Cell Stem Cell
Resource

# Clonal Analysis via Barcoding Reveals Diverse Growth and Differentiation of Transplanted Mouse and Human Mammary Stem Cells

Long V. Nguyen,[1,7] Maisam Makarem,[1,7] Annaick Carles,[2] Michelle Moksa,[2] Nagarajan Kannan,[1] Pawan Pandoh,[3] Peter Eirew,[4] Tomo Osako,[4] Melanie Kardel,[1] Alice M.S. Cheung,[5] William Kennedy,[1] Kane Tse,[3] Thomas Zeng,[3] Yongjun Zhao,[3] R. Keith Humphries,[1] Samuel Aparicio,[4,6] Connie J. Eaves,[1,8,*] and Martin Hirst[2,3,8,*]
[1]Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada
[2]Department of Microbiology and Immunology, Centre for High-Throughput Biology, University of British Columbia, 2125 East Mall, Vancouver, BC V6T 1Z4, Canada
[3]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada
[4]Department of Molecular Oncology, BC Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada
[5]La Ka Shing Faculty of Medicine, University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong
[6]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC V6T 2B5, Canada
[7]These authors contributed equally to this work and are co-first authors
[8]These authors contributed equally to this work
*Correspondence: ceaves@bccrc.ca (C.J.E.), mhirst@bcgsc.ca (M.H.)
http://dx.doi.org/10.1016/j.stem.2013.12.011

## SUMMARY

**Cellular barcoding offers a powerful approach to characterize the growth and differentiation activity of large numbers of cotransplanted stem cells. Here, we describe a lentiviral genomic-barcoding and analysis strategy and its use to compare the clonal outputs of transplants of purified mouse and human basal mammary epithelial cells. We found that both sources of transplanted cells produced many bilineage mammary epithelial clones in primary recipients, although primary clones containing only one detectable mammary lineage were also common. Interestingly, regardless of the species of origin, many clones evident in secondary recipients were not detected in the primary hosts, and others that were changed from appearing luminal-restricted to appearing bilineage. This barcoding methodology has thus revealed conservation between mice and humans of a previously unknown diversity in the growth and differentiation activities of their basal mammary epithelial cells stimulated to grow in transplanted hosts.**

## INTRODUCTION

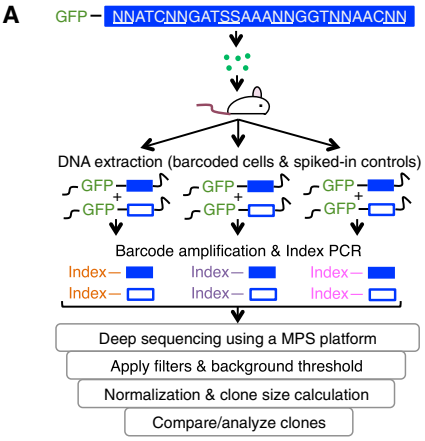
The output of specialized cells in many tissues involves a hierarchical differentiation process (Doulatov et al., 2012; Van Keymeulen and Blanpain, 2012) that originates in a rare tissue-specified stem cell population. Although many approaches have helped to delineate the properties of individual cells within these tissue hierarchies, all have limitations. In vitro systems are generally unable to be sustained for the long periods of mature cell output required to identify the most primitive elements, and in vivo approaches are constrained by the large numbers of animals needed to accrue data about the regenerative properties of single cells. The use of DNA-based genome markers facilitates clonal analyses, historically, most frequently applied to the study of hematopoiesis by detecting the semirandom genomic sites of integration of a viral vector in the progeny of initially transduced hematopoietic stem cells (Schmidt et al., 2009). The resolution of this approach has been increased substantially with the advent of linear amplification-mediated PCR (LAM-PCR) (Kustikova et al., 2008) coupled with massively parallel sequencing (MPS) technology (Arens et al., 2012). However, LAM-PCR-based methodologies are not well suited to deriving clone size estimates and do not sensitively detect small clones. The use of libraries of barcoded vectors can overcome many of these limitations provided appropriate analytical methods are used to circumvent library-specific molecular biases, accumulated errors inherent in MPS, and transduction of cells with more than a single integrated barcode sequence, as recently shown for tracking clones derived from barcoded mouse (Bystrykh et al., 2012; Gerrits et al., 2010; Lu et al., 2011; Naik et al., 2013; Schepers et al., 2008) and human hematopoietc cell transplants (Cheung et al., 2013).

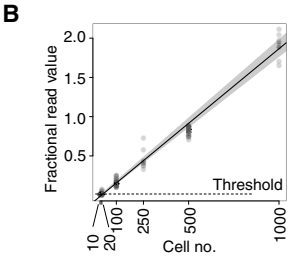
We now describe a multiplexed, high-resolution methodology suitable for deconvoluting barcode sequence data from expanded populations of cells originally transduced with a library of barcoded lentiviruses. We also show how application of this technology to serial transplants of basal mammary epithelial cells isolated from normal adult mouse and human glands can identify a conserved diversity of clonal growth and differentiation patterns and confirm results previously generated from analysis of limiting dilution transplants (Eirew et al., 2008; Lim et al., 2009; Shackleton et al., 2006; Stingl et al., 2006).

