Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Protocol for the use of Polylox – endogenous barcoding for high resolution in vivo lineage tracing

Thomas Höfer (t.hoefer@dkfz.de)
Division of Theoretical Systems Biology, German Cancer Research Center and Bioquant Center, University of Heidelberg, D-69120 Heidelberg

Hans-Reimer Rodewald (hr.rodewald@dkfz.de)
Division of Cellular Immunology, German Cancer Research Center, D-69120 Heidelberg

Weike Pei
Division of Cellular Immunology, German Cancer Research Center, D-69120 Heidelberg,

Thorsten B. Feyerabend
Division of Cellular Immunology, German Cancer Research Center, D-69120 Heidelberg,

Jens Rössler
Division of Theoretical Systems Biology, German Cancer Research Center, D-69120 Heidelberg, Germany

Xi Wang
Division of Theoretical Systems Biology, German Cancer Research Center, D-69120 Heidelberg, Germany

Daniel Postrach
Division of Cellular Immunology, German Cancer Research Center, D-69120 Heidelberg,

---

Method Article

Keywords: barcoding, Polylox, fate mapping, stem cells, single molecule real time sequencing, SMRT sequencing, Cre recombinase, tamoxifen, barcode sequence alignment, generation probabilities, Markov model

Posted Date: September 18th, 2017

DOI: https://doi.org/10.1038/protex.2017.092

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

*Authors Weike Pei, Thorsten B. Feyerabend, Jens Rössler and Xi Wang contributed to this work equally. This protocol describes methods for the generation of highly variable genetic markers in cells in vivo and the experimental and computational procedures for retrieving those barcodes from cells. The method is based on Cre-recombinase dependent _Polylox_ barcoding[1]. In cells or mice carrying the artificial _Polylox_ DNA recombination locus in their germline \(_Rosa26^{Polylox}_), Cre recombinase can induce DNA recombination in the _Polylox_ locus. The barcodes are retrieved by PCR amplification of the _Polylox_ locus from purified cell populations or from single cells, followed by long-range DNA sequencing of the barcode libraries. The recombined _Polylox_ sequences are computationally analyzed to yield barcodes. The generation probability of barcodes is computed to allow filtering for rare barcodes.

Introduction

The development of organs and tissues depends on the differentiation of progenitor cells into mature functional cell populations. In the blood and immune system, hematopoietic stem cells \(HSC) give rise to more than 10 different lineages, yet the ‘structure’ of hematopoiesis and the underlying pathways have remained enigmatic. One of the most direct methods for the in vivo identification of precursor-product relationships is genetic fate mapping \(also termed lineage tracing). A heritable reporter gene is switched on in cells of a specific phenotype or stage, and this marker is perpetuated in all daughter cells from the initially marked cells. Lineage tracing is commonly done using fluorescent protein expression. However, numbers of distinct colors are limited. To achieve higher resolution, which would be required for tissue deconvolution in complex organs, we developed the _Polylox_ system, allowing genetic in-vivo-labeling of cells that is based on variable DNA recombinations. An artificial DNA cassette consisting of ten _loxP_ sites in alternating orientation and nine intervening DNA segments of known sequence are recombined upon transient Cre recombinase expression, with the DNA segments being inverted or excised depending on the orientation of the _loxP_ sites recognized by Cre. The DNA segments of the _Polylox_ locus are called ‘1’-‘9’ in their original orientation and ‘A’-‘I’ when inverted. The position and orientation of the segments translates into the so-called _Polylox_ barcode. While Cre is expressed, the shuffling of the _Polylox_ segments can progress repeatedly, offering a theoretical diversity of ~1.8 mio. codes and a practical diversity \(we identified a locus that had undergone at least 6 subsequent recombination events) of more than 600 thousand different recombined barcodes. The barcodes are thus serving as cellular identifiers for high-resolution lineage tracing and are retrieved for analysis either by classical Sanger sequencing from individual cells, or from bulk populations by SMRT sequencing technology \(Pacific Biosciences). For the determination of lineage relationships, barcode distributions between different cellular populations are compared, and correlation coefficients are calculated as indicators of lineage proximity. Important considerations for the successful use of the _Polylox_ barcoding system are choice of Cre and calculation of barcode generation probability. A suitable Cre allele has to be inducible and tightly controlled for transient but not leaky expression. We have so far successfully tested two tamoxifen inducible Cre alleles, the _Rosa26^{CreERT2}_ \(_B6.129-Gt\(ROSA)26Sor$^{tm1\(cre/ESR1)Tyj/J}$_) for

