNA segments embedded in the genome, with the

help of gRNA[39,40]. Both Polylox and CRISPR–Cas9 rely on the assumption that the DNA recombination is random, which is not entirely true for either system[39–42]. Meanwhile, the viral barcoding method suffers from random multiple labeling as well.

Taken together, the transposon, Polylox, and CRISPR–Cas9 systems enable endogenous activation of cellular labeling. Compared to viral barcoding, these approaches do not require cell isolation, culture, transduction, or transplantation, thus enabling the study of native cellular behaviors in addition to ex vivo and transplantation-mediated studies. Moreover, tissue-specific promoters can be implemented to address tissue-specific questions. However, these approaches can be challenging for cell types that cannot be defined by a single promoter.

To overcome the transgenic requirement, new retrospective methods for clonal tracking take advantage of naturally occurring mutations. These methods rely on rare somatic mutations to reconstruct the lineage relationships between individual cells[43–46]. Because neutral mutations occur during cell division in a seemingly random process, these methods link cells to one another when they share common mutations. However, these methods require enough cell divisions to accumulate rare mutations and cannot track cells that do not carry any mutations. Moreover, they require whole-genome sequencing to identify the rare mutations, which is cost prohibitive for most applications. They are also computationally intensive and require specialized knowledge of lineage reconstruction using a population genetics approach. By contrast, barcoding can label any cells that are accessible to viral infection and integration, is easy and inexpensive, and does not require any advanced computational skills.

Limitations

The barcoding strategy presented here is limited to systems that tolerate cell isolation, short-term culture, and transplantation. In our protocol, cells of interest need to be isolated from their respective tissues in order to be transduced by the lentiviral vector carrying the genetic barcodes. This can be a problem for cells that cannot be readily isolated or that require maintenance of endogenous tissue architecture. In some cases, in vivo injection of barcode-carrying virus can be used as an alternative strategy, although it creates new problems of labeling unwanted cells and uneven transduction. In vivo delivery of the barcodes is not discussed further here.

Cells may potentially change their properties during culture and barcode transduction. Although many studies have shown that lentiviral integration does not cause any apparent change in the biological functions of the transduced cells[17,21,22,47,48], it is still possible that a particular lentiviral vector may be randomly inserted into some genomic region and alter cell behavior. Therefore, experimental replicates and controls must be rigorously used to exclude the possibility that rare viral insertions cause the observed phenotype.

Different cell types have different transduction rates. The technique reported here was optimized for mouse cells but has been used for studying primate cells as well[13,49]. Transduction efficiencies for primary human cells are generally lower than those for mouse cells, in our experience’ but are sufficient to yield meaningful results. Both mouse and human cell lines generally exhibit higher transduction efficiencies than primary cells.

While modern high-throughput sequencing has greatly improved barcode recovery' barcode detection is still limited by experimental loss during barcode extraction and cell collection. For example, cells collected from a part of a solid tissue may not be representative of the whole tissue. In addition, the sizes of some cell populations may numerically exceed the limit of our barcode extraction protocol, and only a fraction of the cells can be analyzed. Sequencing depth, as well as the barcode extraction procedure, may also limit the detection of barcodes with low abundance[50]. Furthermore, detection limits may vary between samples with differing cell numbers.

Experimental design

Plasmid generation—The oligo library template can be obtained from IDT or other vendors. We suggest HPLC purification for best results. The synthetic DNA oligos that we use to generate barcode libraries are composed of several parts: a BamHI restriction enzyme site, a forward primer binding site, a 6-bp library ID, a 27-bp random sequence, a reverse primer binding site, and an EcoRI restriction enzyme site (Fig. 1a). The 6-bp library ID enables different cell populations to be simultaneously barcoded and combined during downstream biological treatment, barcode extraction' and sequencing. This saves much experimental time and resources. The 27-bp random sequence generates a maximum of 4^{27} different barcodes in theory. This number is reduced by excluding sequences with restriction enzyme cutting sites and with characters difficult for PCR and sequencing such as poly-Ns. Longer or shorter random sequences and random sequences with interspersed fixed sequences can also be used. The length should not be so long that it exceeds the sequencing capacity, nor so short such that it limits library diversity. A 6-bp sequence is added to both ends to ensure proper restriction enzyme cutting (Table 1). The primer binding sites enable targeted PCR for barcode extraction (Fig. 1b). The BamHI and EcoRI sites are designed for cloning the double-stranded barcode DNA into lentiviral backbones, such as the pCDH plasmid. Other types of vectors are also applicable, as long as they can insert DNA barcodes into the host cell's genome. The plasmid may also include fluorescent proteins such as GFP to signal the presence of barcodes and to evaluate the transduction efficiency. The primer design can be customized as needed. The 6-bp library ID and 27-bp random sequence can be readily replaced to accommodate alternative barcode designs. Alternative barcode designs include interspersed fixed sequences and a library of known barcode sequences[16]. Implantation of partially or fully pre-designed barcode sequences can avoid restriction cutting sites and poly-N stretches.

Synthesized DNA oligos are made double stranded using a single primer 'Strand2' (Table 2). After cloning, each plasmid library is transformed into competent cells (*Escherichia coli*), and all bacterial colonies are amplified to achieve high barcode diversity. Bacterial cultures are grown overnight in an incubator. Plasmids are isolated from bacterial culture using the Qiagen Plasmid Maxi Kit. Plasmid DNA concentration is then measured using a NanoDrop spectrophotometer. Before proceeding to the next step, the plasmid must be sequenced for evaluating barcode diversity, that is, the number of unique barcodes and their relative abundances in the library[1,17]. A high library diversity (high barcode numbers and equal representation of unique barcodes) is essential to reducing the chance that more than one cell will be labeled by the same barcode. Optimizing the bacterial transformation step is

the key to improving library diversity. The diversity of the library dictates the number of clones that can be tracked in a single experiment, such that each barcode represents a single cell clone with statistical certainty. An exact calculation of this limit was provided in our previous study[1]. A user-friendly calculation tool is provided with this protocol (Supplementary Software). As a rule of thumb, a library of 40,000–50,000 barcodes typically enables ~1,000 cells to be tracked with greater than 95% probability that >95% of the barcodes represent single cells.

Lentivirus packaging—HEK293T cells are used to produce lentiviral particles. HEK293T cells are transfected with the pCDH-barcode plasmid and lentivectors Pax2 and VSV-G, in the presence of SuperFect transfection reagent. The supernatant is collected and the media is changed at 48, 72, and 96 h. The virus should always be kept on ice or at 4 °C after harvest. After pooling and concentrating using 50% PEG-8000, the virus should be divided into aliquots and kept at −80 °C for long-term (up to 1 year) storage. The lentiviral library must be tested on cell lines before use by transduction, barcode extraction and sequencing, to evaluate the viral titer and barcode diversity, that is, the number of unique barcodes and their relative abundances in the library. Results from sequencing plasmids and transduced cell lines can create reference libraries that facilitate downstream bioinformatics analysis[17].

Transduction of experimental cells—The exact transduction time depends on the research purpose, the type of cells, and how well the culture conditions support cellular properties. Using a low viral titer will ensure that each cell receives only one viral insertion and consequently one barcode. Cells that receive more than one barcode will be overrepresented in the results. We previously reported that ~50% transduction efficiency resulted in >95% of cells carrying only a single barcode[1]. Other studies have applied lower transduction efficiency (~15%) to further reduce the chance of double barcoding[17]. After incubation, cells should be washed thoroughly to remove any remaining virus. Labeled cells are now ready for experimental use.

It is important to use the same viral libraries for both the control and experimental groups and to include biological replicates using different viral libraries or viral infection wells, if possible, in order to avoid experimental noise associated with viral infection. We recommend evaluating the percentage of infected cells in each experiment by analyzing an aliquot of the experimental cells. The fractions of cells receiving single and multiple barcodes must be determined experimentally by analyzing the barcode copy number in genomic DNA at the clonal level. Multiple infections of a single cell can label the cell with multiple barcodes, and data from this cell will be overrepresented. The data are acceptable if cells with multiple barcodes are expected to produce results similar to those of cells with single barcodes. The number of experimental cells to be barcoded should be limited on the basis of the barcode diversity in the library[1]. This limit is particularly important for experiments using cell lines for which each barcode is meant to represent a single cell with statistical confidence. In addition to the cell number and viral incubation time, other experimental parameters, such as cell numbers for barcode analysis and time to harvest the

cells, also influence experimental results and can be adapted from previously published studies with similar experimental conditions[9,11,13,14,19,21–24,51–53].

Barcode extraction—Barcodes are recovered by isolating genomic DNA from cells of interest. For a given population' the number of cells required for analyses depend on the desired barcode detection sensitivity. To identify barcodes that are as rare as 1 in 1000, at least 1,000 barcoded cells must be collected for barcode recovery. If possible, >10,000 barcoded cells should be collected for best results. High cell numbers enable the identification of rare barcodes, but too many, as well as too few, cells may reduce barcode recovery rates and present problems during barcode extraction. Sorting is not required for collecting cells, but the collected cells should be prepared for genomic DNA extraction and counted in preparation for the barcode extraction procedure. From the isolated genomic DNA, barcodes are PCR-amplified using designed primers (Table 2) that flank the barcode region and provide binding sites for downstream high-throughput sequencing. These primers also add indexes that enable multiplex sequencing. To ensure precise quantification, the PCR should be halted during the exponential phase of amplification (typically 20–27 cycles) before the curve plateaus (Fig. 3). Samples with different numbers of cells may require different numbers of cycles. Compared to conventional PCR, which uses a predetermined cycle number, stopping the PCR reaction during the exponential phase prevents overamplification and reduces PCR bias, in line with the idea behind quantitative PCR. After PCR, barcode DNA is purified using magnetic beads.

DNA quantification and high-throughput sequencing—The amplified barcodes must be precisely quantified before sequencing. It is important to choose a quantification method that is sensitive and robust. We chose fluorescence-based quantification (Qubit assay), but other methods, such as TapeStation ScreenTape assay, may also suffice.

Barcoded samples prepared using different reverse primers can be pooled for sequencing as one sample to reduce cost. Our library ID design provides an additional option for multiplexing different barcoded samples. Additional index primers and library IDs can reduce sequencing cost at the expense of the additional resources to create them. Although we recommend HPLC-purified primers' desalted primers are also acceptable. The depth of sequencing depends on the number of barcoded cells used during barcode extraction. We recommend sequencing ~100 reads per barcoded cell to ensure precise barcode quantification. Although the barcode is only 33 bp long, we typically perform single-end sequencing for at least 50 bp, so that the sequence from the 34th to the 50th bp can be used as a quality control check.

