CLONAL HETEROGENEITY OF NORMAL AND TRANSFORMED MAMMARY STEM CELLS

by

LONG VIET NGUYEN

B.Sc. (Honours), McGill University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Experimental Medicine)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2014

ABSTRACT

The normal mammary gland contains "stem cells" with extensive *in vivo* growth and bi-lineage differentiation potential and a surface phenotype of basal cells (BCs). BCs also contain cells with more limited growth and differentiation activity *in vitro*. An analogous luminal-restricted progenitor (LPs) subset has surface characteristics of both basal and luminal cells. I hypothesized that the growth and differentiation activity displayed by *individual* mammary epithelial cells from both subsets would be highly diverse, and that the properties of tumours produced from these cells would be affected by their cell of origin. To address this hypothesis, I first developed a lentiviral-mediated barcoding strategy that involves transducing each cell with a unique 27-base pair non-coding DNA sequence so that the number of its clonal progeny can be inferred from high-throughput sequencing data obtained on the progeny of bulk-transduced populations. The use of "spiked-in" control cells carrying a known barcode provided an internal calibration for clone size calculations and allowed clones of ≥100 cells to be reliably detected. Application of this strategy to normal mouse and human mammary cells identified expected bi-lineage clones but an unanticipated predominance of lineage-restricted clones produced in primary transplants. These experiments also revealed that many clones apparent in secondary hosts were not detected in the primary hosts, indicating their origin from cells with very delayed growth activity. Application of the barcoding strategy to normal human BCs and LPs transduced with lentiviruses encoding $KRAS^{G12D}$ ± $PI3KCA^{H1047R}$ ± $TP53^{R273C}$ showed tumour formation in subsequently transplanted immunodeficient mice was rapid (within 8 weeks) and efficient from both cell types (8-12/18 donors, 1/200-1/4,000 transduced cells). However, tumours generated from LPs

contained larger clones than tumours generated from BCs. Surprisingly, none of the LP-derived tumours were ERα^{+} (typical of luminal-like breast cancers) whereas 60% of the BC-derived tumours were. Earlier analysis of xenografts of similarly transduced cells revealed changes in both the number and phenotype of the cells present. Taken together, these findings underscore the diverse regenerative activity of normal mammary cells and provide definitive evidence that the cell of origin can affect the properties of human breast tumours generated using identical oncogenes.

PREFACE

Under the supervision and conceptual guidance of my supervisor, Dr. Connie Eaves, I designed and performed most of the experiments included in this thesis, as well as analysis and interpretation of the data obtained, with the exceptions detailed below. Dr. Martin Hirst provided important oversight of the development of the lentiviral-based cellular barcoding strategy discussed in Chapter 2, as well as library construction and barcode sequencing analysis included in Chapters 2, 3 and 4. Dr. Samuel Aparicio oversaw the characterization of the *de novo* generated tumours discussed in Chapter 4.

Dr. Nagarajan Kannan, Dr. Peter Eirew, and Dr. Maisam Makarem helped with designing experiments and interpreting data in Chapters 2, 3 and 4. Dr. Maisam Makarem performed the syngeneic mouse transplants and helped analyze the corresponding data for some of the experiments in Chapter 3. Dr. Alice Cheung contributed to the analysis of early barcoding experiments discussed in Chapter 2.

Michelle Moksa, Pawan Pandoh, Kane Tse, and Thomas Zeng contributed to the detailed design, construction and validation of the library of barcoded lentiviral vectors. Michelle Moksa, Dr. Melanie Kardel, Dr. Maisam Makarem, and Dr. Davide Pellacani helped with library construction for barcode sequencing. Annaick Carles designed custom scripts for analysis of the raw data from all barcode sequencing experiments. Dr. R. Keith Humphries and Patty Rosten provided instruction and technical assistance on the Southern blot analyses included in Chapter 2. William Kennedy helped with experiments included in Chapter 4. Dr. Tomo Osako reviewed all the tissue samples and immunohistochemistry of normal and pathological tissue included in Chapters 3 and 4.

Glenn Edin produced and titred the lentiviral supernatants, retrieved collagen gels from subrenal transplantation assays, and cultured cell lines used. Darcy Wilkinson consented, collected and prepared cryopreserved samples of dissociated cells from patient reduction mammoplasty samples, as well as performed immunohistochemical staining. Margaret Hale assisted with PCR measurements, molecular construction of oncogene-encoding lentiviral vectors, and Sanger sequencing experiments. Dr. Philip Beer designed and oversaw the construction of the oncogene-encoding lentiviral vectors.

I hope to incorporate some sections of Chapter 1 into a review on tissue stem cells. Sections of Chapter 1 pertaining to clonal diversification and the concept of cancer stem cells was adapted from a review published in Nature Reviews Cancer on which I was the lead author:

Nguyen, L.V., Vanner, R., Dirks, P., and Eaves, C.J. (2012). Cancer stem cells: an evolving concept. Nature reviews Cancer *12*, 133-143.

Work from Chapter 2 has been incorporated into two manuscripts, one published in Cell Stem Cell, and the second in Blood:

Nguyen, L.V., Makarem, M., Carles, A., Moksa, M., Kannan, N., Pandoh, P., Eirew, P., Osako, T., Kardel, M., Cheung, A.M., Kennedy, W., Tse, K., Zeng, T., Zhao, Y., Humphries, R.K., Aparicio, S., Eaves, C.J., and Hirst, M. (2014). Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. Cell stem cell *14*, 253-263.

Cheung, A.M., Nguyen, L.V., Carles, A., Beer, P., Miller, P.H., Knapp, D.J., Dhillon, K., Hirst, M., and Eaves, C.J. (2013). Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. Blood *122*, 3129-3137.

The Cell Stem Cell article also includes the work described in Chapters 3 and 4 pertaining to clonal tracking of normal mouse and human mammary basal cells.

The work in Chapter 4 pertaining to the characterization of human breast tumours generated *de novo* is being prepared for submission as a manuscript for publication in a peer-reviewed journal. The latter experiments in Chapter 4 pertaining to clonal tracking of primary patient breast tumour xenografts is also being prepared as a manuscript for submission to a peer-reviewed journal.

All animal experiments were carried out in accordance with the policies and guidelines presented by the Animal Care Committee of the University of British Columbia. Canadian Council on Animal Care Approval was granted under certificate number A11-0037. Human reduction mammoplasty tissue was obtained with informed consent according to the procedures approved by the Research Ethics Board of the University of British Columbia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

CHAPTER 2:

LIST OF ABBREVIATIONS

2D	**2-dimensional**
3D	**3-dimensional**
APC	**allophycocyanin**
BC	**basal cell**
bFGF	**basic fibroblast growth factor**
bp	**base pairs**
CD	**cluster of differentiation**
cDNA	**complementary deoxyribonucleic acid**
CFC	**colony-forming cell**
CFU	**colony forming unit**
CI	**confidence interval**
CK	**cytokeratin**
Cre	**Cre recombinase**
Cy	**cyanin**
DAPI	**4',6-diamidino-2-phenylindole**
DCIS	**ductal carcinoma in situ**
DMSO	**dimethylsulfoxide**
DNA	**deoxyribonucleic acid**
EDTA	**ethylenediaminetetraacetic acid**
EGF	**epidermal growth factor**
EGFR	**epidermal growth factor receptor**
EpCAM	**epithelial cell adhesion molecule**

ER	**estrogen receptor**
FACS	**fluorescence activated cell sorting**
FBS	**fetal bovine serum**
GFP	**green fluorescence protein**
H&E	**hematoxylin and eosin**
LAM-PCR	**linear-amplification mediated PCR**
LC	**luminal cell**
LCIS	**lobular carcinoma in situ**
LDA	**limiting dilution analysis**
LM-PCR	**ligation-mediated PCR**
LP	**luminal progenitor**
LTR	**long terminal repeat**
MOI	**multiplicity of infection**
MPG	**MNDU3-PGK-GFP**
MMTV	**mouse mammary tumour virus**
MPS	**massively parallel sequencing**
MRU	**mammary repopulating unit**
MUC1	**mucin-1**
ND	**not detected**
NSG	**nonobese diabetic severe combined immunodeficiency interleukin-2Rγc-null**
nrLAM-PCR	**nonrestricted LAM-PCR**
PAGE	**polyacrylamide gel electrophoresis**

PCR	**polymerase chain reaction**
PE	**phycoerythrin**
PerCP	**peridinin chlorophyll**
PR	**progesterone receptor**
RFLP	**restriction fragment length polymorphism**
ROS	**reactive oxygen species**
rtTA	**reverse tetracycline transactivator protein**
Sca-1	**stem cell antigen-1**
SD	**standard deviation**
SIN	**self-inactivating region**
SMA	**smooth muscle actin**
TERT	**telomerase reverse transcriptase**
TDLUs	**terminal ductal lobular units**
WAP	**whey acidic protein**
YFP	**yellow fluorescence protein**

ACKNOWLEDGEMENTS

I would like to thank my supervisor and mentor Dr. Connie Eaves for being an example of an exceptionally powerful, critical, and unrelenting scientific mind – a standard to which I will strive to meet for the rest of my career. She created such as supportive and intellectually nourishing home away from home while simultaneously teaching me what perseverence in research truly means.

I would also like to acknowledge my supervisory committee members Dr. Sandra Dunn, Dr. Martin Hirst, Dr. Aly Karsan, and Dr. Torsten Nielsen for their advice, criticism, scientific expertise, and help in guiding my thesis project.

I am truly indebted to my colleagues in the lab, particularly Dr. Maisam Makarem, Dr. Nagarajan Kannan, Dr. Peter Eirew, Dr. Davide Pellacani, and Sneha Balani, for insightful discussions and for their support through turbulent times.

The MD/PhD program led by Dr. Lynn Raymond and Dr. Torsten Nielsen has played a tremendous role in shaping my outlook on research and medicine. I profoundly thank them for creating a friendly, supportive and stimulating environment that has encouraged me to thrive academically. I also acknowledge Dr. Kalle Gehring and Dr. Guennadi Kozlov from McGill University for helping me gain my first firm footing into the world of research as an undergraduate student.

I also thank Jane Lee from the MD/PhD program, as well as Amanda Kotzer and Alice Chau from the Terry Fox Laboratory for their help in navigating through many administrative hurdles.

Lastly, I would like to thank the UBC Faculty of Graduate and Postdoctoral Studies, the Canadian Institutes for Health Research, and the Vanier Canada Graduate Scholarship program for their generous support.

This thesis is dedicated to my Mom and Dad

CHAPTER 1: INTRODUCTION

1.1 Structure of the normal mammary gland

The adult mammary gland is an intricate bilayered network of cells that form a hollow structure of inter-connecting ducts and lobules, ultimately all joining together at the nipple (Figure 1.1)(Visvader, 2009). The "terminal ductal lobular units" (TDLUs) are comprised of a terminal end duct with surrounding alveoli. In virgin women these TDLUs are smaller and less developed (~48 μm^2 in cross-sectional area, containing ~11 ductules/lobule) as compared to parous women (in whom they are ~60 to 129 μm^2 in cross-sectional area and contain ~47 to 81 ductules/lobule) (Russo and Russo, 2004).