ubiquitous labeling and _Tie2^{MerCreMer}_ \(_Tek$^{tm1.1\(icre/Esr1*)Hrr}$_) for barcode induction in hematopoietic stem cells. Second, _Polylox_ barcodes are not all generated with the same probability. Especially those that require only one recombination event are more likely to be generated than more complex barcodes for which multiple recombination steps are required. This protocol therefore also provides a method for calculating barcode generation probabilities.

Reagents

Reagents for barcode induction Tamoxifen \(Sigma-Aldrich, T5648) Progesterone \(Sigma-Aldrich, P0130) Peanut oil \(Sigma-Aldrich, P2144) **Reagents for organ preparation** DPBS \(Sigma, D8537-500ML) FBS \(Fetal Bovine Serum, Biochrom, S0115) Ficoll-Paque PLUS \(GE Healthcare, 17-1440-02) Cell Strainer 40-µm mesh \(Falcon, REF 352340) Plastic Pasteur pipettes \(Sarstedt, 86/1175-001) **Reagents for purifications of cells by FACS sorting** Mouse IgG \(Jackson ImmunoResearch Laboratories, 015-000-003) Antigen-specific fluorescent labeled antibodies SYTOX Blue \(Life Technologies, S34857) 5 ml tubes with 35 µm cell strainer snap cap \(FALCON, REF352235) **Primer and PCR reagents** Primer #HL14 \(5'-AATCAAGGGTCCCCAAACTCAC-3') Primer #972 \(5'-GAGGCAGCATCTGTCTACAAGAGATGG-3') Primer #984 \(5'-CATCGCATACCATACATAGGTGGAGG-3') Primer #494 \(5'-AGCTACAGCCTCGATTTGTGGTG-3') Primer #2426 \(5'-CGACGACACTGCCAAAGATTTC-3') Primer #2427 \(5'-CATACCTTAGAGAAAGCCTGTCGAG-3') Primer #2450 \(5'-TGTGGTATGGCTGATTATGATCAG-3') Expand Long Template PCR system \(Roche, 11759060001) Proteinase K \(Thermo Fisher Scientific, 25530-049) dNTPs \(Biozym, R0192)

Equipment

Automated cell counter \(e.g. Cellometer Auto 2000 \(Nexcelom)) Cell sorter: BD FACS AriaIII \(BD Biosciences) PCR Thermocycler \(e.g. T3000 Thermocycler, Biometra) Qubit Fluorometer \(Invitrogen) Bioanalyzer \(Agilent) PacBio RS II or PacBio Sequel \(Pacific Biosciences)

Procedure

I. Barcode induction _a) Embryonic mice_ 1. Set up timed matings between _Rosa26$^{Polylox/Polylox}$_ and _Tie2$^{MCM/+}$_ mice. 2. Nine days after the day of the plug \(day 0.5), treat pregnant mice by oral gavage with a single dose of 2.5 mg tamoxifen and 1.25 mg progesterone to induce barcoding in the developing embryos at E9.5. Tamoxifen stock solution is prepared by dissolving 1 g tamoxifen in 4 ml absolute ethanol and 36 ml peanut oil at 55 °C. Stock solution can be stored at -20 °C. 3. The pups are delivered on E20.5 by caesarean section and raised by foster mothers. 4. Genotype the pups for the _Tie2^{MCM}_ allele at 3-4 weeks after birth on genomic DNA prepared from tail biopsies using primers #972, #984 and #HL14. PCR conditions are as follows: 2 min at 94 °C; \(20 s at 94 °C, 30 s at 62 °C, 1 min at 72 °C) 35 times; 5 min at 72 °C. 5. Successful tamoxifen induction of _Rosa26$^{Polylox/+}$Tie2$^{MCM/+}$_ mice is tested by PCR amplification of the _Polylox_ locus from genomic DNA of Ficoll-separated peripheral blood leukocytes.