Analysis of sequencing data—We developed custom Python scripts to extract barcodes from the raw sequencing results (Supplementary Software). The scripts consist of three major steps. In the first step, the code extracts the first 50 bp of each read. This 50 bp should consist of the 6-bp library ID, a 27-bp random sequence, and the 17-bp PCR handle. In the second step, the code aligns the last 17 bp of each read to the expected sequence. The reads containing the expected 17 bp are then separated on the basis of their first 6-bp sequence, that is, library ID. In this step, the code also counts the copy number for each unique sequence. In the third step, the code generates the final results that consist of master

barcodes and their copy numbers. The generation and use of master barcodes are explained in detail below.

Because PCR and sequencing can both generate errors[50], we combine sequences that are closely similar to each other following the conventional strategy used for analyzing high-throughput sequencing data. We use Levenshtein edit distance to quantify the similarity between different sequences. Each nucleotide substitution results in an edit distance of 1. Each indel results in an edit distance of 2 because all sequences are the same length and an indel creates an additional difference at the last base pair. By default, if the edit distance between different sequences is no more than 4 they are considered to be derived from the same sequence (Fig. 4). Our Python scripts enable users to customize the edit distance threshold.

In the second step, we allow a maximum edit distance of 4 when aligning the 17-bp sequence following the barcodes. We exclude reads whose first 6 bp do not match exactly any expected library IDs. In the third step, the code performs pairwise comparison of all the unique sequences and groups the pairs with an edit distance of no more than 4 that share a common sequence. Within each group, the sequence with the highest copy number is kept as the ‘master barcode’. The copy number for each master barcode is the sum of the copy numbers of all barcodes that differ by an edit distance of 4 from the master barcode. The master barcodes are used to represent the original barcodes delivered by the lentivirus. If a reference library from sequencing plasmids and transduced cell lines is used, the master barcode sequences can be drawn from the reference library instead. The sequences of the master barcodes can facilitate comparisons between different samples that are derived from the same barcoded cell population. The third step of the code generates a file reporting the distance between each unique sequence and its master barcode, as well as the distances between different master barcodes. This information can help users adjust the edit distance threshold. Although there is an R package (genBaRcode) available for similar barcode analysis[54], our Python code provides a flexible alternative that is easy to implement for users with little programming skill. Downstream data analysis and visualization are contingent upon the specific biological questions and can be adapted from previous studies[6,7,9,11–17,21,22,49].

Materials

Biological materials

- 5-alpha Competent *E. coli*, high efficiency (New England BioLabs, cat. no. C2987I)
- HEK293T cell line (ATCC cat. no. CRL-3216, RRID: CVCL_0063) **!** **CAUTION** Cell lines should be checked for authenticity and to ensure that they are not infected with mycoplasma.

Reagents

- Lentivirus pCDH plasmid (System Biosciences, cat. no. CD523A-1)
- BamHI restriction enzyme (New England BioLabs, cat. no. R3136S)

- EcoRI restriction enzyme (New England BioLabs, cat. no. R0101S)
- NEBuffer 3.1 (10×; New England BioLabs, cat. no. B7203S)
- DNA polymerase I, large (Klenow) fragment (New England BioLabs, cat. no. M0210S)
- Deoxynucleotide (dNTP) solution mix (New England BioLabs, cat. no. N0447S)
- Reaction Buffer (10×; New England BioLabs, cat. no. B9014S)
- Zymoclean Gel DNA Recovery Kit (Zymo Research, cat. no. D4001; Genesee Scientific, cat. no. 11-300C)
- Oligo library (Integrated DNA Technologies; see Table 1 for ordering details)
- T4 DNA ligase (New England BioLabs, cat. no. M0202S)
- SOC outgrowth medium (New England BioLabs, cat. no. B9020S)
- LB agar plates, (Quality Biological, LB agar with 100 μg/mL ampicillin; VWR, cat. no. 10128-318)
- Premixed LB broth (Miller formulation; VWR, cat. no. 97064-114)
- Agarose RA (Benchmark Scientific, cat. no. A1700)
- GeneJET Gel Extraction Kit (Thermo Fisher Scientific, cat. no. K0691)
- Plasmid Maxi Kit (Qiagen, cat. no. 12162)
- Pax2 lentiviral vector (Addgene, cat. no. 35002)
- pCMV-VSV-G lentiviral vector (Addgene, cat. no. 8454)
- SuperFect transfection reagent (Qiagen, cat. no. 301305)
- PBS (Life Technology, cat. no. 21600-010)
- DMEM (Life Technology, cat. no. 11320033)
- Penicillin–streptomycin (Thermo Fisher Scientific, cat. no. 15140122)
- Fetal bovine serum (FBS; Fisher Scientific, cat. no. SH3007103)
- Poly(ethylene glycol) for molecular biology (molecular weight = 8,000; BioUltra; Sigma-Aldrich, cat. no. 81268-250G)
- Quick-DNA Microprep Kit (Zymo Research, cat. no. D3020)
- Primers, HPLC purified (Integrated DNA Technologies; see Table 2 for ordering details)
- Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Fisher, cat. no. F531L)
- EvaGreen Dye (VWR, cat. no. 89138-984)
- SPRIselect beads (Beckman Coulter, cat. no. B23318)

- Ethyl alcohol (Pure, 200 proof; Sigma-Aldrich, cat. no. E7023-500ML)
- Water (Ultra Pure, sterile; Genesee Scientific, cat. no. 18-194)
- Qubit dsDNA HS Assay Kit (Life Technologies, cat. no. Q32854)
- NextSeq 500 High Output v2 Kit (Illumina, cat. no. FC-404-2005)
- Agarose LE (Apex Bioresearch, cat. no. 20-102)
- GelGreen Nucleic Acid Gel Stain (10,000×; EmbiTec, cat. no. EC-1995)
- 6X Loading Dye (VWR, cat. no. 470105-014)
- Apex DNA Ladder III (Apex Bioresearch, cat. no. 42-425)
- Tris acetate-EDTA (TAE; 50×; Fisher Scientific, cat. no. MP1TAE50X01)
- Isopropanol (Sigma-Aldrich, cat. no. 190764)
- Nuclease-free water (Sigma-Aldrich, cat. no. W4502-1L)

Equipment

- Pipettes (Rainin, cat. nos. 17014382, 17014383, 17014384)
- Falcon tissue culture dishes (polystyrene, sterile; Corning, cat. no. 353003)
- Filter tips for pipettes (Denville Scientific, cat. nos. P1126, P1122, P1121, P1096-FR)
- Sterile syringe filters (0.45-μm filter, 10 mL, 33 mm; Fisher Scientific, cat. no. SLHVM33RS)
- Eppendorf Flex-Tube 1.5-mL microcentrifuge tubes (Eppendorf, cat. no. 022364111)
- Falcon centrifuge tubes (polypropylene, sterile; VWR, cat. no. 21008-918)
- Cell culture plate (6-wells; VWR, cat. no. 62406-161)
- Syringe (50 mL; VWR, cat. no. 80062-745)
- Filter (0.45 μm; VWR, cat. no. 28145-505)
- Cell culture plates (96 wells, untreated; VWR, cat. no.15705-064)
- C1000 Touch ThermoCycler (Bio-Rad, cat. no. 1851148)
- Portable balances (Fisher Scientific, cat. no. S94792C)
- Gelbox and combs, RunOne electrophoresis system with timer (Embi Tec, cat. no. EP-2100)
- Safe Imager 2.0 blue-light transilluminator (Thermo Fisher Scientific, cat. no. G6600)

  ! CAUTION Always wear UV-light-protective safety glasses/face shield.
- Shaker (Innova 44/44R; New Brunswick, cat. no. M1282-0000)

- Tissue culture incubator (Panasonic, cat. no. MCO-170AICUVL-PA)
- NanoDrop 2000c spectrophotometer (Thermo Fisher, cat. no. ND-2000)
- Bench-top centrifuge (Beckman Coulter, cat. no. B30134)
- Centrifuge (Beckman Coulter, cat. no. A99465)
- Swinging-bucket rotors (Beckman Coulter, cat. nos. 366650, 360581)
- Cell sorter (with 530/30-nm FITC laser; BD, model no. FACSAria III)
- PCR plates (96 wells, Olympus FAST-type; Genesee Scientific, cat. no. 24-310)
- ThermalSeal RTS sealing films, sterile (Excel Scientific, cat. no. TSS-RTQS-50)
- Real-time PCR System (ViiA 7; Applied Biosystems, cat. no. 4453545)
- Strip tube magnet (0.2 mL; 10×; Genomics, cat. no. 230003)
- Qubit fluorometer (Thermo Fisher, cat. no. Q33238)
- Gel Extraction Tips (VWR, cat. no. 89179-796)
- Razor blade (VWR, cat. no. 55411-050) or scalpel (VWR, cat. no. 82029-864)
- Computer with Windows 7+ installed

Software

- Microsoft Visual C++ Compiler for Python 2.7 (https://www.microsoft.com/en-us/download/details.aspx?id=44266)
- Anaconda distribution, Python 2.7 version (https://www.anaconda.com/distribution/)

Reagent setup

Oligos and primers—Resuspend IDT DNA oligos (Table 1) and primers (Table 2) to 100 μM in nuclease-free water. Dilute to a 10 μM concentration by adding 10 μL of 100 μM primers to 90 μL of nuclease-free water. DNA oligos and primers can be stored at 10 μM or 100 μM at −20 °C for up to 2 years.

DMEM—Mix 445 mL of DMEM with 50 mL of FBS (10% (vol/vol) final FBS concentration) and 5 mL of penicillin–streptomycin (1% (vol/vol) final penicillin–streptomycin concentration). Store at 4 °C for up to 1 month.

1× TAE buffer—Mix 2 mL of 50× TAE with 98 mL of water for 100 mL of 1× TAE. Store at room temperature (25 °C) until expiration date on packaging.

70% (vol/vol) ethanol solution—Mix 700 μL of ethyl alcohol (pure, 200 proof) with 300 μL of nuclease-free water to obtain 1 mL of 70% (vol/vol) ethanol right before use. ▲ **CRITICAL** Make fresh before use.

Agarose gel—Prepare before use. Mix 1 g for 1% (wt/vol) or 3 g for 3% (vol/vol) agarose with 100 mL of 1× TAE, heat in microwave until agarose completely dissolves, pour the solution into a casting box with the comb positioned, and cool at room temperature for at least 20 min until the gel solidifies.

25× GelGreen DNA dye—Mix 1 μL of 10,000× GelGreen DNA dye with 399 μL of nuclease-free water. Store at 4 °C for up to 6 months.

Equipment setup

Tissue culture incubator—Set incubator to 37 °C with 5% carbon dioxide. Keep the humidifier pan full by adding sterile water.