## RESULTS

### Generation of an Unbiased High-Throughput Multiplexed Cellular Barcoding Methodology

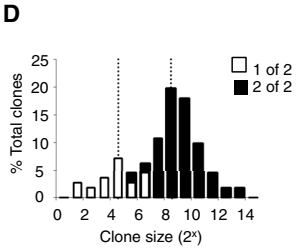
We constructed a library of MNDU3-PGK-GFP (Logan et al., 2004) lentiviral vectors containing a 27-nucleotide noncoding




C

| Controls (cell no.) | Sensitivity | | |
|---|---|---|---|
| | 1st MPS run | 2nd MPS run | 3rd MPS run |
| 1000 | - | - | 11/11 (100%) |
| 500 | 6/6 (100%) | 11/11 (100%) | 11/11 (100%) |
| 250 | - | - | 11/11 (100%) |
| 100 | 6/6 (100%) | 6/11 (55%) | 11/11 (100%) |
| 20 | 4/6 (67%) | 6/11 (55%) | 2/11 (18%) |
| 10 | - | - | 1/11 (9%) |



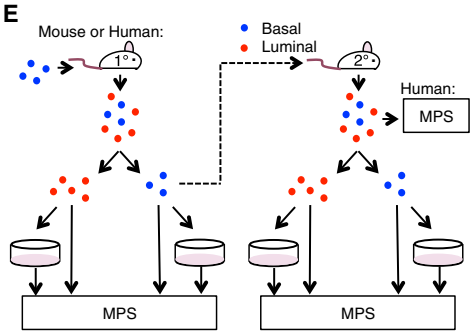




**Figure 1. Cellular Barcoding Using MPS for Quantitative Clonal Tracking Studies**

(A) Workflow of the barcoding methodology. Degenerate nucleotides for the variable regions in the barcode sequence are indicated with an N or S. At the end of the experiment, the barcode composition is determined using spiked-in control cell extracts added to each library. For details, see text and the Experimental Procedures.

(B) Relationship between absolute cell number and the fractional read value (the normalized barcode read abundance value) for the MPS data from the third MPS run. The solid black line indicates the linear regression fitted to the data with the 95% CIs for the relationship parameters shaded in gray. The threshold indicated is the fractional read representation equivalent to approximately 20 cells.

(C) Sensitivity measurements for the 28 sets of replicate control libraries analyzed.

(D) Reproducibility of clone detection with respect to clone size in seven replicate libraries (Table S5). Solid bars indicate clones detected in both libraries and open bars indicate clones detected in only one of the two. Dotted lines indicate the mode of the size distributions.

(E) Experimental design of the mammary cell transplant experiments.

barcode sequence designed as previously described (Gerrits et al., 2010) (Figure 1A). From >$10^6$ bacterial colonies, Sanger sequencing confirmed the accuracy and efficiency of the plasmid cloning (Table S1 available online) and MPS data indicated a diversity of 2 × $10^5$ unique barcodes (Figure S1A) and no systematic bias in read abundance. Analysis of PCR-amplified barcodes from genomic DNA extracted from primary human mammary cells transduced with the virus library generated from these plasmids showed these were contained within a 2-fold range of sequence reads (mean ± SD = 624 ± 157; Figure S1A).

## Establishment of Thresholds for Clone Identification and Absolute Size Determinations

To correct for variables inherent in genomic DNA extraction and/or construction of molecular libraries for sequencing, we added a set of control samples to each experimental sample to serve as an internal calibration. These "spiked-in" controls consisted of defined numbers of cells containing known barcodes and covered the expected range of experimental clone sizes (Table S2). To normalize barcode abundance values that vary in sequence coverage between libraries, each spiked-in control was converted to its fraction of sequences for all of the spiked-in controls in its respective indexed library (Table S3). For a total of 80 sets of spiked-in controls from three experimental data sets, we used regression analyses to establish the relationship between the fractional representation of each barcode and the number of cells from which the spiked-in sequences had been obtained (Figure 1B; Tables S2 and S3; Figure S1B). To establish a confidence threshold for the first two MPS runs, which had an SD of less than three cells across all groups (Table S4), we calculated the fractional representation of sequence reads in the spiked-in controls (the minimal fractional representation after normalization) using a single-cell threshold. In the third MPS run, an increased sensitivity and reduced specificity was observed (revealed by a larger SD, <44 cells, and wider 95% confidence interval [CI] across all groups; Table S4). Hence, a lower threshold corresponding to 20 cells was used in this case. The variation in regression data obtained from these different MPS runs highlights the importance of the spiked-in controls.

Application of these thresholds to virtual data sets from each of the three corresponding MPS runs (a total of 28 sets of control libraries constructed from the spiked-in control cells) showed that all clones of >500 cells were detected (100% sensitivity; Figure 1C), the 100-cell clones were detected with 55%–100% sensitivity, and the 20-cell clones were detected with 18%–67% sensitivity. The specificity of detecting "true" clones was shown to be >99% (Table S4), based on finding only 1 of the