The _Polylox_ PCR is performed with the Expand Long Template PCR system using primers #2450 and #2427, annealing at the 5’ and 3’ anchor regions of the _Polylox_ cassette, respectively. PCR conditions are as follows: 5 min at 95 °C; \(30 s at 95 °C, 30 s at 56 °C, 5 min at 72 °C) repeat 35 times; 10 min at 72 °C. Separate PCR products by gel electrophoresis. _b) Adult mice_ 1. Prepare tamoxifen stock solution by dissolving 1 g tamoxifen in 4 ml absolute ethanol and 36 ml peanut oil at 55 °C. 2. Adult _$Rosa26^{Polylox/+}Tie2^{MCM/+}$_ mice are injected intraperitoneally with 1 mg tamoxifen on five consecutive days. 3. Successful tamoxifen induction of _$Rosa26^{Polylox/+}Tie2^{MCM/+}$_ mice is tested by PCR amplification of the _Polylox_ locus as described above \(Section Ia 5.) **II. Organ preparation** On the day of analysis, mice are sacrificed by carbon dioxide inhalation. Peritoneal cells should be harvested by peritoneal lavage before any other abdominal organs are prepared. _a) Peritoneal Cavity_ 1. Harvest peritoneal exudate cells by lavage of the peritoneal cavity for 5 times with 2 ml of 5% FACS buffer \(DPBS with 5% vol/vol FBS). 2. Small opening and gentle lavage with plastic Pasteur pipettes is helpful to avoid injury and blood cell contamination. _b) Spleen_ 1. Take out the spleen and cut the organ into small pieces in 5% FACS buffer using surgical scissors. 2. To obtain a single cell suspension of splenocytes, the minced spleen is passed through a 40-µm filter using 5% FACS buffer and the plunger of a 5-ml syringe. _c) Bone Marrow_ 1. Prepare femora and tibiae \(humerus, pelvis and vertebrae can also be included to maximize bone marrow recovery). 2. Bone marrow cells are either flushed from the long bones with 5% FACS buffer, or all bones are crashing in a mortar in 5% FACS buffer. 3. Resuspend the extracted bone marrow cells by repeated pipetting and remove remaining bone pieces by filtering through a 40-µm mesh. **III. Purification of cells by FACS sorting** 1. Determine the concentration of the cell suspensions, e.g. on an automated cell counter. This was done using Cellometer Auto 2000 Cell Viability Counter according to the manufacturer’s protocol. 2. Block Fc receptors by incubation of the cells \($5x10^6$ cells/50 µl) for 15 min with 300 µg/ml whole mouse IgG. 3. Add cocktails with titrated concentrations of fluorescent dye-labeled antibodies in 5% FACS buffer and stain the cells for 45 min in the dark on ice. Afterwards, wash cells in an appropriate amount of 5% FACS buffer. 4. For dead cell exclusion e.g. add 100 nM SYTOX Blue at least 5 min before analysis. 5. Cell suspensions should be filtered directly to remove any cell clumps before being placed on the FACS sorter, e.g. by passing the suspension through the mash of a snap cap cell strainer. 6. Stained cells are analyzed on a FACSAriaIII cell sorter running BD FACSDiva software and populations of interest are sorted into 20% FACS collection buffer \(DPBS containing 20% FBS). Alternatively, cells can directly be sorted into PCR buffer. 7. Aliquots of the sorted cells should be reanalyzed to determine sort purity. **IV. PCR amplification of the _**Polylox**_ locus** _a) From single cells via nested PCR_ 1. By FACS sorting individual cells are deposited into 8-tube PCR stripes containing in each well 25 µl of 1x PCR buffer supplemented with 12.6 µg Proteinase K. 2. Perform lysis for 1 h at 55 °C, terminate at 95 °C for 10 min, and cool down to 4 °C before adding the remaining PCR reagents to a final volume of 50 µl. 3. Amplification of the _Polylox_ cassette is done by nested PCR using the Expand Long Template PCR system: - First round PCR: primer #2450 and primer #494 for 5 min at 95 °C; \(30 s at 95 °C, 30 s at 56 °C, 5 min at 72 °C) repeat 35 times; 10 min at 72 °C. - Second round PCR: Use 1-2 µl of the first PCR reaction as template and amplify with primers #2426 and #2427 for 5 min at 95 °C; \(30 s at 95 °C, 30 s at 62 °C, 5 min at 72 °C) 35 times; 10 min at 72 °C. 4. The nested PCR products are analyzed by gel electrophoresis for product length, and sent for Sanger sequencing \(e.g. GATC Biotech). _b) Cell