Software—Download and install Anaconda Distribution, Python 2.7 version. Then download code_demo.zip (Supplementary Software). The zip file includes the following files:

- readme.pdf
- step-1_read-raw-data.py
- step-2_combine-library-ID.py
- step-3_combine-barcodes.py
- step-4-opt_evaluating_barcode_diversity.py
- library_ID.txt
- sample info 041519.txt
- expected_output

Procedure

Plasmid generation ● Timing 2-3 d

1. Order the DNA oligos listed in Table 1 from Integrated DNA Technologies or another vendor. Twenty-four library IDs are provided; fewer can be used if not needed.
2. Perform second-strand synthesis using Strand2 primer (Table 2). Mix and incubate for 2 h at 16 °C. Perform a separate reaction for each virus library from Table 1. A negative control to estimate the background should be set up by removing the enzyme. The experimental setup is shown in the table below:

Component	Amount (μL)	Final (concentration/amount)
Oligos (100 μM; Table 1)	8	40 μM
Strand 2 primer (10 μM, Table 2)	1	0.5 μM
dNTPs (10 μM)	2	1 μM

Component	Amount (μL)	Final (concentration/amount)
10× reaction buffer	2	1×
Klenow enzyme	1	5 U
Nuclease-free water	6	

The negative-control setup is shown in the table below:

Component	Amount (μL)	Final (concentration/amount)
Oligos (100 μM; Table 1)	8	40 μM
Strand 2 primer (10 μM; Table 2)	1	0.5 μM
dNTPs (10 μM)	2	1 μM
10× reaction buffer	2	1×
Klenow enzyme	0	
Nuclease-free water	7 μL	

3. Vortex the SPRIselect magnetic beads before use.

4. Add 1.8× beads to the sample from Step 2 (36 μL of beads for 20 μL of reaction) and gently mix with a pipette 15 times.

   ▲ **CRITICAL STEP** Do not vortex in Steps 4–15.

5. Incubate the sample with beads at room temperature for 5 min.

6. Condense the beads into a pellet with a magnet for 3–5 min.

7. Remove and discard the supernatant without disturbing beads, leaving ~5 μL behind at the bottom of the tube. Keep the beads on the magnet until the elution step; do not disturb the pellet.

8. Pipette 200 μL of 70% ethanol into the sample without disturbing beads; keep the beads on the magnet.

   ▲ **CRITICAL STEP** Prepare fresh 70% (vol/vol) ethanol. Ethanol that has been stored for too long will have an incorrect ethanol/H_2O ratio, which will decrease the DNA yield.

9. Leave the ethanol with the beads for 30 s; then remove the ethanol and discard.

10. Repeat the wash (Steps 8 and 9, for a total of two ethanol washes).

11. Remove as much of the ethanol as possible. Be mindful of small ethanol droplets.

12. Air-dry the pellet for ~1 min.

    ▲ **CRITICAL STEP** The drying time for beads is variable. Be careful not to overdry the pellet, which will lead to cracking and/or breakup, reducing DNA recovery.

13. Add 20 μL of nuclease-free water to all samples and then pipette to mix 15 times. Repeat the mixing to ensure better recovery.

14. Incubate for 5 min.

15. Condense the beads into a pellet with a magnet for 3–5 min.

16. Collect the supernatant into a new tube. (Optional) Capture carry-over beads with a magnet for 3–5 min and then transfer the supernatant into a new tube.

    ■ **PAUSE POINT** Store at −20 °C for long-term storage.

17. Digest the purified product using EcoRI and BamHI.

Component	Amount (μL)	Final (concentration/amount)
DNA (from Step 16)	1	
10× NEBuffer 3.1	5	1×
EcoRI	1	10 U
BamHI	1	10 U
Nuclease-free water	Up to 50	

18. 18 Incubate at 37 °C for 60 min.

19. Add 10 μL of 6× loading dye and 2.4 μL of 25× GelGreen DNA dye to the product.

20. Run all product on a 3% (wt/vol) agarose gel (in 1× TAE) at 100 V for 60 min.

    ▲ **CRITICAL STEP** Depending on gel well volume capacity, the product may need to be loaded into multiple wells.

21. Illuminate the DNA in the gel with a UV transilluminator. The expected band should be ~100 bp. Excise the DNA fragment from the agarose gel with gel extraction tips, a razor blade or a scalpel, and transfer it to a 1.5-mL microcentrifuge tube.

    ! CAUTION Always wear UV-light protective safety glasses/face shield.

22. Purify the DNA from the gel using a Gel DNA Recovery Kit.

23. Add 3 volumes of ADB buffer (provided in the kit) per each 100 mg of agarose excised from the gel (e.g., for 100 mg of agarose gel, add 300 μL of ADB).

24. Incubate at 37–55 °C for 5–10 min until the gel slice has completely dissolved. For DNA fragments >8 kb, following the incubation, add one additional volume (corresponding to the weight of the gel slice) of water to the mixture for better DNA recovery (e.g., 100 μL of agarose, 300 μL of ADB, and 100 μL of water).

25. Transfer the melted agarose solution to a Zymo-Spin column in a collection tube (both from the kit) and centrifuge (10,000*g*, room temperature, 1 min).

26. Discard the flow-through. Add 200 μL of DNA Wash Buffer (from the kit) to the column and centrifuge (10,000*g*, room temperature,1 min). Discard the flow-through. Repeat the wash.

27. Add ≥6 μL of nuclease-free water directly to the column matrix. Place the column into a 1.5-mL tube and centrifuge (10,000*g*, room temperature, 1 min) to elute the DNA.

28. Prepare the lentivirus pCDH vector backbone (or vector of choice). Linearize 5 μg of the pCDH vector by digesting it with EcoRI and BamHI, and run an agarose gel as described in Step 20 but using a 1% (wt/vol) agarose gel. Purify the excised product as described in Steps 21–27. Band size should be ~7,150 bp.

29. Ligate the vector and the barcode insert into a microcentrifuge tube on ice. A negative control with no ligase should be included to assess the background level. The experimental setup is shown in the table below:

Component	Amount	Final (concentration/amount)
T4 DNA ligase buffer	2 μL	1×
Linear lentivirus pCDH vector DNA (Step 28)	50 ng	2.5 ng/μL
Insert DNA (Step 27)	37.5 ng	1.875 ng/μL
Nuclease-free water	Up to 19 μL	
T4 DNA ligase	1 μL	400 U

The negative-control setup is shown in the table below:

Component	Amount	Final (concentration/amount)
T4 DNA ligase buffer	2 μL	1×
Linear lentivirus pCDH vector DNA (Step 28)	50 ng	2.5 ng/μL
Insert DNA (Step 27)	37.5 ng	1.875 ng/μL
Nuclease-free water	Up to 20 μL	
T4 DNA ligase	0 μL	

▲ **CRITICAL STEP** Add T4 DNA ligase last.

30. Gently mix the reaction by pipetting up and down and microcentrifuging briefly (2,000*g*, 25 °C, 3 s).

31. Incubate at 16 °C overnight or at room temperature for 10 min.

32. Heat-inactivate at 65 °C for 10 min.

❚ **PAUSE POINT** Store at −20 °C for long-term storage.

33. Thaw a tube of high-efficiency 5-alpha competent *E. coli* cells on ice until the last ice crystals disappear.

34. Add the entire ligation reaction (20 μL). Carefully flick the tube four or five times to mix the cells and DNA.

▲ **CRITICAL STEP** Do not vortex.

35. Incubate the mixture on ice for 30 min. Do not mix.

36. Heat-shock at exactly 42 °C for exactly 30 s. Do not mix.

37. Place on ice for 5 min. Do not mix.

38. Pipette 950 μL of room-temperature SOC medium into the mixture.

39. Incubate at 37 °C for 60 min, shaking vigorously (225 r.p.m.) or rotating the samples.

40. Pre-warm the 100 μg/mL ampicillin-selection plates to 37 °C.

41. Mix the cells thoroughly by flicking and inverting the tube, then perform several 100-fold serial dilutions in SOC media with 10 μL of transformation reaction from Step 39.

42. Spread 50–100 μL of each dilution onto a pre-warmed selection plate and incubate overnight at 37 °C to estimate the transformation efficiency.

? TROUBLESHOOTING

43. Pipette the remaining transformant into 100 mL of LB broth with 100 μg/mL ampicillin and incubate overnight at 37 °C, shaking vigorously (225 r.p.m.) or rotating the sample.

44. The following day, extract plasmid using a Qiagen Plasmid Maxiprep Kit.

45. Harvest overnight bacterial culture by centrifuging at 6,000× for 15 min at 4 °C.

46. Resuspend the bacterial pellet in 10 mL of Buffer P1 (from the Qiagen Plasmid Maxiprep Kit).

47. Add 10 mL of Buffer P2 (from the Qiagen Plasmid Maxiprep Kit), mix thoroughly by vigorously inverting four to six times, and incubate at room temperature for 5 min. If using LyseBlue reagent, the solution will turn blue.

48. Add 10 mL of prechilled Buffer P3 (from the Qiagen Plasmid Maxiprep Kit) and mix thoroughly by vigorously inverting four to six times. Incubate on ice for 20 min. If using LyseBlue reagent, mix the solution until it is colorless.

49. Centrifuge at ≥20,000*g* for 30 min at 4 °C. Re-centrifuge the supernatant at ≥20,000*g* for 15 min at 4 °C.

50. Equilibrate a Qiagen-tip 500 by applying 10 mL of Buffer QBT and allow the column to empty by gravity flow.

51. Apply the supernatant from Step 49 to the Qiagen-tip and allow it to enter the resin by gravity flow.

52. Wash the Qiagen-tip with 2 × 30 mL of Buffer QC. Allow the Buffer QC to move through the Qiagen-tip by gravity flow.

53. Elute the DNA into a clean 50-mL vessel with 15 mL of Buffer QF.

54. Precipitate the DNA by adding 10.5 mL (0.7 volumes) of room temperature isopropanol to the eluted DNA and mixing. Centrifuge at ≥15,000*g* for 30 min at 4 °C. Carefully decant the supernatant.

55. Wash the DNA pellet with 5 mL of room-temperature 70% (vol/vol) ethanol and centrifuge at ≥15,000*g* for 10 min at 25 °C. Carefully decant the supernatant.

56. Air-dry the pellet for 5–10 min and re-dissolve the DNA in a suitable volume of nuclease-free water. Use a NanoDrop spectrophotometer to measure plasmid concentration.

▲ **CRITICAL STEP** Test the barcode diversity of the plasmid library by following Steps 91–127 before lentiviral packaging. Proceed only if the barcode diversity allows for tracking of single cells in the intended experiment (Step 127)[1].

■ **PAUSE POINT** Store at −20 °C for long-term storage.

? TROUBLESHOOTING

Lentivirus packaging ● Timing 4-5 d

57. Plate 8×10^5 HEK293T cells per well in a 6-well plate. Grow overnight to ~80% confluency.

! CAUTION Except for centrifugation, all downstream procedures in Steps 57–68 must be performed in a biosafety cabinet. Use freshly prepared 10% (vol/vol) bleach solution to disinfect tips after pipetting the virus.