The entire mammary gland is surrounded by a basement membrane consisting primarily of laminin and collagen IV (Novaro et al., 2003) and embedded in a collagen-rich stroma containing fibroblasts, adipocytes, blood and lymph vessels and hematopoietic cells. The connecting tissue between the stroma-embedded lobular units is generally less dense, and more adipose-rich. Cross-sectional histological analysis of the mammary ducts reveals that they are generally composed of two distinct cell layers with apicobasal polarity. The outer "basal" cell (BC) layer consists primarily of cells with myoepithelial features and contractile properties. The inner "luminal" cell (LC) layer includes the cells that can be stimulated to produce milk during pregnancy and lactation (Figure 1.2). Contractile forces exerted by myoepithelial cells within the basal layer causes milk to collect in the ducts and converge proximally to the lactiferous duct before being expelled through the nipple upon suckling from the child. Histological markers characteristic of the inner layer of LCs are CK8/18, CK19, MUC1, and EpCAM (Eirew

et al., 2008; Lim et al., 2009). BCs rest on the basement membrane, and express the following histological markers: SMA, CK5, CK14, EpCAM (low expression), CD10, Thy-1 (CD90), and alpha-6 integrin (CD49f) (Eirew et al., 2008; Lim et al., 2009).

The mouse mammary gland is also a bilayered branching structure, very similar in many respects to the human mammary gland. However, mouse breast tissue differs from its human counterpart in several important respects. Mouse mammary ducts do not terminate in TDLUs, but rather in "terminal end buds", from which further ducts and branches can be stimulated to form under various conditions, including pregnancy (Visvader, 2009). Also the fibrous stroma surrounding mouse mammary ducts is a much thinner layer than in the human breast, and contains a much higher concentration of adipocytes. Markers characteristic of mouse LCs are: CK8, CK18, CK19, EpCAM (high expression), CD24 (high expression), CD29 (mid to low expression), and CD49f (mid to low expression) and of mouse BCs are: SMA, CK5, CK14, EpCAM (mid to low expression), CD24 (mid to low expression), CD29 (high expression), and CD49f (high expression) (Shackleton et al., 2006; Stingl et al., 2006a).

The inner luminal layer of cells and the entire outer layer of BCs of the mammary gland of both species are thought to represent two distinct cell lineages. According to this model, the luminal lineage would include both ductal and alveolar luminal epithelial cells, and the myoepithelial lineage would characterize the differentiated cells within the basal layer of the mammary gland. Only LCs are able to produce milk upon secretory differentiation whereas myoepithelial cells produce and become anchored to the basement membrane, and possess contractile properties (Stingl et al., 2006b). The growth and differentiation of all of these cells can be regulated by external cues, which trigger

signaling pathways and gene expression changes that are modulated internally by transcription factors and epigenetic regulators to determine the molecular state of the cell, and thus, its specific functions (Visvader, 2009).

1.2 Development of the normal mammary gland

Most of our understanding of the development of the mammary gland has come from studies of its formation in the mouse. There, the first evidence of cells destined to generate mammary epithelium is seen between the 10^{th} and 11^{th} days of gestation (E10.5-E11), when 5 pairs of mammary "placodes", which consist of pseudostratified epithelium, become apparent in the ectoderm at sites where future mammary glands will be found (Balinsky, 1950; Howard, 2012). By E12, the placodes transform into hemispheres connected to the epidermis by a narrow neck, structures referred to as "mammary hillocks". Preadipocytes begin to appear in the underlying mesenchyme at E13.5-E14.5, and these later generate the fat pad in which the mammary gland forms. At E15.5 the mammary buds invaginate into the underlying mesenchyme, and rudimentary structures begin to elongate and produce primary and secondary sprouts eventually forming a rudimentary branched structure. Around E17.5 lumen formation begins within the ductal branches of the developing gland, and this process continues until birth (Balinsky, 1950). In male mice, the same process occurs initially up until E14.5. Then the production of androgen causes the mesenchyme to sever the connection between the nascent mammary cells and the overlying epidermal cells resulting in apoptosis of the underlying mammary cells. Interestingly, this step does not occur in the human male

embryo and thus human males are born with a small mammary ductal structure around the nipple (Howard, 2012).

In humans, the prenatal mammary gland is derived from a single epithelial ectodermal bud from a mammary line that spans between the limb buds. Mammary lines are observed in rabbit, rat and human embryos, but not in the mouse. By the time of birth, the mammary ducts in humans have already formed short distal branches called ductules that are also 1-2 epithelial cell layers in depth (Russo and Russo, 2004). During gestation, the maternal hormones induce features in the developing mammary epithelial cells that are typical of an apocrine secretory epithelium, including the formation of lipid-filled cytoplasmic vacuoles. However, by 3-4 weeks after birth, these secretory features have subsided (Russo, 1987; Russo and Russo, 2004).

In both mice and humans, the next significant series of changes in the growth of the mammary gland occur during puberty under the influence of estrogen and progesterone hormones. At this time, the cells in the rudimentary mammary gland begin to proliferate more actively and produce more glandular tissue in concert with parallel growth of the animal including the stroma surrounding the mammary gland itself. The mammary ducts grow in size and number and form more developed, club-shaped terminal end buds (Russo and Russo, 2004). In mice, the ducts extend to the outer limits of the mammary fat pad in the 10-12 week old adult, and secretory differentiation is not observed until pregnancy (Howard, 2012).

During puberty in humans, the terminal end buds continue to generate new branches or “alveolar buds” that cluster around the terminal end ducts, which together constitute the TDLU. TDLUs develop radially from the nipple to comprise approximately

5-10 units by adulthood. Later in life, with the occurrence of peri-menopause associated with ovarian follicular atresia and resulting irregular menstrual cycles, the mammary glandular epithelium begins to regress. In post-menopausal women, the mammary glandular epithelium undergoes a significant regression with a concomitant decline in the number of the larger and more developed TDLUs (such that smaller TDLUs commonly found in virgin adult women are predominant in the remaining mammary gland) (Russo and Russo, 2004).

1.3 Mouse mammary stem and progenitor cells

In 1959, DeOme et al developed a method to transplant fragments of mouse mammary tissue into the cleared inguinal fat pads of virgin recipient mice (Deome et al., 1959). Virtually all portions of the mammary gland from the adult donor mice demonstrated the ability to recapitulate an entire functional mammary structure upon transplantation (Figure 1.3), suggesting that the mammary glandular epithelium of adult mice retains the capacity for regeneration, for several passages until regenerative senescence eventually ensues (Daniel et al., 1968; Daniel and Young, 1971).

Almost four decades later Kordon and Smith used mouse mammary tumor virus (MMTV), known to integrate semi-randomly into the genome of transduced cells, to uniquely mark cells transplanted into the cleared fat pads of recipient mice. The clonal repertoire of the regenerated mammary epithelium was discerned by enzymatic restriction followed by the discrimination of unique clones visually as different fragment sizes on a Southern blot. Their findings suggested that a single cell could regenerate an entire functional mammary gland (Kordon and Smith, 1998).

Subsequent studies utilized a limiting dilution approach to quantify the frequency of mouse mammary cells with *in vivo* regenerative activity when transplanted into the cleared fat pad of 3-week-old virgin female mice (Shackleton et al., 2006; Stingl et al., 2006b). This led to definition of mammary repopulating units (MRUs) based on the detection of a fully regenerated branching mammary gland structure produced by the test cells 6-8 weeks after they had been transplanted. By performing transplants at various cell doses, a cell dose at which ~37% of transplants are negative and, assuming the data fit a Poisson distribution, this allows the number of input cells that contain a single MRU to be identified (Table 1.1). However, a single-hit Poisson model to calculate the frequency of MRU may not appear to apply when transplant datasets are limited by the number of cell doses and technical replicates feasible with this approach (Bonnefoix and Callanan, 2009). In this case, an extreme limiting dilution analysis model may serve as a better alternative (Hu and Smyth, 2009).

Advances in purifying subsets of dissociated mammary epithelial cells allowed cell suspensions that were enriched in MRU to be obtained. CD45, CD31 and Ter119 were used to deplete for hematopoietic, endothelial and stromal cells, respectively. Following this, various combinations of markers such as EpCAM, Sca-1, CD49f, CD29 and CD24 were tested for their presence or absence on MRUs by selective fluorescent activated cell sorting (FACS) of viable positive and negative populations and subjecting these to the *in vivo* transplant "MRU assay". The highest reported purity of MRUs obtained in these initial experiments was approximately 1 in 20 to 1 in 50 $CD45^{-}CD31^{-}Ter119^{-}Sca\text{-}1^{lo}CD49f^{hi}$ cells (Stingl et al., 2006a), which had a smaller 95% CI compared to other studies (Welm et al., 2002). This made single cell transplants feasible and such

experiments confirmed that an entire functional mammary gland could indeed be regenerated from a single transplanted cell (Shackleton et al., 2006; Stingl et al., 2006a).

Although transplant studies suggest that MRU activity is an exclusive property of a subset of basal mammary epithelial cells, other models used to track the clonal expansion potential of mammary epithelial cells have revealed that LCs can also produce cells of the luminal lineage for extensive periods of time *in vivo*. This was shown first by lineage-tracing studies in transgenic mice carrying an inducible whey acidic protein (WAP)-specific promoter to drive expression of Cre to activate expression of LacZ. Tracking the progeny of these $LacZ^{+}$ cells in the mammary gland over several cycles of pregnancy, lactation, and involution revealed a "parity-induced" luminal progenitor (LP) cell that proliferated during pregnancy to produce mature LCs in both the ducts and alveolae. After involution, most of the labeled cells were found to undergo apoptosis, although a few remained and could subsequently repeat this regenerative process in a second and third cycle of pregnancy and lactation (Wagner et al., 2002). Later, similar cells were shown to exist in nulliparous mice as well (Booth et al., 2007). Consistent with these early findings, more recent lineage-tracing studies using CK8 and CK18-specific inducible promoters to mark LCs with YFP suggested that LCs in pubertal and adult virgin female mice clonally expanded over a period of 10 weeks, and could do so through three cycles of pregnancy, lactation and involution (Van Keymeulen et al., 2011). Interestingly, E12.5 Axin2-CreERT2/Rosa26-LacZ mouse embryos marked cells with restricted luminal differentiation throughout postnatal development, suggesting that cells of the luminal lineage may, under normal physiological conditions, be contributed by cells of a LC-specified fate as early as E12.5 in the developing mammary placode (Table

1.2) (van Amerongen et al., 2012). This claim, however, has not yet been reinforced by any other studies in the field. Later, it was shown that a rare subset of ER^+ and ER^- LP cells ($CD45^-CD31^-Ter119^-EpCAM^{++}CD49f^+Sca1^+CD49b^+$, and $CD45^-CD31^-Ter119^-EpCAM^{++}CD49f^+Sca1^-CD49b^+$ cells, respectively) could be activated to repopulate cleared mammary fat pads after an initial engraftment in renal transplants (collagen/matrigel gels transplanted under the renal capsule) (Shehata et al., 2012). Similarly, mouse MRU activity can be activated within the progeny of occasional (approximately 6%) purified LCs from adult virgin mice when these cells are cultured at clonal frequencies for 7 days in a 3D Matrigel system (Makarem et al., 2013).

Although transplantation of BCs has established that MRUs have the potential for bi-lineage differentiation, lineage-tracing studies reveal that under normal physiological conditions there are $CK14^+$, $CK5^+$, $Lgr5^+$, or $Axin2^+$ BCs that if labeled in prepubescent mice behave as if restricted to the myoepithelial lineage (van Amerongen et al., 2012; Van Keymeulen et al., 2011). On the other hand, when these cells are labeled on postnatal day 14 to 16 or in adult mice, they subsequently generate both BCs and LCs (van Amerongen et al., 2012). One study reports that of the clones generated from CK5-labeled cells, bi-lineage clones are predominant. Interestingly, of these bi-lineage clones, some contained an equal proportion of BCs and LCs whereas others were enriched for LCs (Table 1.2) (Rios et al., 2014).

Mammary epithelial cells thus demonstrate different potentials for growth and differentiation under different conditions, which appear to be highly dependent on age, parity, surrounding environment, and whether the cells are assayed by transplantation or *in situ* labeling (Joshi and Khokha, 2012; Visvader and Lindeman, 2011).