populations \(10,000-50,000 cells) by direct PCR_ 1. Sort cells into 1.5 ml Eppendorf tubes containing 300 µl of 20% FACS buffer. 2. Use 2-5% of the sorted cells for reanalysis and harvest the remaining cells by centrifugation for 5 min at 1500 g and 4 ºC. 3. Turn the tube 180° in the rotor and spin again for 5 min at 1500 g. 4. Carefully remove the supernatant and resuspend with 25 µl lysis buffer \(1x PCR buffer 1 from the Expand Long Template PCR system in water and 12.6 µg Proteinase K). Digest at 55 °C for 1 hour and terminate the reaction at 95 °C for 10 min. DNA can be at -20 °C, but the experiment should continued swiftly. 5. PCR amplification of the _Polylox_ locus is as described above \(Section Ia 5.) 6. Visualize PCR products \(one third of the PCR reaction) by gel electrophoresis. _c) Cell populations \(>50,000 cells) via phenol-chloroform DNA extraction_ 1. Sort cells into 2 ml 20% FACS collection buffer and harvest the cell pellet by centrifuge at 400 g for 10 min after reanalysis for sort purity. 2. Carefully remove the supernatant leaving 50-100 µl of the supernatant to avoid disturbing of the cell pellet. 3. Resuspend the pellet in 500 µl freshly prepared proteinase K-solution \(10 mM Tris-HCl pH 8, 5 mM EDTA, 1% SDS, 0.3 M Na-acetate, 0.2 mg/ml proteinase K) and incubated at 37 °C for at least 3 hours. 4. From the lysate, purify genomic DNA by phenol-chloroform extraction. Briefly, extract once with an equal volume of a 1:1 mixture of phenol and chloroform. Then extract the aqueous phase at least twice with an equal volume of chloroform. Precipitate the DNA by adding 1/10 volume of Na-acetate \(3M) pH 5 and 1 volume of isopropanol and 5-20 µg glycogen for 1 hour at –20 °C. Spin at 16000 g for 10 min and wash the DNA pellet twice with 75% ethanol. Dissolve the air-dried pellet in water. Measure DNA concentration. 5. The _Polylox_ cassette is readily amplified from 100-200 ng of template DNA \(representing 1.7 - 3.5 x10^4 cells) by PCR using the conditions described above \(Section Ia 5.) 6. PCR products \(one third of the PCR reaction) are observed by gel electrophoresis. **V. _**Polylox**_ Sequencing** _a) Sanger sequencing from single cells_ _Polylox_ amplicons from single cells were sequenced by primer walking, using a commercial service \(e.g. GATC biotech, Germany). PCR products were either purified by QIAquick PCR Purification Kit, or DNA clean up was done by the provider of the sequencing service. _b) SMRT sequencing from cell populations_ 1. Use the entire PCR product that is left after gel electrophoresis for an initial AMPure PB bead clean up step. Perform according to the manufacturer’s protocol. 2. For quality control determine DNA concentration by Qubit measurement and perform a Bioanalyzer run according to the manufacturers’ protocols. 3. Continue library preparation for PacBio SMRT sequencing according to the “Amplicon Template Preparation and Sequencing” protocol released by the company. http://www.pacb.com/wp-content/uploads/Procedure-Checklist-Amplicon-Template-Preparation-and-Sequencing.pdf 4. Perform primer annealing and polymerase binding according to the manufacturer’s protocol. 5. Load the library by Magbead mode \(PacBio RS II) or diffusion mode \(PacBio Sequel) for SMRT sequencing. We usually recorded for 4 hours. **VI. Barcode identification** The SMRT sequencing yields data files with CCS reads. The following part explains the retrieval of _Polylox_ barcodes from PacBio CCS Reads \(RPBPBR). _a) Software and hardware_ • Data \(PacBio CCS reads, either in fasta or fastq format) • _Polylox_ adapters and segments \(provided in the data folder of RPBPBR toolkit) • Bowtie2 software \(http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) • SAMtools \(http://www.htslib.org/) • RPBPBR toolkit \(https://github.com/sunlightwang/RPBPBR) • Hardware: Computer running either Linux or Mac OS X \(10.6 Snow Leopard or later; at least 4 GB of RAM \(8 GB per core preferred); at least quad-core CPU • Perl \(https://www.perl.org/) by default already installed in Linux