58. Aspirate and discard media for the HEK293T cells. Add 1 mL of DMEM to each well.

59. For each well, mix in a tube 2 μg of ligated lentivirus pCDH vector from Step 56, 1.3 μg of Pax2, and 0.7 μg of pCMV-VSV-G in a 100-μL final volume of DMEM.

60. Add 6 μL of SuperFect transfection reagent to the tubes from Step 59, vortex for 10 s, incubate at room temperature for 5–10 min, and then disperse drop by drop to the cells evenly across the plate. Put the plate into a 37 °C incubator.

61. After 8–12 h, aspirate and discard the media, wash the HEK293T cells with PBS, and add 2 mL of fresh DMEM to each well.

62. Harvest 2 mL of supernatant containing the virus at 48, 72, and 96 h after Step 61, pool supernatants from the same well together on ice, and store at 4 °C. Replace with 2 mL of fresh DMEM at each time point.

63. Spin down for 10 min at 300*g* and 4 °C. Collect the supernatant into a new tube.

64. With a 0.45-μm filter mounted on a syringe, filter the supernatant into a .25 volume of 50% PEG-8000 (e.g., 6 mL of virus supernatant into 2 mL of 50% PEG-8000).

65. Mix by inverting and incubate at 4 °C overnight.

66. The following morning, spin down the tubes at 1,500*g* at 4 °C for 20 min and discard the supernatant.

67. Spin down again at 300*g* at 4 °C for 5 min to remove as much of the remaining liquid as possible.

68. Resuspend the virus stock in 30 μL of PBS, make 3-μL aliquots (or your preferred size), and store at −80 °C.

▲ **CRITICAL STEP** Avoid multiple freeze-thaw cycles to prevent virus stock degradation.

▲ **CRITICAL STEP** Titer the virus stocks and test the barcode diversity by performing Steps 69–89 on a cell line similar to the experimental cells. Make sure that the viral concentration and exposure time are appropriate for transduction and result in 15–50% GFP expression.

▲ **CRITICAL STEP** Test the diversity of the lentiviral library by performing Steps 69 −127 on a cell line. Proceed only if the barcode diversity allows for tracking single cells in the intended experiment (Step 127)[1].

■ **PAUSE POINT** Store virus stocks at −80 °C for up to 1 year.

? TROUBLESHOOTING

Transduction of experimental cells ● Timing 13-16 h

! CAUTION Except for centrifugation, all downstream procedures after Step 69 must be performed in a biosafety cabinet. Use freshly prepared 10% (vol/vol) bleach solution to disinfect tips after pipetting the virus.

69. Prepare experimental cells in a new tube. Cells can be grown in suspension or adherent.

70. Spin down the collected cells at 300*g* (speed may vary depending on cell type) at 4 °C for 5–10 min.

71. Aspirate the supernatant with a pipette, leaving ~20 μL above the pelleted cells.

72. To allocate cells into different samples, resuspend the cells in the desired culture medium to achieve a 30-μL volume per sample. Add each 30-μL cell suspension sample to an individual well in a 96-well plate.

73. Add barcode virus from Step 68 at the determined viral concentration to each well containing the cells. If the samples will be pooled for sequencing (Step 119), each sample needs to receive a different virus library ID. Mix with the pipette.

▲ **CRITICAL STEP** Always keep the virus stock on ice.

74. Incubate at 37 °C for the viral exposure time determined in Step 68.

75. Collect each sample by adding 200 μL of PBS (or medium of choice) to each well, mix by pipetting, and transfer the cells to a new tube for each sample.

▲ **CRITICAL STEP** Adherent cells will need to be detached from the plate.

76. To collect residual cells, wash each well with 200 μL of PBS and transfer the contents to the respective tubes from Step 75.

77. Repeat Step 76 three more times (1-mL final volume).

78. Spin down each tube at 300*g* at 4 °C for 5 min.

79. Aspirate and discard the supernatant, leaving ~50 μL above the pelleted cells. Add 1 mL of PBS (or medium of choice) and mix by pipetting thoroughly.

80. Repeat Steps 78 and 79 for a total of three washes.

81. The cells are now barcoded and ready for an experiment. (Optional) Culture a small aliquot of experimental barcoded cells for 3–5 d and analyze GFP expression, using flow cytometry to estimate the transduction rate.

? TROUBLESHOOTING

Barcode extraction ● Timing 4-5 h

82. Pellet experimental barcoded cells by centrifuging at 300*g* at 4 °C for 5 min. Remove and discard the supernatant.

▮ **PAUSE POINT** Store the cell pellet at −80 °C for up to 6 months.

83. Extract genomic DNA using a Quick-DNA Micro Prep Kit. Start by adding 400 μL of Genomic Lysis Buffer (from the Quick-DNA Micro Prep Kit) to the cell pellets (add 800 μL of lysis buffer if the sample contains >0.8 million cells).

84. Mix completely by vortexing for 4–6 s and then incubate for 5–10 min at room temperature.

85. Transfer the mixture to a Zymo-Spin IC column in a collection tube. Do not work with >1 million cells. Centrifuge at 10,000*g* for 1 min at room temperature. Discard the collection tube with the flow-through.

86. Transfer the Zymo-Spin IC column to a new collection tube. Add 200 μL of DNA Pre-Wash Buffer (from the kit) to the spin column. Centrifuge at 10,000*g* for 1 min at room temperature.

87. Add 500 μL of g-DNA Wash Buffer (from the kit) to the spin column. Centrifuge at 10,000*g* for 1 min at room temperature.

88. Transfer the spin column to a clean microcentrifuge tube. Add 20 μL of DNA Elution Buffer (from the kit) to the spin column. Incubate for 2–5 min at room temperature and then centrifuge at 10,000*g* for 1 min at room temperature to elute the DNA.

89. Load the flow-through again on the same column, incubate for another 5 min, and then centrifuge at 10,000*g* for 1 min at room temperature to elute the DNA for a second time.

90. Use a 20-μL pipette to measure the eluted volume.

▲ **CRITICAL STEP** Use a Qubit spectrophotometer to measure genomic DNA concentration (Steps 110–118). For the subsequent PCR reaction, use no more than 1,000 ng of gDNA in the 40-μL PCR reaction. Multiple samples can be combined into one PCR reaction if their library IDs are different. Make sure that each cell sample is equally represented by using equivalent amounts of gDNA in the combined PCR reaction.

❚ **PAUSE POINT** Store gDNA at −20 °C for up to 1 year.

? TROUBLESHOOTING

91. Perform barcode recovery by qPCR using Phusion High-Fidelity PCR Master Mix with HF Buffer. Keep the reagents on ice.

92. The amounts given below are for a 40-μL PCR reaction; use a negative control without DNA template to assess background noise and contamination. The experimental setup is shown in the table below:

Component	Amount (μL)	Final (concentration/amount)
Template (from Step 90)	16	
Forward primer (10 μM; Table 2)	2	1 μM
Reverse primer of choice (10 μM; Table 2)	2	1 μM
Phusion Mix (2×; from the kit)	20	1×
EvaGreen dye	0.4	1×

The negative-control setup is shown in the table below:

Component	Amount (μL)	Final (concentration/amount)
Nuclease-free water	16	
Forward primer (10 μM; Table 2)	2	1 μM
Reverse primer of choice (10 μM; Table 2)	2	1 μM
Phusion Mix (2×; from the kit)	20	1×
EvaGreen dye	0.4	1×

93. Mix, briefly spin down (2,000*g*, 25 °C, 3–5 s), and load on a qPCR machine.

Cycle no.	Denature	Anneal	Extend
1	98 °C, 30 s		
2-27	98 °C, 10 s	65 °C, 30 s	72 °C, 30 s
28			72 °C, 10 min

▲ **CRITICAL STEP** Stop the PCR when the fluorescence increases by 900,000 a.u. and/or after 5–8 PCR cycles from the start of the exponential phase. The

PCR curve should still be in the exponential phase when the PCR is stopped; this typically occurs between 20 and 27 cycles.

▲ **CRITICAL STEP** Run samples with similar cell numbers together. Samples with substantially different numbers of cells need to be processed separately, because they will need different numbers of PCR cycles.

? TROUBLESHOOTING

94. Vortex the SPRIselect magnetic meads before use.

95. Add 1.8× beads to the sample from Step 93 (72 μL of beads for a 40-μL PCR reaction) and gently mix with pipette 15 times.

▲ **CRITICAL STEP** Do not vortex for Steps 95–107.

96. Incubate the sample with the beads at room temperature for 5 min.

97. Condense the beads into a pellet with the magnet for 3–5 min.

98. Remove and discard the supernatant without disturbing the beads, leaving ~5 μL behind at the bottom of the tube. Keep the beads pelleted until the elution step; do not disturb the pellet.

99. Pipette 200 μL of 70% (vol/vol) ethanol without disturbing the beads, and keep them pelleted.

▲ **CRITICAL STEP** Prepare fresh 70% (vol/vol) ethanol. Ethanol that has been stored for too long will have an incorrect ethanol/H_2O ratio, which will impair DNA yield.

100. Leave the ethanol on the beads for 30 s; then remove and discard the ethanol.

101. Repeat the wash (Steps 99 and 100) for a total of two ethanol washes).

102. Remove as much of the ethanol as possible. Be mindful of small ethanol droplets.

103. Air-dry the pellet for ~1 min.

▲ **CRITICAL STEP** The drying time for beads is variable. Be careful not to overdry the pellet, which will lead to cracking and/or breakup, thus reducing DNA recovery.

104. Add 20 μL of nuclease-free water to all samples and then pipette 15 times to mix. Repeat the mixing to ensure better recovery.

105. Incubate for 5 min.

106. Condense beads into a pellet with the magnet for 3–5 min.

107. Collect the supernatant into a new tube.

108. (Optional) Capture carry-over beads with the magnet for 3–5 min and transfer the supernatant to a new tube.

109. Quantify the concentration of the purified PCR product using a Qubit fluorometer.

▲ **CRITICAL STEP** Incubate with SPRIselect beads (Steps 94–108) longer to increase recovery rate.

❚ **PAUSE POINT** Store purified PCR product at −20 °C for up to 1 year.

? TROUBLESHOOTING

DNA Quantification ● Timing 0.5-1 h

110. Quantify the concentration of the barcode library.

111. Prepare 0.5-mL tubes. The number of tubes necessary is the number of samples plus two. Label them with the appropriate sample IDs and ‘S1’ and ‘S2’.

112. Using the Qubit dsDNA HS Assay kit, prepare a working solution by mixing dsDNA HS Buffer and dsDNA HS Reagent (from the kit) at a ratio of 199:1. Make enough for 200 μL × (sample number + 2). For example, for 2 samples, prepare at least 200 μL × (2 + 2) = 800 μL of working solution.