Mouse mammary epithelial cells with colony-forming activity in 2D assays *in vitro* have also been identified. Early studies showed that ~7% of bulk mouse mammary cells had this activity if assayed under conditions of low oxygen (Smalley et al., 1998). This frequency of mammary colony-forming cells (CFCs) was mouse strain-dependent (Table 1.3), and reached ~30% in later studies (Shackleton et al., 2006). Our group has shown that the low (5%) O_2 requirement is exclusive to basal CFCs and that their growth is further optimized by the addition of Rock inhibitor (Y-27632) to the culture medium (Makarem et al., 2013). Prospective purification of LC and BC subsets ($EpCAM^{++}CD49f^{+}$ and $EpCAM^{+}CD49f^{++}$, respectively) allowed for CFC frequencies to be measured within these subsets, as mouse colonies derived from luminal and basal progenitors are not distinguishable on the basis of their morphology or the immunophenotypic features of their progeny (Shackleton et al., 2006; Stingl et al., 2006a).

In liquid suspension cultures maintained in a system that discourages cell adherence, mammary cells will produce "mammospheres". These can arise from single cells in which case each mammosphere is clearly a clone. However, because of the high motility of mammary epithelial cells and their strong tendency to adhere to one another, mammospeheres generated from larger numbers of cells are typically aggregates of variably expanded clones (Booth et al., 2007). Thus, the report that mammospheres from $AXIN2^{+}$ WNT-responsive cells can be serially propagated *in vitro* and retain cells with MRU activity is difficult to interpret (Zeng and Nusse, 2010).

In 3D matrigel culture, mouse mammary cells form elaborate, branched alveolar structures from both individual BCs and LCs (Dontu et al., 2003; Makarem et al., 2013).

When irradiated NIH-3T3 cells are added to the cultures, many more daughter CFCs are produced (~10,000 and ~100-fold increase from BCs and LCs of adult mice, respectively) (Makarem et al., 2013). The BCs also produce expanded numbers of MRU at high frequency in these cultures and, as noted above, even some LCs are activated to display MRU activity.

Historically, the organization of the mammary gland was conceptualized as a hierarchy wherein self-renewing MRU (enriched within the basal fraction) produce myoepithelial and luminal progenitors, which in turn give rise to mature terminally differentiated cells within their respective lineages (Figure 1.4)(Asselin-Labat et al., 2008; Stingl et al., 2006b). However, recent *in situ* lineage tracing studies suggest that although MRU indeed have bi-lineage regenerative potential, under normal physiological conditions, cells within the myoepithelial and luminal lineages can be self-sustaining, from cells which retain the potential for long-term regeneration (Fu et al., 2014). Furthermore, although differentiation was originally conceived as a unidirectional and irreversible process, with the demonstration of induced pluripotency of adult somatic cells (Takahashi and Yamanaka, 2006), and activation of LCs to demonstrate MRU activity (Makarem et al., 2013; Shehata et al., 2012), the original unidirectional hierarchical concept of mammary cell differentiation may be viewed as overly rigid.

1.4 Human mammary stem and progenitor cells

It has long been known that the adult human mammary gland is of polyclonal origin. Studies that mapped patterns of X-chromosome inactivation, a process occurring early in prenatal development known as lyonization, revealed that the gland contains contiguous patches of tissue of different clonal origin (Tsai et al., 1996). This suggests that single

bipotent primitive cell types give rise to discrete regions of the gland, and further that cells which remain in these regions throughout life, contribute to normal tissue turnover.

With the development of methods to viably dissociate mammary epithelial cells into a single cell suspension from donor reduction mammoplasty samples, it became possible to assay for cells with colony forming activity in 2D serum-free cultures containing epidermal growth factor (EGF). Initial frequencies of the CFCs detected were low (Table 1.4) suggesting that CFCs are rare, and thus might represent intermediate progenitor cells that retain the ability to produce mature and terminally differentiated cells (Emerman et al., 1996; Stingl et al., 1998; Stingl et al., 2001; Stingl et al., 2005).

Subsequent addition of irradiated fibroblasts to the 2D assays and the use of collagen-coated tissue culture plates allowed the frequency of CFCs to be markedly increased. Under these conditions, three morphologically and molecularly distinct colonies could be recognized after 10-14 days (Stingl et al., 1998; Stingl et al., 2001). These consisted of : (i) “luminal” colonies apparent as compact, round colonies with a smooth boundary, and containing predominantly cells that stain positive for markers characteristic of cells within the luminal layer *in vivo* (i.e., EpCAM, CK8/18, CK19, and MUC1), with a few cells staining for CK14, a marker characteristic of BCs; (ii) “myoepithelial” colonies consisting of more dispersed teardrop-shaped cells that stain uniformly for CK14 and do not express luminal markers; and (iii) “bi-lineage” colonies that contain a mixture of both luminal and myoepithelial cells as just defined (Stingl et al., 2005). By prospective purification of LC and BC subsets (Table 1.4), luminal colonies could be shown to be derived almost exclusively from a subset of LCs and hence separately from the myoepithelial and bi-potent CFCs, which were present almost

exclusively in the BC fraction (Kannan et al., 2013; Raouf et al., 2008). Although it was not possible to separate the latter two CFCs types phenotypically, myoepithelial restricted CFCs were found to be selectively increased with serial *in vitro* passage.

Human mammary epithelial cells, like their murine counterparts, when cultured in suitable media in ultra-low adherent tissue culture plates also form "mammospheres". Critical additives are EGF (and/or basic fibroblast growth factor, bFGF), insulin, hydrocortisone and B27 (Dontu et al., 2004). Characterization of mammospheres by immunohistochemistry reveal that they are negative for MUC1, SMA, CK18, but are positive for CD49f, CK5, and CD10, with a random distribution of staining for EpCAM (~50% of cells) and CK14 (~30% of cells). Mammospheres can be stimulated to differentiate if plated on collagen-coated tissue culture plates, and covered with a layer of prolactin-supplemented matrigel. After 7 days, three types of colonies form: luminal ($EpCAM^+$), myoepithelial ($CD10^+$) and bilineage (positive for both markers). Several other 3D culture conditions which use various extracellular matrix proteins have been developed to look for differentiation and branching. Specifically, such studies have shown that branching is an exclusive property of the LP subset of cells ($EpCAM^+CD49f^+$) that reside in putative stem cell zones of ducts but not lobules, whereas mature LCs and BCs produce small and large spherical colonies, respectively (Villadsen et al., 2007).

The development of *in vivo* assays has allowed very primitive human mammary cells to be detected with greater specificity. One assay adapted a heterotopic model originally developed for xenografting normal and cancer tissue (Bogden et al., 1979; Buck, 1963; Lee et al., 2005). It involves embedding dissociated mammary epithelial

cells in a solid collagen gel that is then implanted under the renal capsule of immunodeficient mice for 4-8 weeks. At the end of that time, histologically normal bi-layered structures can be observed to have formed (Figure 1.5). Moreover, these structures contained mammary CFC whose presence is demonstrable when the cells in the removed gel are dissociated enzymatically and assayed in standard 2D cultures. The colonies they generated are similar to those observed from cells dissociated from primary reduction mammoplasty samples (Eirew et al., 2010; Eirew et al., 2008). Limiting dilution analysis revealed that the number of CFCs present in the gels after 4 weeks correlated linearly with the number of cells originally transplanted, suggesting that a single entity was responsible for their production and hence rationalized the use of limiting dilution methods to quantify its frequency. By analogy with the mouse studies, the human mammary cell responsible for generating detectable CFCs in this assay was termed a human MRU. Limiting dilution experiments indicate that the frequency of MRUs in bulk dissociated reduction mammoplasty samples is 1/1,000 to 1/10,000 (Table 1.5), and the average number of CFC obtained per MRU is 4.1 ± 0.6 CFC (Eirew et al., 2008). Cell purification experiments further showed that MRU activity is exclusively detected in the BC subset ($EpCAM^{lo}CD49f^{+}$) with minimal to no MRU detected in either the mature LC or LP cell subsets ($EpCAM^{+}CD49f^{-}$ and $EpCAM^{+}CD49f^{+}$, respectively, Table 1.5) (Eirew et al., 2008).

A second *in vivo* assay allows for orthotopic xeno-transplantation of human mammary epithelial cells (Sheffield and Welsch, 1988) by first clearing, then humanizing the mouse mammary fat pad with human mammary fibroblasts (Kuperwasser et al., 2004). Although this method yielded similar histologically normal bi-layered structures

after 8 weeks, it is more cumbersome due to the multiple surgeries required (Kuperwasser et al., 2004; Proia and Kuperwasser, 2006). This procedure was later refined to involve only a single surgery in a study that found the frequency of MRU detected is also lower than for the kidney capsule MRU assay (Table 1.5), although the phenotype of the cells responsible is the same (Lim et al., 2009). Interestingly, a later study showed that mammary cells expressing high levels of CD44, whether initially isolated from human tissue as such, or after culture under non-adherent conditions, could reconstitute the humanized mouse mammary fat pad (Chaffer et al., 2011), but how these cells compared to those previously defined to enrich for MRU activity in either the heterotopic or orthotopic xenograft models, has yet to be shown.

Similar to the mouse, organization of the human mammary gland has historically been conceptualized as a hierarchy in which self-renewing MRUs (with a basal phenotype) produce bi-potent as well as myoepithelial and luminal-restricted progenitors that in turn give rise to mature terminally differentiated cells within their respective lineages (Figure 1.4). Although *in situ* lineage tracing studies are not feasible in humans, a recent study showed that human mammary LCs can produce multilayered acinar structures *in vivo* that contain cells of both the luminal and myoepithelial lineages, suggesting a degree of developmental plasticity not previously known (Shehata et al., 2012). Thus, differentiation within the normal human mammary gland may also not necessarily be as rigid and irreversible as initially thought.

1.5 Regulation of the biological properties of mouse and human mammary stem and progenitor cells

The mammary gland undergoes continuous dynamic remodeling throughout life, including periods of significant expansion in both the myoepithelial and luminal compartments, as seen in a pronounced fashion during puberty. Pregnancy and lactation, however, initiate a particular expansion of LCs within the alveolae leading to milk production. Accumulating data are identifying the nature and role of specific extrinsic cues that regulate these changes (Visvader, 2009).

Estrogen and progesterone are both steroid hormones produced primarily in the ovaries which regulate mammary cell growth and differentiation throughout development, particularly during puberty, pregnancy and lactation. Estrogen, particularly the active compound 17β-estradiol (estradiol), exerts its effects through activation of the estrogen receptor (ER) α and β, members of a superfamily of nuclear receptors that act as ligand-inducible transcription factors. ERα is required for ductal elongation and invasion into the mammary fat pad in mice during puberty (Brisken and O'Malley, 2010; Mallepell et al., 2006; Tanos et al., 2012), as *Esr1*$^{(-/-)}$ mice (that lack expression of ERα) develop only rudimentary ducts. Moreover, this phenotype persists even when the mammary cells from *Esr1*$^{(-/-)}$ mice are transplanted into the cleared fat pads of *Esr1*$^{(+/+)}$ mice indicating that it is a cell-autonomous requirement (Mueller et al., 2002). A similar phenotype is observed in ovariectomized mice or mice with a defect in aromatase cytochrome P450, an enzyme involved in the biosynthesis of estradiol (Bocchinfuso and Korach, 1997; Fisher et al., 1998; Imagawa et al., 1990). However, this phenotype, as might be predicted, can be rescued at least partially by exogenous administration of estrogen (Daniel et al., 1987).

Esr2$^{(-/-)}$ mice (that lack expression of ERβ) in contrast appear to develop normally functioning mammary glands (Krege et al., 1998), suggesting it is ERα that is primarily involved in regulating mammary gland development.