or Mac OS X computers _b) Software setup_ Commands given in the protocol are runnable at the UNIX shell prompt, which are prefixed with a '$' character. _To install the SAMtools_, download the SAMtools \(http://www.htslib.org/download/) and unpack the SAMtools tarball: $ tar jxvf samtools-1.5.tar.bz2 Then cd to the SAMtools source directory and build the samtools binary $ cd samtools-1.5 $ ./configure –prefix=/path/to/install $ make $ make install Copy the samtools binary to some directory in your PATH \(e.g. $HOME/bin): $ cp samtools $HOME/bin Or add the directory containing samtools binary to your PATH environment variable $ export PATH=/path/to/install/bin:$PATH _To install Bowtie2_, download the latest binary package for Bowtie2 \(https://sourceforge.net/projects/bowtie-bio/files/bowtie2/) and unpack the Bowtie2 zip archive: $ unzip bowtie2-2.3.2-legacy-macos-x86_64.zip Copy the Bowtie executables to a directory in your PATH \(e.g. $HOME/bin): $ cd bowtie2-2.3.2-legacy $ cp bowtie2* $HOME/bin Or add the directory containing bowtie2 binaries to your PATH environment variable $ export PATH=/path/to/bowtie2/binary/directory:$PATH _To install RPBPBR toolkit_, clone the latest binary package from RPBPBR github site \(https://github.com/sunlightwang/RPBPBR/) $ git clone https://github.com/sunlightwang/RPBPBR.git Add the directory containing RPBPBR binaries to your PATH environment variable $ export PATH=/path/to/RPBPBR/bin/:$PATH _c) Procedure _ _To run RPBPBR on the example data files_, cd to the RPBPBR example directory $ cd /path/to/RPBPBR/example Then execute RPBPBR on each example file: $ RPBPBR test1.fastq test1 fastq $ RPBPBR test2.fa test2 fasta RPBPBR takes PacBio CCS reads \(in either fasta or fastq format) and directly reports the number of barcodes in the PacBio library of interest for downstream analysis. By default, RPBPBR takes 4 cores per process; however, the number of cores is adjustable in the script. Using 4 cores, the running time of RPBPBR varies from < 1 hour to several hours depending on the amount of reads to be processed. Usage: RPBPBR < input.fasta/fastq > < out.prefix > < type:fasta/fastq > \[keep-temp] < input.fasta/fastq > required, the PacBio read file in fasta or fastq format. < out.prefix > required, the prefix of output file, and also the name of a temporary directory to be created during the process. < type:fasta/fastq > required, the format of the PacBio read file, only can be fasta or fastq, other formats not acceptable. \[keep-temp] optional, if not specified or with value 0, the temporary directory created during the process will be removed after the process is done; otherwise, it will be kept. _d) Output _ Output file name: < out.predix >.barcode.count.tsv Output file is a tabular text file, each line gives the count \(in the second column) of each barcode listed in the first column. Total: total PacBio reads that have been processed. Intact: the number of PacBio reads with both 5' and 3' adapter sequences. Barcodes*: starting from 5' and end with 3', barcode segments are connected with hyphens. * In the barcode string, X represents non-recognized segments due to low sequencing quality. **VII. Barcode filtering** This section explains the computation of barcode generation probabilities based on which rare barcodes can be identified. _a) Software_ • Data \(Excel Sheet from RPB, see VI. Barcode identification) • Matlab \(R2013b) • barcode_pipeline.m script • data.mat o barcode_library _// all possible barcodes_ o path_matrix _// Pgen for a distinct # of recombinations_ o min_list _// list of minimal number of recombinations_ _b) Software setup_ Copy barcode_pipeline.m and data.mat in desired path. Import the following files from experimental data into your Matlab Workspace: • A list of found barcodes in cell format • A m by n matrix of reads, m is the number of barcodes, n the number of populations • A list of population names in cell format _c) Procedure_ To run barcode_pipeline.m type: barcode_pipeline \(< found_codes >, < found_reads >,<