113. In the tube labeled ‘S1’, mix 190 μL of working solution with 10 μL of dsDNA HS Standard 1 (from the kit).

114. In the tube labeled ‘S2’, mix 190 μL of working solution with 10 μL of dsDNA HS Standard 2 (from the kit).

115. In the sample tubes, mix 199 μL of working solution with 1 μL of sample from Step 107.

116. Vortex all the tubes, microcentrifuge (2,000*g*, 25 °C) them for a few seconds, and then incubate them at room temperature for 2 min.

117. Turn on the Qubit fluorometer, select ‘DS-HS-DNA’ and run a new calibration using the S1 and S2 tubes. Then load the sample tubes, following the instructions on the screen.

118. Record the measurement results.

High-throughput sequencing · Timing 2-3 d

▲ **CRITICAL** Genomic DNA from cells tagged with different library IDs may use the same reverse primer in qPCR amplification and can be combined into one sequencing sample. Genomic DNA from cells tagged with the same library ID should be amplified with different reverse primers if they will be combined in one sequencing sample (Table 2, qPCR). Table 2 provides three examples of reverse primer designs. Additional reverse primers can be designed as needed.

119. When pooling sequencing samples, use the same amount of DNA from each cell source. Try to save half of each sample in case of a failed run. The remaining sample can be stored at −20 °C for up to 1 year.

120. Sequence the samples using a NextSeq 500 High Output v2 Kit.

- Sequencing primer: see Table 2. Note that Custom Index Primer is not a standard Illumina sequencing primer.
- Sequencing cycle: set read1 for at least 50 cycles; set Index i7 for 6 cycles.
- Sequencing depth: plan for 2 million reads per cell source.
- Sequencing sample name: use the 6-bp reverse primer index (Table 2) as the sample name.

? TROUBLESHOOTING

Analysis of sequencing data ● Timing 1 d

121. Install the Anaconda Distribution, Python 2.7 version (https://www.anaconda.com/distribution/).

122. From the 'Start Menu', locate the 'Anaconda2 (64-bit)' folder, start 'Anaconda Prompt', and then type the following command into the terminal window:

pip install python-Levenshtein

Press 'Enter' and then type the following command:

pip install biopython

Press 'Enter'. In a few minutes, messages will show up indicating that the packages have been successfully installed.

123. In the 'Anaconda2 (64-bit)' folder, start the Spyder software. Use the software to open '*.py' Python scripts in the 'code demo' folder (Supplementary Software).

124. Open the step-1_read-raw-data.py file and edit 'variables subject to change' accordingly:

- 'fastq_location'. This specifies the directory where the raw sequencing data are stored. Sample data is provided as the GCCAAT_S1_R1_001.fastq.gz file online (see Data Availability). 'GCCAAT' is the sample index corresponding to the reverse primer that was used to extract the barcodes.
- 'date_today'. The code uses this information to document data analysis history.
- 'step1_output'. This specifies the folder in which to store the output of this script. Default is 'step-1 '. If the folder does not exist, the code will create one.

Run the file. A GCCAAT_041519.txt file will be generated in the output folder. It includes the first 50 bp of all reads in this sample.

125. Open the step-2_combine-library-ID.py file and edit 'variables subject to change' accordingly:

- ‘step1_location’. This specifies the folder in which to store the step-1 output. It is the input for this step.
- ‘sample_info’. This is a tab-delimited text file. A template can be found in the ‘code demo folder’.
  - ‘reverse primer’ and ‘primer index’. These hold information about the reverse primer used to extract the barcodes from the sample. Note that primer index was used to name the sequencing sample.
  - ‘sample’ and ‘lib ID’. These are the cell sample name and the corresponding virus library ID. Note that cell samples with different virus library IDs can be combined in the same sequencing sample.
- ‘distance_allowed’. This specifies the edit distance allowed when determining whether the 34th to the 50th base pair of each read is the adaptor sequence. Default is 4.
- ‘step2_output’. This is the folder in which to store the output of this script. Default is ‘step-2’. If the folder does not exist, the code will create it.

  Run the file. Two types of output files will be generated in the output folder:
  - [sample].bin. For this demo, ‘sample1.bin’ and ‘sample2.bin’. They store the intermediate files that serve as input for the next step.
- [index]_stats.txt. For this demo, GCCAAT_stats.txt. It is the quality report for each sequencing sample in this step.

? TROUBLESHOOTING

126. Open the step-3_combine-barcodes.py file and edit ‘variables subject to change’ accordingly:

- ‘step2_location’. Specifies folders storing the step-2 output. It is the input for this step.
- ‘distance_allowed’. Specifies edit distance allowed when determining whether two barcodes are legitimately the same. Default is 4.
- ‘step3_output’. Specifies folder to store the output of this script. Default is ‘step-3’. If the folder does not exist, the code will create it.

  Run the file. Four types of output files will be generated in the output folder:
- [sample]_[number of reads].txt. This is a tab-delimited text file; the first column is the barcode sequence, and the second column is the number of reads of this barcode.

- [sample]_[number of reads].bin. This is a binary file storing a dictionary, whose key is master barcodes, and item is a dictionary in which the key is barcodes that are legitimately the same as the master barcode, and item is the corresponding number of reads.
- [sample].xlsx. This is an Excel file with two worksheets. The ‘intraclonal’ sheet reports the copy number and the Levenshtein distance with the master barcode for each unique barcode; the ‘interclonal’ sheet reports Levenshtein distance between different master barcodes.
- step3_stats.txt. This is the quality report for this step.

Evaluation of diversity ● Timing 0.5 h

127. This step applies only to evaluating plasmid library diversity post plasmid generation (Step 56) and evaluating lentiviral library diversity post lentiviral packaging (Step 68). Open the step-4-opt_evaluating-barcode-diversity.py file and edit ‘variables subject to change’ accordingly:

- library_file. This is a tab-delimited text file generated by step-3_combine-barcodes.py.
- ‘ intended_cells’. This specifies the amount of cells to be tracked in an intended experiment.
- ‘simulation_events’. This specifies the number of Monte Carlo simulation experiments. Default is 1,000. Although a larger number of experiments take longer time, the user should try different values until a stable result is obtained.

Run the file. The code will print on the screen whether the given library is sufficiently diverse to track the intended number of cells with >95% probability that >95% of the barcodes represent single cells.

? TROUBLESHOOTING

Troubleshooting

Troubleshooting advice can be found in Table 3.

Timing

Steps 1–56, plasmid generation: 2–3 d

Steps 57–68, lentivirus packaging: 4–5 d

Steps 69–81, transduction of experimental cells: 13–16 h

Steps 82–109, barcode extraction: 4–5 h

Steps 110–118, DNA quantification: 0.5–1 h

Steps 119 and 120, high-throughput sequencing: 2–3 d

Steps 121 −126, analysis of sequencing data: 1 d

Step 127, evaluation of diversity: 0.5 h

Anticipated results

Plasmid generation

Barcode oligos should be inserted at the BamHI and EcoRI restriction enzyme sites. Plasmids should be ~100 bp larger (~7,250 bp if using the pCDH lentivirus vector) and circularized if ligation is successful. PCR using qPCR primers from Table 2 should produce an ~150-bp product. Barcode diversity should be sufficient for the intended experiment as tested in Step 127.

Lentivirus packaging

Virus should have an appropriate viral titer that ensures 15–50% transduction efficiency and should have enough barcode diversity for the intended experiment as tested in Step 127.

Transduction of experimental cells

After completing Steps 68 and 81, transduction should produce 15–50% GFP expression via flow cytometry to reduce the chance of double barcoding. When testing barcode diversity using control cell lines, a higher percentage of GFP^+ cells is acceptable.

Barcode extraction

qPCR amplification of barcodes should produce a typical exponential curve, which should be absent for the negative control (Fig. 3).

DNA quantification and high-throughput sequencing

Barcode DNA should be ~150 bp and yield >1 ng/μl per sample. High-throughput sequencing results provide one .fastq file for each reverse primer used in the experiment. The file name typically begins with a 6-bp reverse primer index (Table 2).

Analysis of sequencing data

Barcode quantification results will be generated (Fig. 5). In addition, step-2_combine-library-ID.py and step-3_combine-barcodes.py will both generate a stats.txt file for quality check.

In stats.txt files generated in Step 125:

1. ‘% valid reads based on 17bp ending’ should be at least 70–80%.
2. ‘Numbers of reads with expected virus ID’ should be higher than those with unexpected library IDs. (Fig. 5a).

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A sample dataset has been deposited in Figshare: https://doi.org/10.35092/yhjc.11374446. This dataset was used to generate Figs. 4 and 5.

Code availability

The Python scripts have been provided in the Supplementary Software of this paper. The code in this paper has been peer-reviewed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank all members of the Lu lab for helping to optimize the protocol and C. Lytal for help editing the text. We thank the USC Stem Cell Flow Cytometry Facility and CHLA Sequencing Core for technical support. This research was funded primarily by a National Institutes of Health (NIH) R00 early investigator grant (NIH-R00-HL113104) and R01 grants (R01HL135292 and R01HL138225). R.L. is a Scholar of the Leukemia & Lymphoma Society and a Richard N. Merkin Assistant Professor. The project described was supported in part by award no. P30CA014089 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

1. Lu R, Neff NF, Quake SR & Weissman IL Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat. Biotechnol 29, 928–933 (2011). [PubMed: 21964413]
2. 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 42, 71–79 (1985). [PubMed: 4016956]
3. Keller G, Paige C, Gilboa E & Wagner EF Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. Nature 318, 149–154 (1985). [PubMed: 3903518]
4. Lemischka IR, Raulet DH & Mulligan RC Developmental potential and dynamic behavior of hematopoietic stem cells. Cell 45, 917–927 (1986). [PubMed: 2871944]
5. Schepers K et al. Dissecting T cell lineage relationships by cellular barcoding. J. Exp. Med 205, 2309–2318 (2008). [PubMed: 18809713]
6. Lyne A-M et al. A track of the clones: new developments in cellular barcoding. Exp. Hematol 68, 15–20 (2018). [PubMed: 30448259]
7. Naik SH et al. Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature 496, 229–232 (2013). [PubMed: 23552896]
8. Cheung AMS et al. Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. Blood 122, 3129–3137 (2013). [PubMed: 24030380]
9. Nguyen LV et al. Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. Cell Stem Cell 14, 253–263 (2014) [PubMed: 24440600]