Progesterone acts on the progesterone receptor (PR), also a nuclear receptor that is important for regulating ductal branching and alveolar differentiation during pregnancy and lactation (Brisken et al., 1998). Progesterone induces a ~10-fold increase in MRUs during the dioestrus phase and during pregnancy in mice (Joshi et al., 2010). Exposure to exogenously administered progesterone induces a similar expansion in MRU numbers. However, MRUs have an ER$^-$PR$^-$ phenotype (Asselin-Labat et al., 2008), suggesting that progesterone must exert its effect on MRUs indirectly via a paracrine mechanism.

The levels of RANK receptor and its ligand, RANKL protein, have both been shown to increase during pregnancy in mouse mammary epithelial cells. RANKL increases in mature ductal LCs both during pregnancy or when progesterone and estrogen are administered, suggesting these hormones are the mediators (Lee et al., 2013). Basal cells express the RANK receptor, suggesting that the ability of progesterone to stimulate an increase in MRUs involves the activation of RANKL production by mature LCs which then stimulates RANK$^+$ MRUs (Lee et al., 2013). Consistent with the role of RANK and RANKL in mammary gland development, *Rank*$^{(-/-)}$ and *Rankl*$^{(-/-)}$ mice have been found to exhibit defective alveolar differentiation during pregnancy and lactation, which can be rescued by forced expression of RANK or RANKL (Fata et al., 2000). RANKL activation of RANK initiates multiple signaling pathways, particularly the AKT and ERK pathways, which are implicated in mammary epithelial cell proliferation (Beristain et al., 2012).

A similar paracrine mechanism exists with WNT signaling, where WNT4 expression is induced in mouse LCs during pregnancy or when mice are supplemented with progesterone plus estrogen, to act on the LRP5 Wnt receptor. LRP5 is expressed on BCs, and once activated can transduce downstream signals leading to upregulation of WNT target genes *Axin2* and *Mmp7*, and *Axin2* and *Tcf1* in LCs and BCs, respectively (Joshi et al., 2010).

E-twenty six transcription factor (ELF5) is thought to be important for alveolar differentiation, since decreased methylation of the *ELF5* promotor in LCs coincides with the increased expression of *ELF5* and subsequent differentiation to $ELF5^{+}$ mature alveolar cells, induced by progesterone during pregnancy (Oakes et al., 2008; Zhou et al., 2005). Also, loss of ELF5 blocks alveolar LC differentiation resulting in an increase in the number of LP cells. In contrast, in BCs, ELF5 is methylated, suggesting it is expressed in a lineage-specific fashion to direct luminal alveolar cell fate (Lee et al., 2011). GATA3 is another transcription factor that appears to function in a similar capacity as ELF5, to promote luminal alveolar cell differentiation, since a loss of GATA3 results in a block in differentiation concomitant with an expansion of LP cells, and its forced expression in BCs increasingly directs MRU toward luminal differentiation (Asselin-Labat et al., 2007).

NOTCH signaling has also been implicated in regulating the commitment of bi-potent mammary cells to the luminal lineage, as well as restricting MRU expansion. Knockdown of CBF1, the canonical NOTCH effector molecule, was reported to increase mouse MRU numbers *in vivo* (Bouras et al., 2008). Similarly, in non-adherent suspension cultures, NOTCH-activating DSL peptide increased the frequency of secondary human

mammosphere-initiating cells and promoted branching in 3D matrigel cultures, whereas these effects were lost by inhibiting NOTCH signaling with a NOTCH4 blocking antibody or a γ-secretase inhibitor (Dontu et al., 2004). At the point of commitment to the luminal lineage, disruption of NOTCH3 signaling by knockdown of NOTCH3 receptor, treatment with a γ-secretase inhibitor, or forced expression of a dominant negative form of human mastermind-like-1 (MAML) in purified bi-potent CFCs, which is critical for the transcriptional activation of Notch signaling, suppressed their ability to produce luminal CFCs but had no effect on the ability of luminal-restricted CFCs to complete their differentiation program (Raouf et al., 2008).

Mammalian Pygopus 2 (PYGO2), a member of the Pygopus family of transcriptional coactivators of the WNT/β-catenin signaling pathway (Jessen et al., 2008), binds to H3K4me3 histone marks (that are associated with active transcription) to recruit histone-modifying enzymes that will produce further active histone marks and lead to upregulation of gene expression (Gu et al., 2009; Gu et al., 2012). Using a *Pygo2*$^{(-/-)}$ mouse, it was shown that PYGO2 mediates cross-talk between NOTCH and WNT signaling pathways to suppress luminal differentiation of MRU-enriched BCs.

The polycomb-group (PcG) proteins also regulate epigenetic changes of chromatin, and include proteins such as BMI1 and EZH2. PcG proteins are widely implicated in maintaining stem cell identity in many tissues (Aloia et al., 2013; Luis et al., 2012). In mice, BMI1-deficiency impairs mammary gland development and reduces the number of MRU (Pietersen et al., 2008). Conversely forced overexpression of BMI1 in human mammary cells increased mammosphere formation (Liu et al., 2006). EZH2, a H3K27 methyltransferase, also appears to play a key role in maintaining mouse

mammary stem cells, since *MMTV-Cre/Ezh2*$^{f/f}$ mice show a 14-fold decrease in MRU frequency and impaired luminal alveolar cell differentiation. EZH2 expression in mammary epithelial cells is induced (at least partially) by progesterone, leading to consequent H3K27me3 modifications that are thought to regulate stem and progenitor activity, and differentiation (Pal et al., 2013).

As noted, EGF is necessary for the propagation of both mouse and human mammary epithelial cells *in vitro*, especially when the cultures are initiated at low density and paracrine and autocrine effects are thereby reduced. Interestingly, ERBB2, also referred to as human epidermal growth factor receptor (HER-2/neu) is a receptor for the family of EGF-related ligands and is not expressed on normal mammary epithelial cells (Asselin-Labat et al., 2006). However, it is overexpressed in 20-30% of human breast tumours (Press et al., 2005). Activation of the EGF receptor (EGFR) by EGF stimulates the PI3K/AKT signaling pathway, one of the most commonly deregulated pathways in human breast tumours (~37% prevalence of PI3K mutations) (Cancer Genome Atlas, 2012). Activation of EGFR also stimulates the RAS/ERK pathway, with both pathways promoting cell survival. Although RAS mutations are not as prevalent in human breast cancers, several serine/threonine kinases such as MAP3K1 and MAP2K4 are candidate tumour suppressor genes, and have a high prevalence of mutation in human breast tumours (Cancer Genome Atlas, 2012). Wild-type MAP3K1 regulates JNK activation and its E3 ubiquitin ligase domain ubiquitylates c-Jun and ERK1/2 to simultaneously induce pro-survival signals and promote apoptosis (Pham et al., 2013). MAP2K4 (also known as MKK4) is known to activate JNK1 and p38 signaling pathways that regulate cell proliferation and apoptosis (Teng et al., 1997). Thus loss of function of these kinases

enhances pro-survival signals. In combination with other perturbations, their disruption is associated with the acquisition of malignant properties.

1.6 Approaches for clonally tracking cells with regenerative activity *in vivo*

1.6.1 Direct single cell assays

Methods for tracking the clonal growth behaviour of individual cells require a strategy to obtain or mark the starting population so that the progeny of every clone can be distinguished. The most direct approach is to transplant a single cell and then monitor its progeny over time or after a defined period of time. This approach is, however, only practical if the behavior of interest can be elicited from a sufficient proportion of the test population to make the negative transplants an economically and practically acceptable outcome. This has been achieved in a few instances (Benz et al., 2012; Shackleton et al., 2006; Stingl et al., 2006a; Yamamoto et al., 2013) but is generally not feasible, necessitating alternative approaches.

1.6.2 Limiting dilution approaches

One method to examine the outputs from clonally expanding cells is to use a limiting dilution approach. Generally, when several doses of cells are assayed for a property that can be detected as a positive or negative readout, limiting dilution analysis provides an estimate of the cell dose that is at limit. Based on a Poisson distribution, the limiting dose is equivalent to the cell dose at which 37% of the replicates are detected as negative (Figure 1.6)(Eirew et al., 2010). However, there is typically a wide 95% confidence interval associated with this calculation, and at limit there is a reasonable chance of the

cell dose containing >1 cell of interest. Thus, to be confident the observed readout is from a single clonally expanded cell, these assays need to be performed at sub-limiting doses. These, like single-cell transplants, can still be too cumbersome for many types of analyses.

1.6.3 Tracking of endogenous markers

Another approach to tracking the clonal outputs of cells is to identify a unique mark that can be detected and discriminated from other clonally expanded cells. One of the earliest methods involved tracking chromosomal rearrangements that could be identified in karyotype preparations from cells. This was used to infer the clonal origin of various leukemias, including chronic myeloid leukemia in which the translocation between chromosomes 9 and 22 resulted in a minute "Philadelphia" chromosome (Nowell and Hungerford, 1960). Similar to this, "neutral" chromosomal rearrangements induced by exposure to sublethal doses of ionizing radiation have been used to identify primitive hematopoietic cells with multi-lineage clonogenic regenerative potential *in vivo* (Barnes et al., 1959).

Another strategy that has been useful for examining the clonality of TDLUs and cancerous lesions in the breast, and other diseases such as hyperplasias in the lung (Niho et al., 1999), and Langerhans'-cell histiocytosis (Willman et al., 1994), is restriction fragment length polymorphism (RFLP) analysis. RFLP analysis is based on detecting X-chromosome-linked polymorphic markers. These markers can only be identified in women, as they have two X-chromosomes, one of which is inactivated due to methylation of one of the two chromosomes, by a process called lyonization (van Dijk et

al., 2002; Vogelstein et al., 1985; Vogelstein et al., 1987). Lyonization occurs randomly within cells early in the development of the female embryo, and the pattern set at that time is then permanently propagated in all daughter cells (Riggs, 1975). Thus most tissues are generated from a random mixture of many of these fetal cells and the state of activation of the maternally and paternally-derived X-chromosome allows the clonal composition of adult tissues to be assessed. This strategy depends on the assessment of gene alleles that are polymorphic either at the protein (e.g., G6PD (Fialkow et al., 1967)) or DNA level (Vogelstein et al., 1985). The latter can be detected by DNA probes that recognize the polymorphism, after the DNA has been digested by a methylation-sensitive restriction endonuclease that does not digest the allele on the inactivated chromosome. This results in two DNA fragments with different lengths that can be distinguished by gel electrophoresis. There are, however, several limitations to this method. Phosphoglycerate kinase (PGK), hypoxanthine phosphoribosyltransferase (HPRT) and the hypervariable locus DXS255 (M27beta) are some markers that have been used in the past, but the low incidence of polymorphisms in these genes in the general female population can pose severe constraints on the sample numbers. The human androgen receptor gene (HUMARA) is a better alternative, since it contains a highly polymorphic trinucleotide repeat that is polymorphic in ~90% of the female population (Vogelstein et al., 1987). Another limitation of RFLP analysis is the efficiency of the restriction endonuclease. If the digest is incomplete, the assessment of clonality by gel electrophoresis is compromised, and although quantitative PCR-based methods have been employed to monitor the completeness of enzymatic restriction, such approaches have not garnered much success at circumventing this limitation (van Dijk et al., 2002). Finally, such

methods are weighted toward detecting single predominant clones, but cannot resolve contributions of multiple clones expanding in the same location.