population_annotation >) The pipelines running time varies between < 1 min to 10 minutes depending on the number codes. Barcode_pipeline first checks all experimentally found barcodes with the complete list of possible barcodes, potentially purging erroneous barcodes from the list. Then the minimal number of recombinations for every barcode is determined \(look up in pre-calculated table). From this minimal recombination number a frequency distribution of recombination events is calculated, which is then used to calculate the generation probabilities for all barcodes. _d) Output_ Output is a Matlab-structure with the following fields: • purged_codes _// list of all possible codes in the data-set_ • purged_reads _// matrix containing the reads_ • purged_freq _// frequencies – normalized for populations_ • recombination_frequencies _// distribution of recombinations 0-10_ • minimal_recom _// list of minimal recombinations for all codes_ • pgen _// Pgen for every code_ • annotation _// population-annotation list_

Timing

In vivo barcode induction takes 1-5 days, depending on the tamoxifen dosage and duration of application. Cell purification via FACS sorting is done within one day. DNA preparation and PCR amplification take one day. Library preparation and SMRT sequencing require 4-5 days. Computational analysis requires approximately 1 day.

Troubleshooting

To achieve a useful range of recombination, the specified tamoxifen dose may need to be adjusted depending on Cre driver, developmental stage and tissue to be labeled. This can be evaluated by the fragmentation pattern of the _Polylox_ locus after tamoxifen treatment. Cell harvest and _Polylox_ PCR are the most critical steps. We strongly recommend to test _Polylox_ PCR on different cell numbers before starting experiments.

Anticipated Results

This protocol yields the recovered barcodes in the sampled cell populations, along with read counts and barcode generation probabilities.

References

1. W. Pei, T. B. Feyerabend, J. Rössler, X. Wang, D. Postrach, K. Busch, I. Rode, K. Klapproth, N. Dietlein, C. Quedenau, W. Chen, S. Sauer, S. Wolf, T. Höfer, and H.R. Rodewald. _Polylox_ barcoding reveals hematopoietic stem cell fates realized in vivo. Nature 2017