10. Gerrits A et al. Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood 115, 2610–2618 (2010). [PubMed: 20093403]
11. Nguyen L et al. Functional compensation between hematopoietic stem cell clones in vivo. EMBO Rep. 19, e45702(2018). [PubMed: 29848511]
12. Brewer C, Chu E, Chin M & Lu R Transplantation dose alters the differentiation program of hematopoietic stem cells. Cell Rep 15, 1848–1857 (2016). [PubMed: 27184851]
13. Wu C et al. Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. Cell Stem Cell 14, 486–499 (2014). [PubMed: 24702997]
14. Lu R, Czechowicz A, Seita J, Jiang D & Weissman IL Clonal-level lineage commitment pathways of hematopoietic stem cells in vivo. Proc. Natl Acad. Sci. USA 116, 1447–1456 (2019). [PubMed: 30622181]
15. Bystrykh LV, de Haan G & Verovskaya E Barcoded vector libraries and retroviral or lentiviral barcoding of hematopoietic stem cells in Hematopoietic Stem Cell Protocols (eds Bunting KD & Qu C-K) 345–360 (Springer New York, 2014).
16. Bystrykh LV & Belderbos ME Clonal analysis of cells with cellular barcoding: when numbers and sizes matter in Stem Cell Heterogeneity: Methods and Protocols (ed Turksen K) 57–89 (Springer New York, 2016).
17. Naik SH, Schumacher TN & Perie L Cellular barcoding: a technical appraisal. Exp. Hematol 42, 598–608 (2014). [PubMed: 24996012]
18. Thielecke L et al. Limitations and challenges of genetic barcode quantification. Sci. Rep 7, 43249 (2017). [PubMed: 28256524]
19. Woodworth MB, Girskis KM & Walsh CA Building a lineage from single cells: genetic techniques for cell lineage tracking. Nat. Rev. Genet 18, 230 (2017). [PubMed: 28111472]
20. Kebschull JM & Zador AM Cellular barcoding: lineage tracing, screening and beyond. Nat. Methods 15, 871–879 (2018). [PubMed: 30377352]
21. Nguyen LV et al. DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. Nat. Commun 5, 5871 (2014). [PubMed: 25532760]
22. Nguyen LV et al. Barcoding reveals complex clonal dynamics of de novo transformed human mammary cells. Nature 528, 267 (2015). [PubMed: 26633636]
23. Wang T, Wei JJ, Sabatini DM & Lander ES Genetic screens in human cells using the CRISPR-Cas9 aystem. Science 343, 80 (2014). [PubMed: 24336569]
24. Shalem O et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343, 84 (2014). [PubMed: 24336571]
25. Dykstra B et al. Long-term propagation of distinct hematopoietic differentiation programs in vivo. Cell Stem Cell 1, 218–229 (2007). [PubMed: 18371352]
26. Sieburg HB et al. The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets. Blood 107, 2311–2316 (2006). [PubMed: 16291588]
27. Osawa M, Hanada K, Hamada H & Nakauchi H Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. Science 273, 242 (1996). [PubMed: 8662508]
28. Weber K, Thomaschewski M, Benten D & Fehse B RGB marking with lentiviral vectors for multicolor clonal cell tracking. Nat. Protoc 7, 839 (2012). [PubMed: 22481527]
29. Cornils K et al. Multiplexing clonality: combining RGB marking and genetic barcoding. Nucleic Acids Res. 42, e56–e56 (2014). [PubMed: 24476916]
30. Livet J et al. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature 450, 56 (2007). [PubMed: 17972876]
31. Rios AC, Fu NY, Lindeman GJ & Visvader JE In situ identification of bipotent stem cells in the mammary gland. Nature 506, 322 (2014). [PubMed: 24463516]
32. Schmidt M et al. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat. Methods 4, 1051 (2007). [PubMed: 18049469]
33. Harkey MA et al. Multiarm high-throughput integration site detection: limitations of LAM-PCR technology and optimization for clonal analysis. Stem Cells Dev 16, 381–392 (2007). [PubMed: 17610368]

34. Wu C et al. High efficiency restriction enzyme-free linear amplification-mediated polymerase chain reaction approach for tracking lentiviral integration sites does not abrogate retrieval bias. Hum. Gene Ther. 24, 38–47 (2013). [PubMed: 22992116]
35. Wu C et al. Tracking retroviral-integrated clones with modified non-restriction enzyme LAM-PCR technology. Mol. Ther 19, S45 (2011).
36. Zhou S et al. Quantitative shearing linear amplification polymerase chain reaction: an improved method for quantifying lentiviral vector insertion sites in transplanted hematopoietic cell systems. Hum. Gene Ther. Methods 26, 4–12 (2014).
37. Sun J et al. Clonal dynamics of native haematopoiesis. Nature 514, 322–327 (2014). [PubMed: 25296256]
38. Pei W et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature 548,456–460 (2017). [PubMed: 28813413]
39. Frieda KL et al. Synthetic recording and in situ readout of lineage information in single cells. Nature 541, 107 (2016). [PubMed: 27869821]
40. Kalhor R et al. Developmental barcoding of whole mouse via homing CRISPR. Science 361, eaat9804 (2018). [PubMed: 30093604]
41. Rufer AW & Sauer B Non-contact positions impose site selectivity on Cre recombinase. Nucleic Acids Res 30, 2764–2771 (2002). [PubMed: 12087159]
42. Shen MW et al. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature 563, 646–651 (2018). [PubMed: 30405244]
43. Lee-Six H et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473–478 (2018). [PubMed: 30185910]
44. Osorio FG et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. Cell Rep. 25, 2308–2316.e4 (2018). [PubMed: 30485801]
45. Chapal-Ilani N et al. Comparing algorithms that reconstruct cell lineage trees utilizing information on microsatellite mutations. PLoS Comput. Biol. 9, e1003297 (2013). [PubMed: 24244121]
46. Wasserstrom A et al. Reconstruction of cell lineage trees in mice. PLoS ONE 3, e1939 (2008). [PubMed: 18398465]
47. McKenzie JL, Gan OI, Doedens M, Wang JCY & Dick JE Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nat. Immunol 7, 1225 (2006). [PubMed: 17013390]
48. Gonzalez-Murillo A, Lozano ML, Montini E, Bueren JA & Guenechea G Unaltered repopulation properties of mouse hematopoietic stem cells transduced with lentiviral vectors. Blood 112, 3138–3147 (2008). [PubMed: 18684860]
49. Wu C et al. Clonal expansion and compartmentalized maintenance of rhesus macaque NK cell subsets. Sci. Immunol 3, aat9781 (2018).
50. Bystrykh LV, Verovskaya E, Zwart E, Broekhuis M & de Haan G Counting stem cells: methodological constraints. Nat. Methods 9, 567 (2012). [PubMed: 22669654]
51. Merino D et al. Barcoding reveals complex clonal behavior in patient-derived xenografts of metastatic triple negative breast cancer. Nat. Commun 10, 766 (2019). [PubMed: 30770823]
52. Guernet A et al. CRISPR-barcoding for intratumor genetic heterogeneity modeling and functional analysis of oncogenic driver mutations. Mol. Cell 63, 526–538 (2016). [PubMed: 27453044]
53. Verovskaya E et al. Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. Blood 122, 523–532 (2013). [PubMed: 23719303]
54. Thielecke L, Cornils K & Glauche I genBaRcode: a comprehensive R package for genetic barcode analysis. Bioinformatics 10.1093/bioinformatics/btz872 (2019).

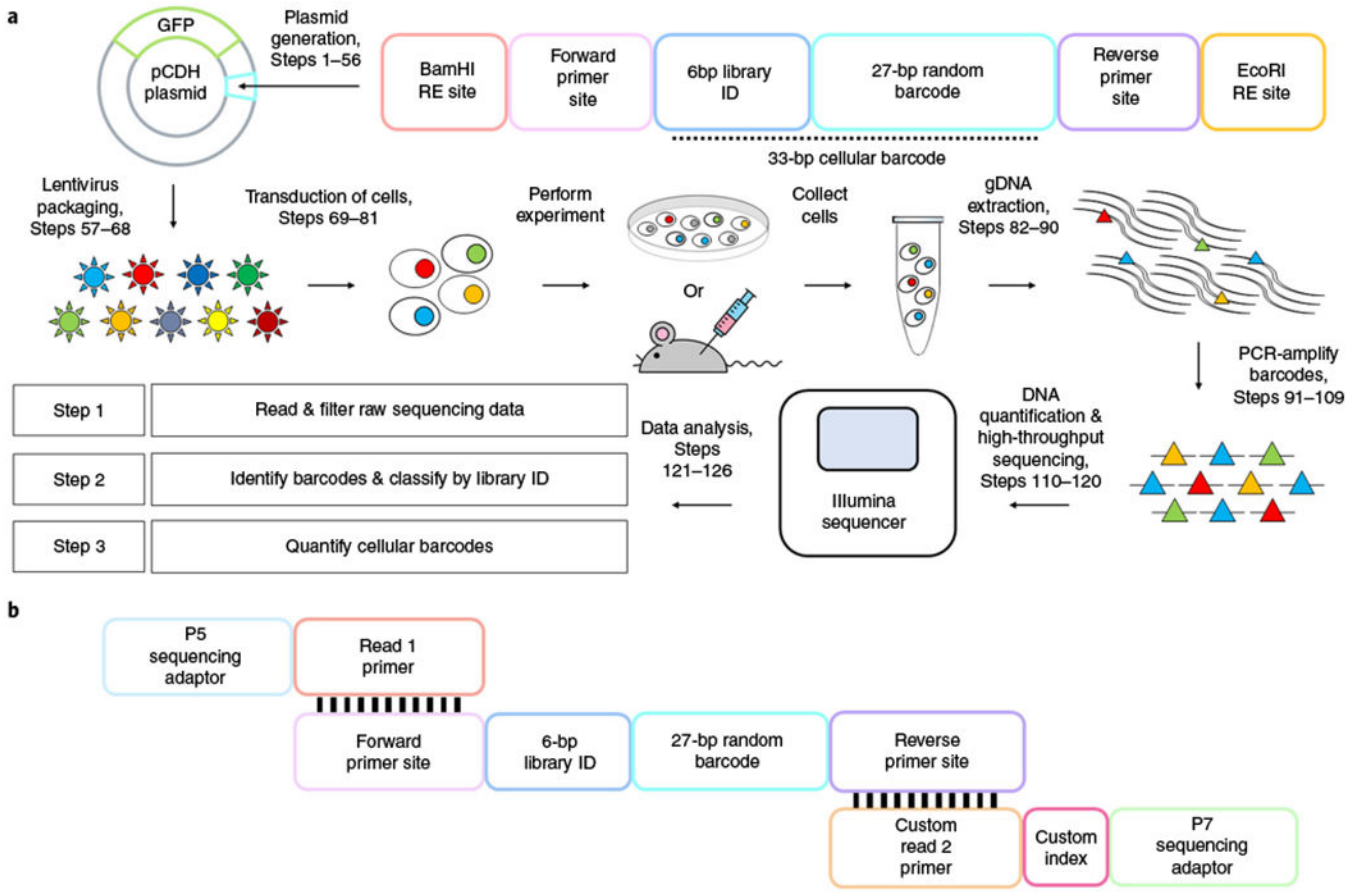

Fig. 1 |. Experiment workflow.

a, Synthesized semi-random barcode oligos (Table 1) are cloned into plasmids before packaging into a lentiviral vector. Cells of interest are then transduced. To retrieve barcodes, genomic DNA is extracted before qPCR amplification and high-throughput sequencing. Raw sequencing data are processed by a custom data analysis pipeline to quantify the abundance of each barcode. **b**, PCR strategy. The 33-bp cellular barcode, comprising a 6-bp library ID and a random 27-bp barcode, is flanked by an Illumina TruSeq read1 sequence and a custom read2 sequence so that a single PCR reaction can add the Illumina P5 and P7 adaptors to the ends of each barcode. See Table 2 for primer sequences. RE, restriction enzyme.