1.6.4 Tracking of exogenously-introduced markers

The introduction of unique marks into the genome of cells has been a powerful development to circumvent the limitations associated with RFLP analysis. These unique marks have been traditionally integrated into the genome of cells using retroviruses or lentiviruses, which are known to permanently integrate into the genome in a semi-random fashion. This viral transduction step allows for their site of integration(s) to be used as uniquely discriminating clonal marks, assuming that the site of integration itself does not alter the growth properties of the cells. While this was originally considered to be a very rare event, some studies indicate that it occurs more often than initially anticipated, sufficient to represent a significant clinical problem in gene therapy trials (Schroder et al., 2002; Wu et al., 2003).

Nevertheless, vector integration site analyses have been highly useful in analyzing the clonal repopulation dynamics of subsets of cells in the blood-forming system. The method generally involves fragmentation of the genomic DNA followed by PCR amplification with primers that recognize known viral DNA sequences flanking the sites of restriction (Bystrykh et al., 2012; Gentner et al., 2003; Mikkers et al., 2002). Thus, the restriction enzymes used must have at least one site of restriction within the vector backbone that is integrated into the genome. Early methods relied on gel electrophoresis to identify the DNA fragments that corresponded to specific clones (Figure 1.7). However, this approach limits the number of clones that can be tracked simultaneously, due to the low resolving power of a DNA gel. Subsequently developed methods rely on

the ligation of an adaptor to the ends of the fragmented DNA in order for the fragments containing part of the viral DNA and part of the host genomic DNA to be sequenced. Consequently, a comparison of the sequenced genomic DNA to the reference genome allows the sites of integration to be discriminated with higher confidence, thereby enabling larger numbers of clones to be tracked simultaneously (Harkey et al., 2007).

However, the dependency of site integration analysis methods on restriction endonucleases and subsequent PCR also brings certain limitations. The size of the DNA fragments produced from enzymatic restriction may introduce PCR amplification bias, since only 100-500bp fragments will be amplified efficiently. Other fragments will be underrepresented or absent, thus skewing the clonal data obtained (Bystrykh et al., 2012; Harkey et al., 2007). The selection of which restriction endonucleases to use is also important, in order to maximize the coverage of the genomic DNA that is digested (to yield smaller fragments). This in turn is influenced by the representation of such restriction sites, and where they may be located within a compacted chromatin structure. These considerations are important in order to ensure as many cells of each clone are represented as possible, and also to sufficiently digest the genomic DNA into small DNA fragments, such that they can be PCR amplified. It has been suggested that multiple restriction enzymes can be used in combination, and *in silico* analysis has suggested that a combination of Tsp5091, MspI, MaeII, HspA1, NlaIII, MaeI, MboI enzymes will cover ~98% of the genome (Bystrykh et al., 2012).

There have been numerous adaptations of this approach in terms of how the DNA is PCR-amplified and how the adaptors are ligated onto the DNA fragments for sequencing (Bystrykh et al., 2012). One of these approaches, called linear-amplification

mediated PCR (LAM-PCR) or ligation-mediated PCR (LM-PCR) has been suitable for high-throughput analysis by coupling the strategy with massively parallel sequencing (MPS) platforms (Harkey et al., 2007; Kustikova et al., 2005; Schmidt et al., 2001). However, many of the limitations inherent to restriction-based and PCR-based site integration methods remain. To circumvent the bias introduced from digestion with restriction endonucleases, nonrestricted LAM-PCR (nrLAM-PCR) was developed, in which a step of short linear DNA synthesis creates multiple short amplicons that can then be directly ligated with adaptors for subsequent amplification before sequencing (Gabriel et al., 2009). However, it is difficult to precisely control the length of DNA fragments generated. Thus, although this step may reduce bias from use of restriction enzymes, it does not substantially reduce PCR amplification bias.

An alternative approach to uniquely labeling cells is to introduce a unique non-coding DNA barcode into the viral vector that gets integrated into the cells to be tracked (Gerrits et al., 2010; Schepers et al., 2008). In order to employ this method of "cellular barcoding", several essential considerations must be made. The barcode oligonucleotide sequences need to allow for a sufficiently diverse library of plasmids to be generated, such that when many cells are transduced, the probability of two cells acquiring the same barcode are minimal. The library of barcoded plasmids, once constructed need to be validated to have no significant bias in the representation of specific barcode sequences within the library. Also, when the cell population is transduced with the barcoded viruses, appropriate measures need to be in place to minimize the occurrence of multiple integrations, which would alter interpretation of the data obtained and depend on the use

of computational methods to detect and remove likely redundant barcoded clones (Bystrykh et al., 2012; Nguyen et al., 2014).

By its very design, the cellular barcoding approach circumvents many of the biases of PCR and restriction endonuclease-based methods. The tracking of cellular barcodes is dependent on PCR, but does not use restriction endonucleases. However, the identical length of each cellular barcode eliminates any PCR amplification bias due to fragment size. Furthermore, cellular barcodes can be designed to be contained within a single sequencing read on a MPS platform, making this approach suitable for high-throughput analysis (Cheung et al., 2013; Lu et al., 2011; Naik et al., 2013; Nguyen et al., 2014; Perie et al., 2014).

1.6.5 Lineage-tracing using genetic mouse models

Clonal tracking *in situ* can be accomplished by labeling clones with a fluorescent reporter or mulitple reporters that emit at different wavelengths or stain differentially in "knock-in" or inducible knock-in transgenic mice. The former involves using a lineage-specific promoter (e.g. CK5) to drive the expression of a recombinase enzyme, such as Cre-recombinase (Cre), Cre^{ER}. The latter makes use of a system like the reverse tetracycline transactivator protein (rtTA) that is activated by doxycycline. The recombinase then acts to turn on the reporter gene (e.g., GFP, YFP or LacZ) whose expression is determined by the lineage-specific promoter. A short low dose pulse to activate the recombinase is required to label single cells so that the progeny they produce, presumably within the same region, can be analyzed to infer the magnitude of clonal expansion (by counting the number of labeled cells within a single region of the tissue), and extent of lineage differentiation (by co-staining with antibodies that recognize lineage-specific

markers)(Blanpain, 2013; Blanpain and Simons, 2013; Hope and Bhatia, 2011). However, this approach has several caveats. Expression of the recombinase enzymes can be leaky, thus resulting in unintended labeling of cells, thereby confounding clonal analyses. Also, tamoxifen, used to induce the Cre^{ER} recombinase, may influence the normal development and physiology of estrogen-responsive tissues, such as the mammary gland (Rios et al., 2014). Clonal analysis is usually performed on tissue sections that capture only a single cell layer within the tissue, and clonal growth that occurs in 3D may not be accurately captured. Virtual 3D reconstruction of tissue sections analyzed using confocal microscopy has been proposed as a solution to this last caveat. Interestingly, this latter approach has already revealed results that conflict with previous lineage-tracing studies (Rios et al., 2014).

The latter option for clonal tracking *in situ* involves labeling many clones simultaneously, but with multiple different colours, or shades of colours that can be discriminated by microscopy. Early methods involved generating chimaeric mice by mixing stem cells of two different reporters (Hadjantonakis et al., 2002) or crossing two established mouse lines that express different reporters (Feng et al., 2000). These methods were fairly laborious and also limited the number of reporters that could be simultaneously tracked. Initially developed for clonal tracking in the nervous system, "Brainbow" 1.0 and 2.0 were developed to allow cells to be labeled with reporter combinations that would produce up to ~100 different fluorescent readouts. Brainbow 1.0 was designed with several incompatible *loxP* sites which sandwich different fluorescent reporter genes so that once Cre is driven, recombination occurs randomly to yield a myriad of different colour combinations. Brainbow 2.0 has a similar design, except the

fluorescent reporter genes are oriented in both forward and reverse directions, and the incompatible *loxP* sites designed to randomly flip the genes into different orientations. When the number of reporters and *loxP* sites cloned in tandem is increased, this further increases the number of possible colours with which cells can be labeled (Livet et al., 2007).

When the recombinase for the brainbow reporter is placed under the control of a lineage-specific promoter, it is possible to label the cells of any tissue. This was demonstrated in the "R26R-Confetti" mouse (eg. CK5-Cre/R26R-Confetti)(Snippert et al., 2010). The advantage of the brainbow or confetti mouse is that by initially labeling cells with tens of different colours, many clones can be tracked simultaneously and the probability that two adjacent cells will be initially labeled with the same colour is low (Snippert and Clevers, 2011).

1.7 Development and evolution of malignant human mammary populations

1.7.1 Clinical staging and treatment of breast cancer

Breast cancer has historically been classified as Stage 0 to Stage IV depending on the extent to which the disease has progressed. These stages also have prognostic implications, and are used to guide clinical decisions on treatment. The American Joint Committee on Cancer designates breast cancer stages by TNM classification, which refers to characteristics of the primary tumour, lymph nodes, and metastases (Edge et al., 2010).

Stage 0 includes three types of breast carcinoma *in situ*: ductal carcinoma *in situ* (DCIS), lobular carcinoma *in situ* (LCIS) and Paget disease of the nipple. DCIS is a

noninvasive condition when abnormal cells are detected in the lining of the a breast duct but is still confined by the basement membrane. DCIS rarely presents as a palpable mass, and as such the majority of cases are detected by mammography (Fonseca et al., 1997). LCIS describes a situation where abnormal cells are detected in the lobules but has not yet invaded through the basement membrane. Women with LCIS also have a higher risk of developing invasive ductal carcinomas later in life (Fisher et al., 2004). Paget disease of the nipple refers to cases where abnormal cells are detected in the skin of the nipple, and frequently also have either DCIS or an invasive breast cancer within the same affected breast (Caliskan et al., 2008).

Breast cancer is classified as Stage IA when the primary tumour is ≤20mm (in its longest dimension), and Stage IB when there are micrometastases detected in the surrounding lymph nodes. Stage IIA involves a primary tumour that is >20mm and ≤50mm with no nodal metastases, or ≤20mm with metastases of the ipsilateral axillary lymph node(s) whereas Stage IIB involves a primary tumour that is >50mm with no nodal metastases or >20mm and ≤50mm with metastases of the ipsilateral axillary lymph node(s). Stage IIIA involves a greater extent of metastases to the surrounding lymph nodes whereas Stage IIIB involves a primary tumour of any size that has extended to the chest wall and/or skin and resulted in ulceration or visible skin nodules. Stage IIIC involves more extensive metastases, such as to the ipsilateral infraclavicular or supraclavicular lymph nodes. Lastly breast cancer is diagnosed as Stage IV when any distant metastases (>0.2mm) are detected and confirmed by histology (Edge et al., 2010).

Breast cancers are clinically divided into three groups that influence treatment options and prognosis: hormone receptor positive/HER2 negative ($ER^+PR^+HER2^-$),

HER2 positive, or triple negative ($ER^-PR^-HER2^-$) cancer. Stage I to III breast cancers typically require treatment using multiple modalities that are decided clinically based on a number of variables such as the extent of the cancer, patient age and health. Generally, however, Stage I and II breast cancer can be treated with breast conserving surgery followed by radiotherapy, or mastectomy. In either case, an attempt is made to remove the tumour completely along with a border of tumour-free tissue. The prognosis for a patient with early stage (Stage 0, I or II) breast cancer is generally good with the majority of patients experiencing long-term remission (NCI, 2014).

Advance stage breast cancer (Stage III or IV) ER^+ and/or PR^+ breast cancers are typically treated with 5 years of tamoxifen therapy in pre-menopausal women or 5 years of sequential tamoxifen and aromatase inhibitor in post-menopausal women. $HER2^+$ breast cancer is generally treated with trastuzumab (trade name Herceptin). Chemotherapy reagents are used in combination, and include antimetabolites such as 5-fluorouracil, anthracyclines such as doxorubicin and epirubicin, mitotic inhibitors such as paclitaxel and docetaxel, and alkylating agents such as cyclophosphamide and carboplatin. Patients with $ER^-PR^-HER2^-$ breast cancers are thought not to benefit from hormone therapy, and so chemotherapy is generally recommended (NCI, 2014; Shenkier et al., 2004).

1.7.2 Molecular subtypes of breast cancer correlate with clinical outcome

In 2000 and 2001, Perou et al. and Sorlie et al. reported their results of transcriptome analyses of a few hundred breast carcinomas using microarrays. Hierarchical clustering analysis produced 6 subtypes: Luminal A, Luminal B, Luminal C, $ERBB2^+$ ($HER2^+$) normal breast-like, and basal-like (Perou et al., 2000; Sorlie et al., 2001). These

molecular subtypes have been later refined to 4 molecular "intrinsic" subtypes where Luminal B combined the previously defined Luminal B and Luminal C subtypes, and the normal-like subtype is not included as a separate subtype because it is unclear whether they indeed represent a distinct group of breast tumours with any prognostic implications (Prat and Perou, 2011). Recently, an additional subtype of claudin-low breast cancers have been introduced based on studies of human and mouse tumours (Herschkowitz et al., 2007), as well as breast cancer cell lines (Prat et al., 2010). These are consistent clinically with $ER^-PR^-HER2^-$ invasive ductal carcinomas that are frequently metastatic and are typically associated with poor prognosis (Prat and Perou, 2011).

Luminal A breast cancers are characterized predominantly by expression of ESR1 (ERα), as well as other genes such as Bcl-2 (Kim et al., 2012). Luminal B breast cancers are characterized by increased expression of proliferation genes, such as MKI67 (Ki67) and PCNA (Gonzalez-Angulo et al., 2012). $HER2^+$ breast cancers have high expression of several genes including ERBB2 (HER2) and GRB7, whereas basal-like breast cancers have high expression of basal-associated KRT5, and also have a high prevalence of TP53 mutation (Cancer Genome Atlas, 2012; Perou et al., 2000).

Comparison of the molecular "intrinsic" subtypes to overall and relapse-free survival has shown statistically significant differences between all subtypes, with $HER2^+$ and basal-like breast cancers having the shortest survival times. The ERBB2 oncoprotein is a known prognostic marker associated with poor survival in breast cancer patients, in the absence of anti-HER2 targeted therapy. Recently, these intrinsic subtypes of breast cancer have been further refined to a set of 50 genes (PAM50) (Bastien et al., 2012) that is now being investigated in clinical trials for its predictive value in assessing clinical

outcome (Ellis et al., 2011). There have been several other prognostic tests developed for assessing likelihood of relapse that are also based on gene expression (Paik et al., 2004; Sotiriou et al., 2006; Tutt et al., 2008; van de Vijver et al., 2002).

A recent study by Curtis et al. analyzed the genomic and transcriptomic profiles of nearly 2,000 breast tumours, including a validation cohort of tumours. Using integrative clustering analysis, they discriminated 10 subgroups (termed IntCluster 1 to 10). These 10 subgroups have different gene expression profiles, copy number aberrations and distinct clinical outcomes (Curtis et al., 2012; Dawson et al., 2013). Interestingly, IntClust 4 is characterized by good prognosis, for a subset of ER^+ and ER^- tumours, indicating that not all ER^- tumours are associated with poor prognosis. Furthermore, IntClust 5 is composed of $HER2^+$ and luminal cases, suggesting a subset of ER^+ tumours may be responsive to trastuzumab therapy. IntClust 3 is composed of luminal A tumours, with good prognosis, whereas IntClust 2 is an ER^+ subgroup with poor clinical outcome, driven by known and putative drivers such as CCND1, EMSY, PAK1 and RSF1. IntClust 1, 6, 7, 8 and 9 are associated with intermediate prognosis, whereas basal-like tumours were associated with IntClust 10, which had high genomic instability, and relatively good prognosis after the first 5 years. This subgroup frequently had cis-acting alterations: chromosome 5 loss, 8q gain, 10p gain and 12p gain (Curtis et al., 2012).

1.7.3 Investigation of genomic diversification in human breast cancers

Advances in sequencing technology have now allowed genomic profiles of breast tumours to be analyzed (Shah et al., 2009). Single nucleotide variants, as well as insertions/deletions (indels) are detected at various frequencies, and used to infer clonal

genotypes based on algorithms used to model clonal and subclonal mutation clusters. From these models, it is also possible to infer early "driver status" mutations from those that occur later in the evolution of the tumour (Nik-Zainal et al., 2012; Shah et al., 2009; Shah et al., 2012). For example, p53 has been found to be the most prevalent mutation in triple negative/basal-like breast cancers, and frequently occurs in the highest clonal frequency group, as would be expected of a driver mutation. However, in some tumours it is found in the lower-abundance clonal frequency group, indicating it is not a founding mutation. In this way the mutations different tumour subclones acquire allow for a model of branched clonal evolution to be reconstructed (Greaves and Maley, 2012). However, in order to definitively confirm the existence of different clonal genotypes, tumours need to be sequenced at single-cell resolution.

The study of clonal diversification involves detecting the evolution of genomic clones over time. One approach is to infer the genomic clonal diversity through allele frequency detected in metastases to those in the primary tumour. These types of studies have been performed for human breast (Nik-Zainal et al., 2012; Shah et al., 2009), kidney (Gerlinger et al., 2012), lung (Takahashi et al., 2007a) and pancreatic cancers (Yachida et al., 2010). A different approach is to xenograft a patient's primary tumour into immunodeficient mice (Cariati et al., 2011; DeRose et al., 2011; Ding et al., 2010; Zhang et al., 2013) and then examine the clonal evolution that occurs with serial passages. However, caveats of the latter approach are that clonal selection may occur in the subset of clones that can engraft, or that clonal evolution in the mouse is not reflective of the selection pressures that drive evolution in the patient. The advantage of generating xenografts, however, is that mice can be subjected to different targeted or chemotherapy

agents in order to study drug efficacy on eliminating possibly rare tumour-propagating cells (Aparicio and Caldas, 2013; Nguyen et al., 2012).

1.7.4 Forward-engineering breast cancer from normal human mammary epithelial cells

A popular notion for which there is long-standing evidence is that malignant breast cancer cells reprogram their molecular state to resemble that of a fetal mammary stem cell, including the expression of genes that may enhance their self-renewal and proliferative activity (Spike et al., 2012; Veltmaat et al., 2003). Nevertheless, it seems plausible that some of the characteristics of the malignant cells that initiate and propagate malignant populations will at least, at the early stages of tumour development, reflect properties of the cell in which a self-sustaining malignant (invasive) potential was first acquired. This is commonly referred to as the "cell of origin" (Visvader, 2011).

The "most primitive" cell within the normal mammary gland (the cell that retains the greatest capacity for long-term regeneration of mammary tissue), is widely considered to be a likely target for initial predisposing events, since these cells already possess some of the defining properties of tumour-initiating cells (Nguyen et al., 2012). For human breast cancers, the likely cell of origin would then appear to be a BC, since these contain the MRUs and are normally $ER^-PR^-HER2^-$, similar to basal-like breast cancers that are the most aggressive and have poor prognoses. However, recent studies indicate that the LPs of the mammary epithelium have more developmental plasticity than previously recognized and hence may also accumulate mutations independently (Molyneux et al., 2010; Regan et al., 2012). In addition, recent studies indicate that LPs possess other

features that predispose them to accumulating mutations. These include their acquisition of very short telomeres and evidence of telomere associated DNA damage (Kannan et al., 2013) as well as elevated levels of reactive oxygen species (ROS) and ROS-associated DNA damage (Kannan et al., submitted). Certainly in the mouse, it is clear that both luminal and basal cells can become transformed (Chaffer et al., 2011; Keller et al., 2012; Molyneux et al., 2010; Proia et al., 2011; Regan et al., 2012).

An early approach to understanding the cellular origin of breast cancer made use of *in vitro* TERT-immortalized normal human mammary epithelial cells. However, these lines were not derived from prospectively purified subsets of mammary epithelial cells and had already undergone several passages *in vitro* (Hahn et al., 1999; Kendall et al., 2005). They thus set the stage for more sophisticated experiments with primary cells but provided little useful information about the relevance of the cell of origin. Subsequent studies with unseparated populations of primary normal human mammary epithelial cells were important in establishing that these could be transformed with a variety of oncogenes to produce different phenotypes. For example, *ESR1* + *BMI1* + *TERT* + *MYC* produced tumours that were ER^+, estrogen-dependent and genetically stable (Duss et al., 2007), whereas *OCT4* alone produced tumours that were poorly differentiated (Beltran et al., 2011). *TERT* + *SV40* + *HRAS* produced tumours that were well-differentiated and poorly tumorigenic (Chaffer et al., 2011), whereas $TP53^{R175H}$ + *CCND1* + *PIK3CA* + $KRAS^{G12V}$ (Keller et al., 2012; Proia et al., 2011) and *SV40* + $KRAS^{G12V}$ (Keller et al., 2012) both caused the formation of squamous tumours (expressing CK14, p63 and vimentin) with papillary regions (expressing ER, CK8/18 and CK19).

To date there have been several oncogene combinations used to transform purified subsets of primary normal human mammary epithelial cells. $EpCAM^{+}CD49f^{-}$ LCs transformed with *SV40* + $KRAS^{G12V}$ resulted in higher levels of ER and reduced expression of CK14 compared to $EpCAM^{+}CD49f^{+}$ cells. $CD10^{+}$ BCs transformed with either $TP53^{R175H}$ + *CCND1* + *PIK3CA* + $KRAS^{G12V}$ or *SV40* + $KRAS^{G12V}$ produced tumours that were poorly tumorigenic, and tumours with exclusively basal features (squamous, metaplastic and giant cell differentiation; lack of ER and luminal cytokeratin expression, and robust expression of CK14), respectively (Keller et al., 2012; Proia et al., 2011). These findings suggest that both the combination of oncogenes used for transformation, and the cell of origin may influence the characteristics of the resultant tumour.

1.8 Thesis objectives

The mammary gland is comprised of cells belonging to two distinct lineages. However, although the majority of these cells are thought to be terminally differentiated, a substantial and as yet poorly defined proportion appear to retain readily activated regenerative activity. These findings raise the possibility that the size and content of the clones these cells can produce *in vivo* are dependent on the conditions under which they are stimulated to divide. However, the diversity of this regenerative potential remains poorly characterized. This is an important issue as the properties inherent in normal cells likely determine the minimal changes required to enable them to become malignant.

In this thesis, I will examine both of these critical aspects of growth regulation in the mammary gland and during malignant transformation. ***Whereas limiting dilution***

transplants of mouse and human mammary epithelial cells have shown that MRUs demonstrate bi-lineage differentiation, I hypothesize that under non-limiting conditions some clones will display lineage-restricted differentiation. Furthermore, I hypothesize that upon transformation with defined oncogenes, these normal patterns of mammary epithelial cell differentiation will be perturbed, and inform on how the cell of origin may influence the phenotypic and functional properties of human breast tumours generated de novo.

As a first step, I therefore sought to develop a more powerful approach to analyzing clonal growth dynamics *in vivo* that could be applied to large numbers of co-transplanted cells with regenerative potential, thus more rapidly achieving a situation that might simulate the state of the adult mammary gland. The method of choice for this work required the development and validation of a library of barcoded lentiviral vectors and a method to reliably convert the sequence data into clone size data. This work is described in Chapter 2.

The syngeneic mouse transplant model for detecting the *in vivo* regenerative activity of mouse MRUs is significantly more robust in terms of mature cell production compared to the human xenograft model. Therefore I chose to first test my hypothesis about the diversity of regenerative cell behavior by applying the barcoded library I generated to purified mouse mammary cells assayed in this system. This work forms the substance of Chapter 3.