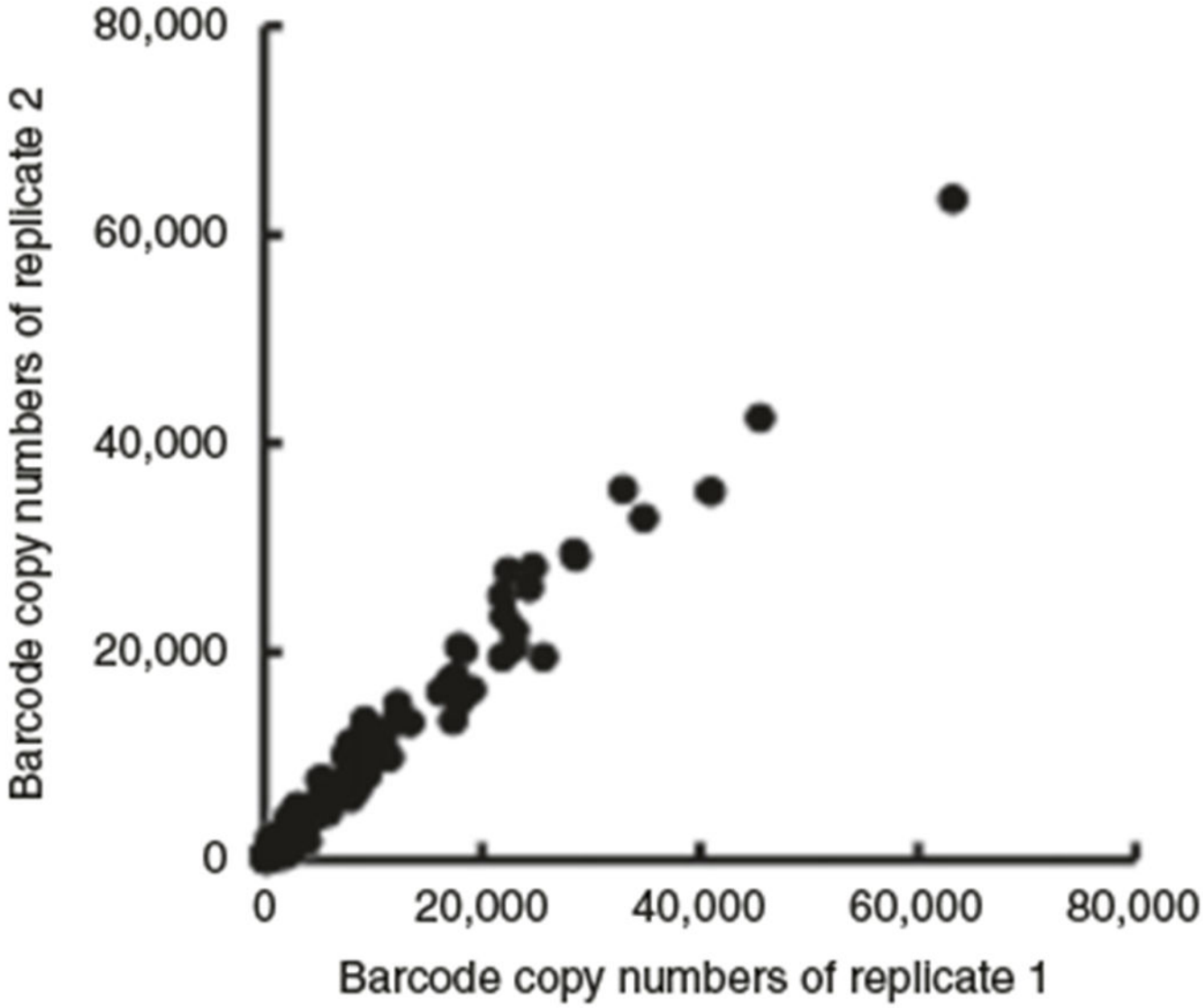

Fig. 2 |. Comparing barcode extraction replicates.

Primary mouse hematopoietic stem cells were barcoded and transplanted into recipient mice. Four months after transplantation, the mice were bled, and white blood cells were collected and processed according to Steps 69-126. Cell lysates were divided into two replicate samples and processed separately for genomic DNA extraction, barcode amplification, and sequencing. Each dot represents a barcode. Barocde abundance is highly consistent between the two replicated samples. Pearson correlation: 0.99; $P = 5.3 \times 10^{-144}$. Animal procedures were approved by the Institutional Animal Care and Use Committee of the University of Southern California, and the mice were maintained at USC's Research Animal Facility.

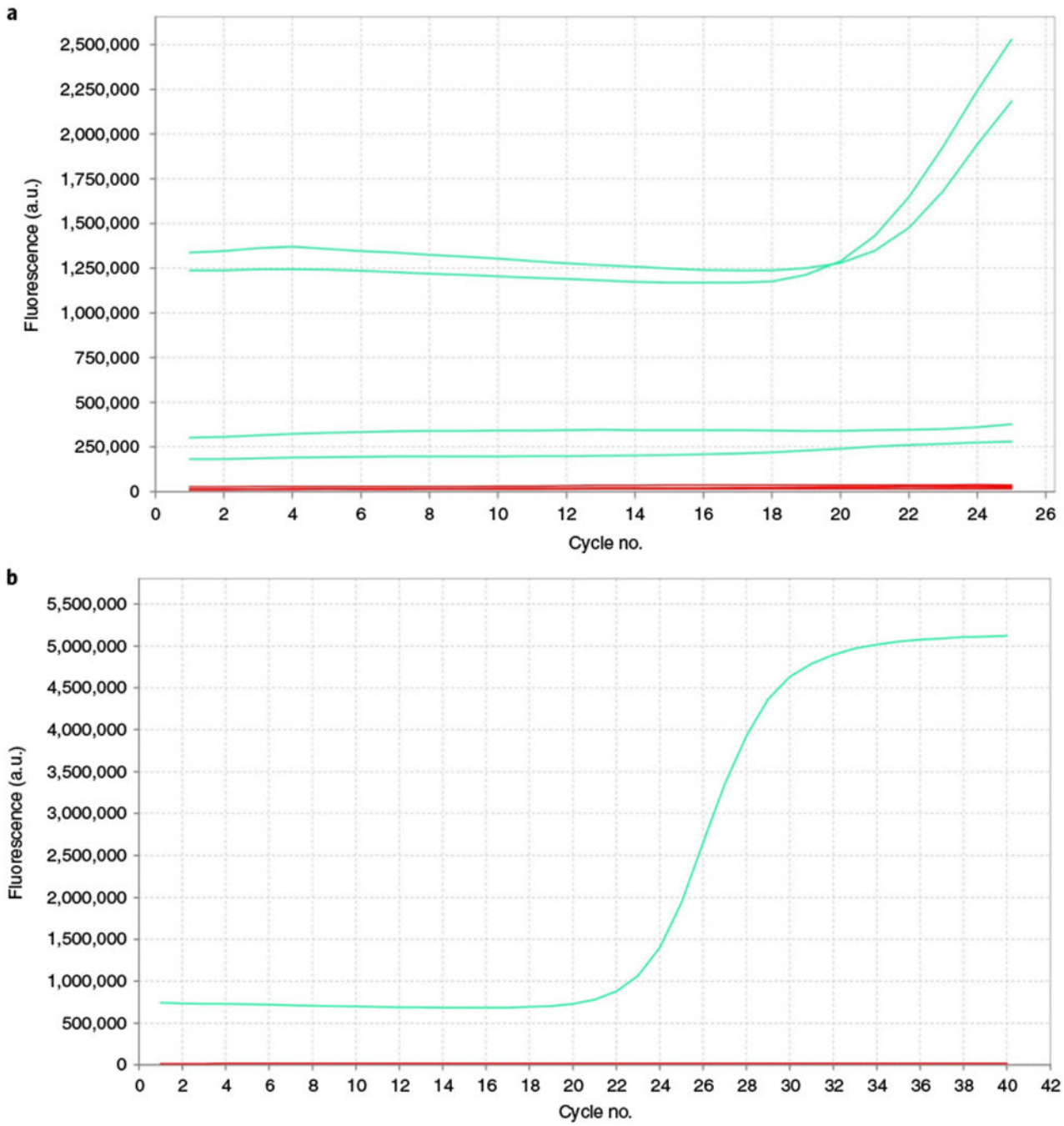

Fig. 3 |. qPCR amplification of barcode.

BB88 cells were barcoded and 50,000 GFP$^+$ cells were sorted via FACS 1 week after transduction. gDNA was isolated and amplified using primers from Table 2. A multi-component plot of barcode amplification is shown. EvaGreen fluorescent dye (green lines) was used to quantify DNA amounts; thus no Rox signal was observed (red lines). **a**, Two samples with similar amounts of genomic DNA were amplified together, and their exponential curves emerged at similar numbers of PCR cycles. We stopped the reaction at cycle 25, which is about halfway through the exponential phase. This was to avoid over-

amplification and to reduce background signals. No-DNA template control samples showed no amplification (two flat green lines). **b**, One sample was amplified to saturation. This is an example of over-amplification. a.u., arbitrary units.