It is possible that any normal viable cell, regardless of whether it demonstrates regenerative activity can be the target for oncogenic transformation. The role of the cell of origin in determining the molecular and functional characteristics of breast tumours is

still largely unknown, but is likely to be dependent on both the cell of origin and the perturbations with which the cells are transformed. Thus, I next sought to examine the role of the cell of origin on the molecular and functional properties of breast tumours generated *de novo* and compare the perturbations of clonal growth with those of their normal counterparts and those of tumours originating in patients and subsequently xenografted under the same conditions. The results of these experiments are described in Chapter 4.

1.9 Figures & tables

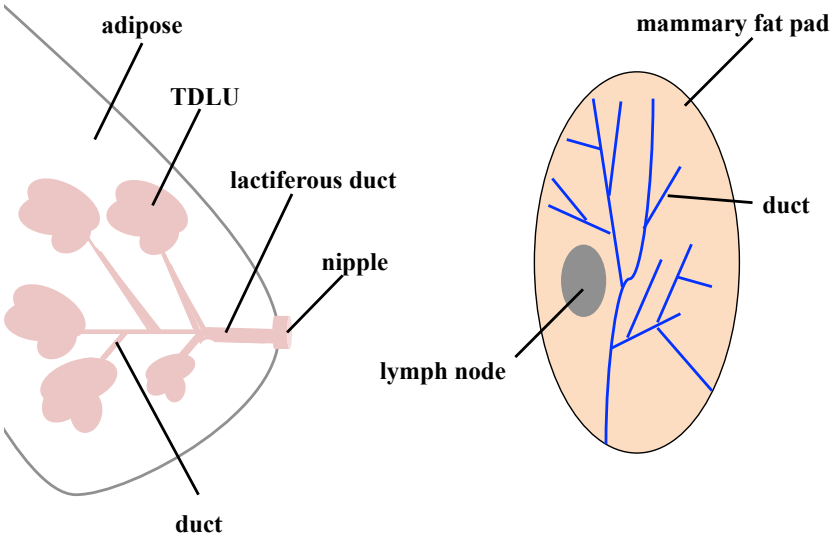

Figure 1.1 Structure of the mammary gland in human and mouse

Schematic representations of the human (left) and mouse (right) mammary glands.

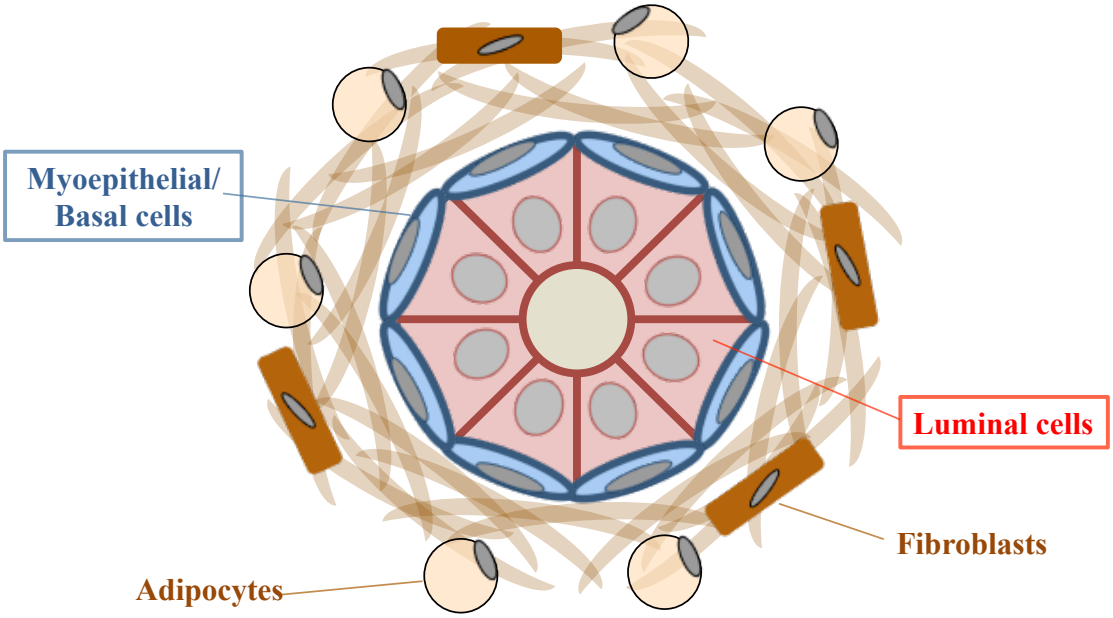

Figure 1.2 Cross-sectional representation of a mammary duct

The mammary duct depicted is embedded in a dense, fibrous stroma as is characteristic of the human mammary gland. In the mouse, the environment surrounding the mammary duct is richer in adipose tissue.

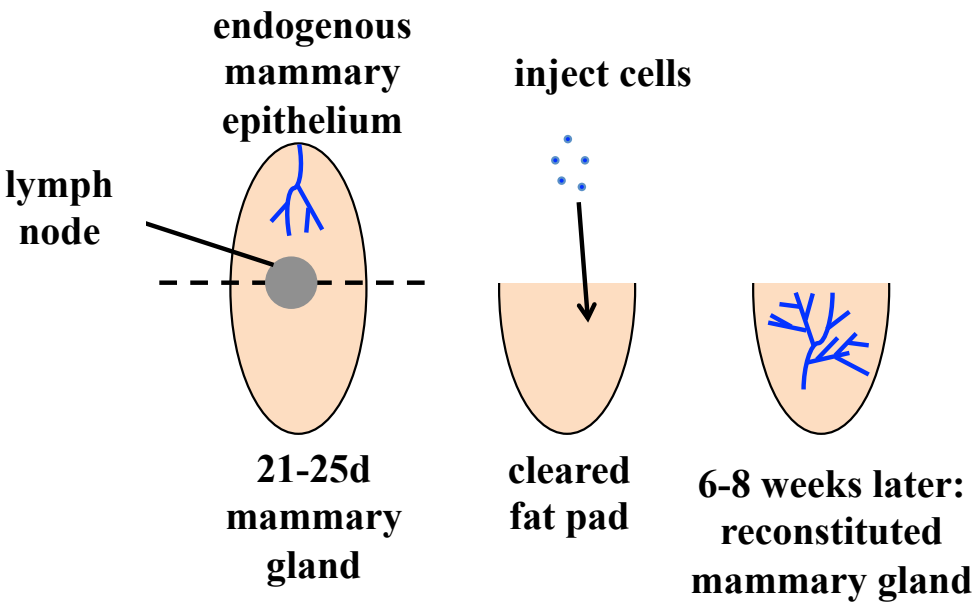

Figure 1.3 Cleared mouse mammary fat pad transplant procedure

A schematic depiction of the procedure shows how a single cell suspension of mouse mammary epithelial cells is injected into a pre-cleared fat pad to assess whether a macroscopically visible branced mammary tree can be visualized 6-8 weeks later as evidence of the presence in the initial innoculum of cells with *in vivo* mammary gland regenerative activity.

Table 1.1 Reported MRU frequencies in different subsets of mouse mammary epithelial cells

Donor mouse strain	Subset analyzed	Frequency (95%CI)	Reference
B6;129S-GtROSA26	**Sca1^{+}**	**1/1** (1/1-1/2,750)	(Welm et al., 2002)
	Sca1^{-}	**1/74,812** (1/26,774-1/209,044)	
	SP (5d culture)	**<1/75,000**	
FVB	**SP** (no culture)	**1/34,440** (1/14,324 - 1/82,806)	(Alvi et al., 2003)
FVB	**Bulk**	**1/1,400** (1/600-1/3,000)	(Stingl et al., 2006a)
	CD45^{-}Ter119^{-}CD31^{-}CD140a^{-} **CD24^{++}CD49f^{+}**	**<1/230**	
	CD45^{-}Ter119^{-}CD31^{-}CD140a^{-} **CD24^{+}CD49f^{++}**	**1/62** (1/37-1/100)	
	CD45^{-}Ter119^{-}CD31^{-}CD140a^{-} **CD24^{lo}CD49f^{lo}**	**1/3,400** (1/1,300-1/9,100)	
C57BL/6	CD45^{-}Ter119^{-}CD31^{-}CD140a^{-} **CD24^{++}CD49f^{+}**	**<1/6100**	
	CD45^{-}Ter119^{-}CD31^{-}CD140a^{-} **CD24^{+}CD49f^{++}**	**1/91** (1/51-1/160)	
not specified (FVB or C57BL/6)	CD45^{-} **CD49f^{+}Rho^{-}**	**1/40** (1/13-1/127)	
	CD45^{-} **CD49f^{+}Rho^{+}**	**1/1,100** (1/440-1/2,700)	
	CD45^{-} **CD49f^{+}Rho^{-}**	**<1/63**	
	CD45^{-} **CD49f^{+}Rho^{lo}**	**1/1,100** (1/400-1/3,100)	
	CD45^{-} **CD49f^{+}Rho^{hi}**	**>1/3,200**	

(Table continued on subsequent page…)

Donor mouse strain	Subset analyzed	Frequency (95%CI)	Reference
FVB	$CD45^-Ter119^-CD31^-$ **$Sca\text{-}1^{lo}CD49f^{hi}$**	**1/19** (1/9-1/41)	(Stingl et al., 2006a)
not specified (FVB/NJ, C57BL/6, BALB/c)	$CD45^-CD31^-Ter119^-$	**1/4,900** (1/3,200-1/7,500)	(Shackleton et al., 2006)
	$CD45^-CD31^-Ter119^-$ **$CD29^{lo}CD24^-$**	**1/147,000** (1/37,000-1/590,000)	
	$CD45^-CD31^-Ter119^-$ **$CD29^{lo}CD24^+$**	**<1/7200**	
	$CD45^-CD31^-Ter119^-$ **$CD29^{hi}CD24^-$**	**1/2,900** (1/1,100 - 1/7,800)	
	$CD45^-CD31^-Ter119^-$ **$CD29^{hi}CD24^+$**	**1/590** (1/300-1/1,100)	
	$CD45^-CD31^-Ter119^-$ **$FSC^{lo}Sca\text{-}1^{hi}$**	**1/30,000** (1/10,000-1/93,000)	
	$CD45^-CD31^-Ter119^-$ **$FSC^{lo}Sca\text{-}1^{mid\text{-}lo}$**	**1/8,900** (1/5,100-1/16,000)	
	$CD45^-CD31^-Ter119^-$ **$FSC^{hi}Sca\text{-}1^{lo\text{-}hi}$**	**1/37,000** (1/5,200-1/260,000)	
	$CD45^-CD31^-Ter119^-$ **Hoechst-MP**	**1/2,900** (1/1,600-1/5,100)	
	$CD45^-CD31^-Ter119^-$ **Hoechst-SP**	**1/3,300** (1/470-1/23,000)	
	$CD45^-CD31^-Ter119^-$ **$CD29^{lo}CD24^+$** (double sorted)	**<1/1,100**	
	$CD45^-CD31^-Ter119^-$ **$CD29^{hi}CD24^-$** (double sorted)	**<1/1,100**	
	$CD45^-CD31^-Ter119^-$ **$CD29^{hi}CD24^+$** (double sorted)	**1/64** (1/46-1/90)	
C57BL6/J	**bulk** ($EpCAM^+$ cell equivalent)	**0.8%** (0.3%-2%)	(Makarem et al., 2013)
	$CD45^-CD31^-Ter119^-BP\text{-}1^-$ **$EpCAM^+CD49f^+$**	**0.3%** (0.1%-1%)	
	$CD45^-CD31^-Ter119^-BP\text{-}1^-$ **$EpCAM^{++}CD49f^+CD61^+$**	**0.01%** (0.004%-0.04%)	