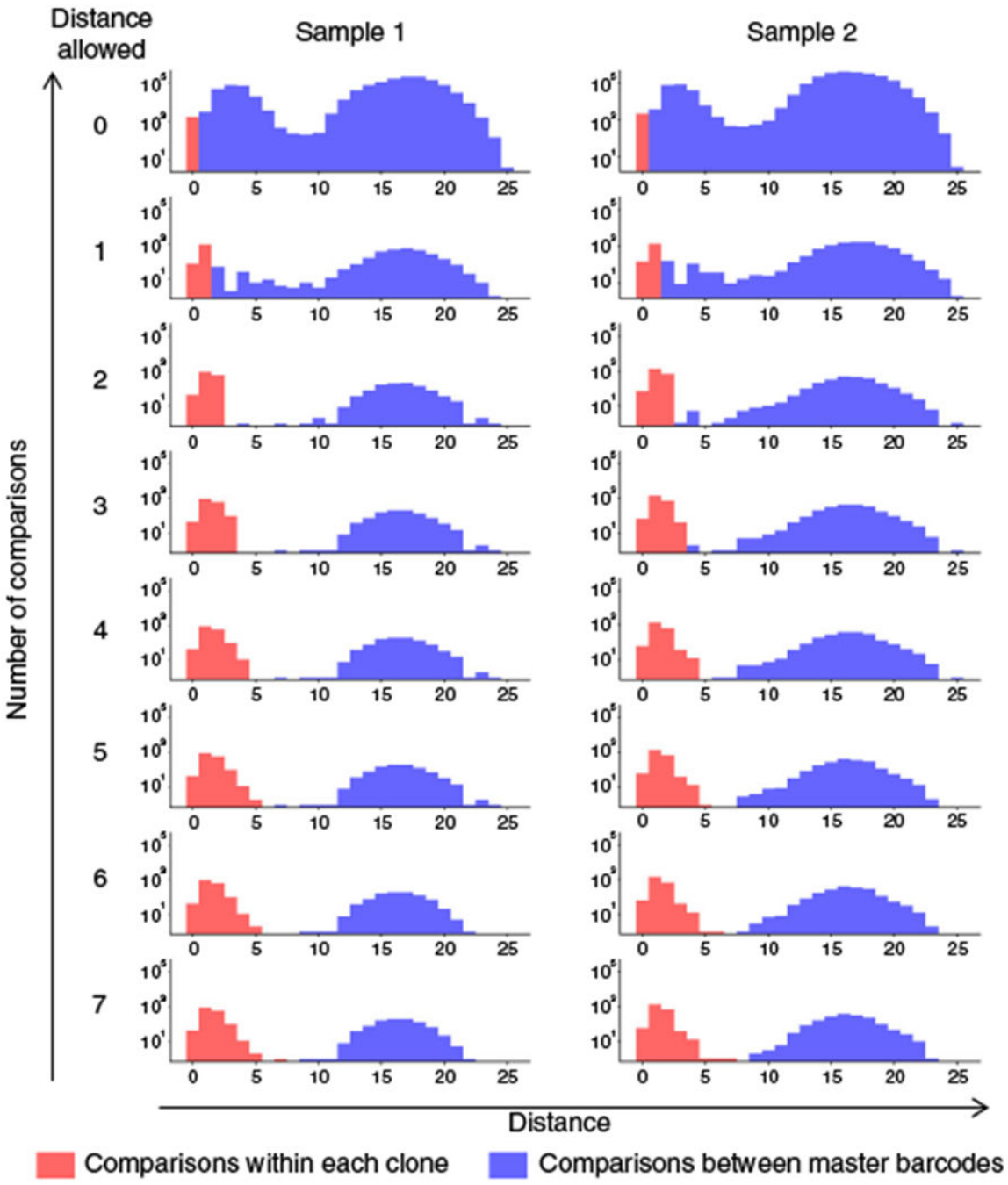

Fig. 4 |. Optimizing the edit distance thresholds.

Histograms show the distances between unique sequences and their corresponding master barcodes (red), as well as the distances between different master barcodes (blue). Each row shows one edit distance threshold. Data from two independent samples are shown in the two columns. The threshold of edit distance of 4 was chosen as the point at which the distances between master barcodes are higher than and separated from the distances between unique sequences and their master barcodes.

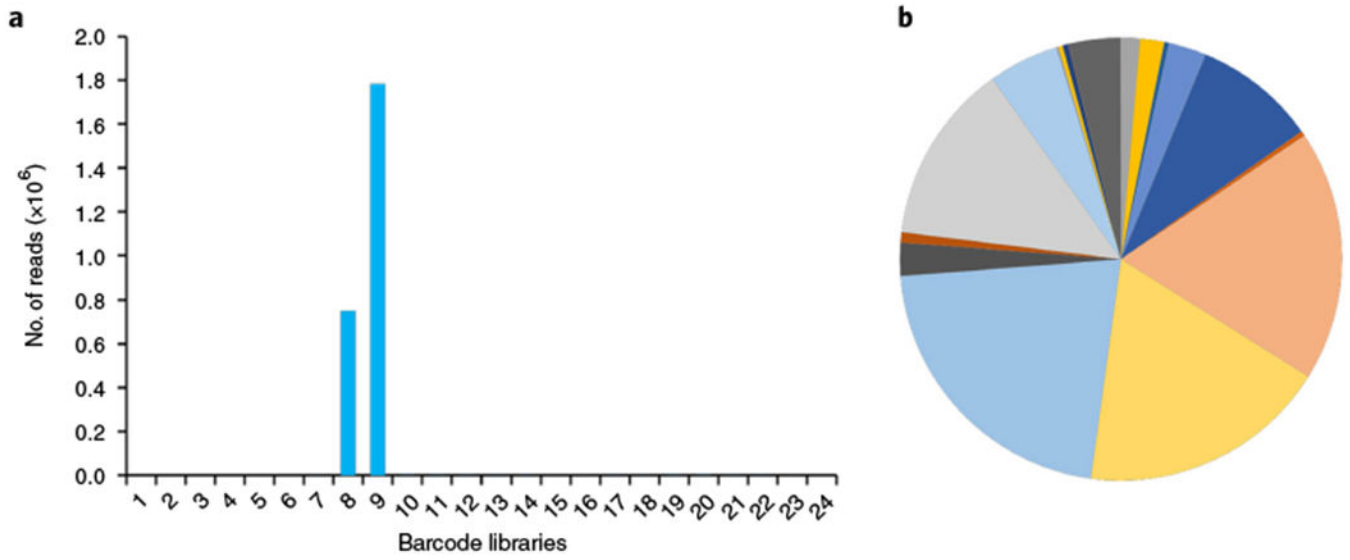

Fig. 5 |. Python pipeline outputs.

Primary human acute lymphoblastic leukemia (ALL) cells were barcoded and transplanted into non-obese diabetic scid-gamma (NSG) mice. Two months after transplantation, the mice were bled, and ALL cells were collected and processed according to Steps 69–126. ALL cells barcoded with virus libraries 8 and 9 were used for this example. **a**, Custom algorithms written in Python code group reads on the basis of their library IDs. **b**, The Python algorithm quantifies each barcode with consideration to sequencing errors. Each color represents a unique barcode, and the size represents its relative abundance. Shown are data from library 9 in Fig. 5a. Animal procedures were approved by the Institutional Animal Care and Use Committee of the University of Southern California, and the mice were maintained at USC's Research Animal Facility.

Table 1 |

Barcode oligos for each library ID

Virus library	Library ID	DNA oligo (5'-3')
1	CGTGAT	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
2	ACATCG	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
3	GCCTAA	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAANNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
4	TGGTCA	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCANNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
5	CACTGT	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTCACTGTNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
6	ATTGGC	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTATTGGCNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
7	GATCTG	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCTGNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
8	TCAAGT	GCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAAGTNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
9	CTGATC	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATCNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
10	AAGCTA	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCTANNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
11	GTAGCC	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAGCCNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
12	TACAAG	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTTACAAGNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
13	ATGACA	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTATGACANNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
14	AGCGGT	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCGGTNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
15	ACTCAG	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTACTCAGNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
16	TAACGT	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTTAACGTNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
17	TGTTAC	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTTACNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
18	TCCGTA	GCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCGTANNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
19	GAGTTC	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGTTCNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
20	GTCGAG	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCGAGNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
21	GCAACT	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAACTNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
22	CAGTGC	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTGCNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
23	CTTATG	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTATGNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG
24	CGACCT	CGCCGCGGATCCACACTCTTTCCCTACACGACGCTCTTCCGATCTCGACCTNNNNNNNNNNNNNNN NNNNNNNNNNNNNNAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAATTCCGGCG

The core of each oligo consists of a 6-bp library ID and a 27-bp random sequence represented as Ns. The core is flanked by forward and reverse primer binding sites, as well as by restriction enzyme sites. An additional 6-bp sequence is added at both ends to ensure proper restriction enzyme cutting.

Table 2 |

Primer list

Procedure	Primer	Sequence (5’-3’)
Plasmid generation	Strand2	CGCCGGAATTCCAAGCAGAAGACGGCATACGA
qPCR	Forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCC TACACGACGCTCTTCCGATCT
	R1 (GCCAAT)	CAAGCAGAAGACGGCATACGAGATGCCAATACGGCAT ACGAGCTCTTCCGATCT
	R2 (GATCTG)	CAAGCAGAAGACGGCATACGAGATGATCTGACGGCAT ACGAGCTCTTCCGATCT
	R3 (TCAAGT	CAAGCAGAAGACGGCATACGAGATTCAAGTACGGCAT ACGAGCTCTTCCGATCT
Sequencing	Custom index primer	AGATCGGAAGAGCTCGTATGCCGT

Table 3 |

Troubleshooting table

Step	Problem	Possible reason	Possible solution
42	<100 colonies in 1:10,000 dilution	Poor bacterial transformation efficiency	Improve the transformation efficiency during cloning
56/127	Step-4-opt_evaluating_barcode_diversity output: 'No, please increase library diversity'	Plasmid library barcode diversity is too low for desired cell number	Redo transformation Re-ligate DNA oligo library into vector Reorder DNA oligos from manufacturer
68/127	Step-4-opt_evaluating_barcode_diversity output: 'No, please increase library diversity'	Lentiviral library barcode diversity too low for desired cell number	Repackage lentiviral library
68,81	GFP^+ percentage < 15% GFP+ percentage > 50%	Low transduction efficiency Transduction efficiency is too high	Increase the viral titer Experiment should be stopped. Set up a new transduction and incubate the cells with virus for a shorter time or reduce the viral titer
90	Low yield of genomic DNA (<1 ng/μL)	Too few barcoded cells to start with	Start with at least 10,000 barcoded cells
93	Amplification beyond the exponential phase No amplification curve after 27 cycles	PCR was terminated too late Missing fluorescent dye, such as EvaGreen Too few barcoded cells in this sample	Repeat the qPCR reaction Add a fluorescent dye If the sample does not have enough barcoded cells, it may not be worth further investigation
	Curve plateaus before fluorescence increases 900,000 a.u.	Too much template	Use no more than 1,000 ng of the template for qPCR
109	Low DNA yield (<1 ng/μL) after bead purification	Incorrect volume of beads; bead pellet dried out	Repeat the qPCR and bead purification steps
120	Number of reads is less than expected	Poor quantification of DNA library	Use a reliable method to quantify DNA concentration
	Failed to demultiplex	Incorrect index primer used for sequencing	Re-sequence the library using the correct custom index primer provided in Table 2
125	Many reads seen with unexpected library IDs	Other cell sources present in the sample	Check the experimental setup, modify the sample information file, and rerun the data analysis
	More 'combined codes' than expected	Misread allowance was set too stringently	Set different values for 'distance_allowed' in the code. Use '4' by default