Table 1.2 Reported lineage-tracing studies detecting long-term clones *in situ* with bi-lineage or lineage-restricted differentiation

Mouse model	Pulse timing	Chase observations	Reference
WAP-Cre/Rosa-lacZ	Adult	• Labels LP cells that contribute to alveolar differentiation during pregnancy and lactation • A minor subset remain after involution, and continue to contribute to luminal and alveolar differentiation through several cycles of pregnancy	(Wagner et al., 2002)
WAP-Cre/Rosa-lacZ	Adult	• Explant cultures from virgin female mice cultured with insulin, hydrocortisone and prolactin induced secretory differentiation • Estrogen and progesterone do not induce secretory differentiation	(Booth et al., 2007)
K14-Cre/Rosa-YFP	E17	• YFP^+ myoepithelial and luminal cells	(Van Keymeulen et al., 2011)
	puberty and adult	• YFP expression restricted to the myoepithelial lineage (through two cycles of pregnancy, lactation and involution)	
K5-Cre^{ER}/Rosa-YFP	P1 and P28	• YFP expression restricted to the myoepithelial lineage (through puberty and pregnancy)	
Lgr5-GFP-Cre^{ER}/Rosa-Tomato	P28	• Lgr5 expression is observed predominantly in the myoepithelial lineage, and rare LCs	
K8-Cre^{ER}/Rosa-YFP	P1, P28 and adult	• YFP expression restricted to the luminal lineage (after 10 weeks or through puberty or pregnancy)	
K18-Cre^{ER}/Rosa-YFP	P28 and P56	• YFP expression restricted to the luminal lineage (after 10 weeks or through puberty or pregnancy)	

(Table continued on subsequent page…)

Mouse model	Pulse timing	Chase observations	Reference
Axin2-Cre^{ERT2}/Rosa26-mT/mG	E12.5	• GFP^{+} LCs only	(van Amerongen et al., 2012)
	E14.5	• GFP^{+} LCs only	
	E17.5	• GFP^{+} LCs only	
Axin2-Cre^{ERT2}/Rosa26-lacZ and Axin2-Cre^{ERT2}/Rosa26-mT/mG	P14 to P16	• LacZ^{+} and GFP^{+} cells predominantly myoepithelial, with detectable luminal cells in 50% of labeled mice	
Axin2-Cre^{ERT2}/Rosa26-mT/mG	Pre-puberty	• In 10-12 week virgin mice, GFP+ myoepithelial and luminal cells were detected	
Axin2-Cre^{ERT2}/Rosa26-lacZ and Axin2-Cre^{ERT2}/Rosa26-mT/mG	adult virgin	• 48h to 66d later, predominantly labeled myoepithelial cells, with few LCs detected in 38% of mice	
Axin2-Cre^{ERT2}/Rosa26-mT/mG	adult virgin	• GFP^{+} myoepithelial and luminal cells (through two cycles of pregnancy and lactation)	
Elf5-rtTA/TetO-Cre/Rosa26-Confetti	puberty	• labeled clones restricted to the luminal lineage (throughout puberty and adulthood)	(Rios et al., 2014)
Elf5-rtTA/TetO-Cre/Rosa26-Confetti	P63	• labeled clones restricted to the luminal lineage	
K5-rtTA/TetO-Cre/Rosa26-Confetti	puberty	• labeled clones contributed to both myoepithelial and luminal lineages	
K5-rtTA/TetO-Cre/Rosa26-Confetti	adult	• After 8 weeks, labeled clones were myoepithelial-restricted or bi-lineage • Bi-lineage clones had equal proportions of myoepithelial and luminal cells or were enriched for LCs	
K14-Cre^{ERT2}/Rosa26-Confetti	puberty	• After 8 weeks, equal numbers of labeled myoepithelial and luminal cells were detected	
Lgr5-GFP-IRES-Cre^{ERT2}/Rosa26-tdTomato	adult	• After 8 weeks, both labeled myoepithelial and luminal cells were detected	

Table 1.3 Reported CFC frequencies in mouse mammary epithelial cell subsets

Donor strain	Subset analyzed	Frequency	Reference
Parkes mice	**bulk**	**7.0%** ± 1.8%	(Smalley et al., 1998)
	33A10+ (luminal)	**3.2%** ± 1.6%	
	JB6+ (myoepithelial)	**2.9%** ± 1.8%	
FVB	bulk	**1/63** (1/45-1/108, 95% CI)	(Stingl et al., 2006a)
not specified (FVB/NJ, C57BL/6, BALB/c)	$CD45^-CD31^-Ter119^-$ **$CD29^{lo}CD24^-$**	**~1%**	(Shackleton et al., 2006)
	$CD45^-CD31^-Ter119^-$ **$CD29^{hi}CD24^-$**	**~3%** ± ~2%	
	$CD45^-CD31^-Ter119^-$ **$CD29^{hi}CD24^+$**	**~23%** ± ~4%	
	$CD45^-CD31^-Ter119^-$ **$CD29^{lo}CD24^+$**	**~7%** ± ~4%	
Gata-$3^{+/f}$	$CD31^-CD45^-$ **$CD29^{lo}CD24^+CD61^-$**	**~12%** ± ~4%	(Asselin-Labat et al., 2007)
	$CD31^-CD45^-$ **$CD29^{lo}CD24^+CD61^+$**	**~32%** ± ~6%	
C57BL6/J	**bulk** ($EpCAM^+$ cell equivalent)	**27%** ± 2%	(Makarem et al., 2013)
	$CD45^-CD31^-Ter119^-BP\text{-}1^-$ **$EpCAM^+CD49f^+$**	**22%** ± 2%	
	$CD45^-CD31^-Ter119^-BP\text{-}1^-$ **$EpCAM^{++}CD49f^+$**	**9%** ± 1%	
	$CD45^-CD31^-Ter119^-BP\text{-}1^-$ **$EpCAM^{++}CD49f^+CD61^+$**	**13%** ± 3%	

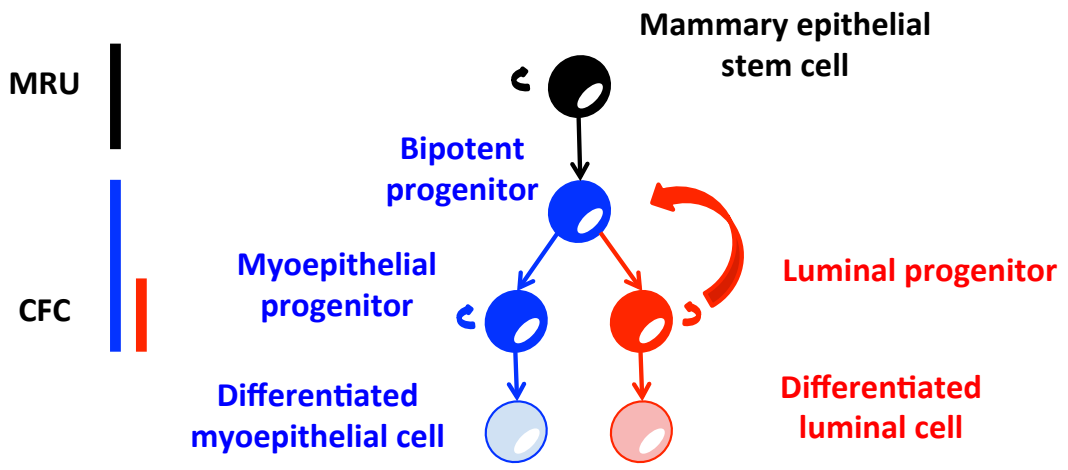

Figure 1.4 Mammary epithelial cell differentiation hierarchy

Shown is a simplified model of the mammary epithelial cell differentiation hierarchy in mice and humans. However, in the mouse bi-potent progenitors have not been detected in CFC assays (because their correct differentiation *in vtiro* is not preserved), and in humans, there is no evidence of self-sustaining long-term myoepithelial or luminal progenitor cells from *in situ* lineage tracing as there have been in mouse models.

Table 1.4 Reported CFC frequencies in human mammary epithelial cell subsets

Population	Frequency	Reference
$MUC1^{-}CD10^{-}$	**0.1% - 1.0%**	(Stingl et al., 1998)
$MUC1^{+}CD10^{-}$	**1% - 2.1%**	
$MUC1^{-}CD10^{+}$	**0.2% - 2.1%**	
$MUC1^{+}CD133^{+}EpCAM^{+}CD49f^{+}CD10^{-}CD90^{-}$	**~30%**	(Stingl et al., 2005)
$MUC1^{-}CD133^{-}EpCAM^{+}CD49f^{+}CD10^{+}CD90^{+}$	**~50%**	
$EpCAM^{+}CD49f^{-}$	**1/700**	(Villadsen et al., 2007)
$EpCAM^{+}CD49f^{+}$	**19/700**	
$EpCAM^{lo}CD49f^{+}$	**2/700**	

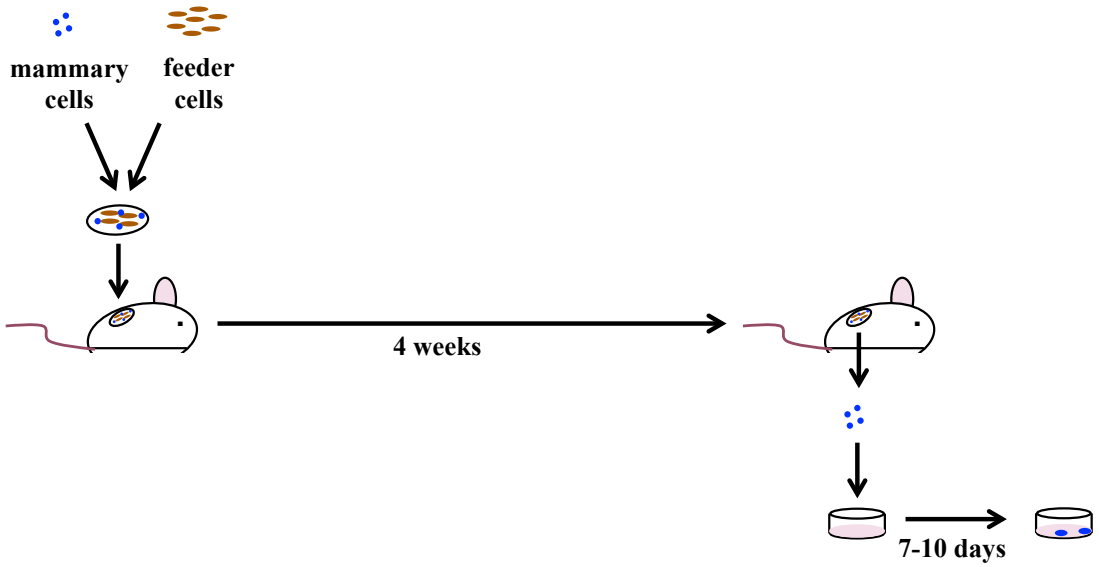

Figure 1.5 Heterotopic xenograft assay to detect human MRU activity

Shown is a depiction of the human MRU assay. A suspension of human mammary epithelial cells are embedded into a solid collagen gel along with irradiated C3H-10T1/2 mouse embryonic fibroblasts, then transplanted under the renal capsule of immunedeficient mice. After 4 weeks the gels are retrieved, the regenerated cells dissociated, and plated into a 2D tissue culture plate to measure in CFC frequency *in vitro*. The number of regenerated CFC was shown to correlate linearly to the number of input MRU, and thus can be used to estimate the input MRU frequency (Eirew et al., 2010; Eirew et al., 2008).

Table 1.5 Reported MRU frequencies in human mammary epithelial cell subsets

Reference	Recipient mouse	Subset analyzed	Frequency 1 in X (95%CI)
(Eirew et al., 2008)	NSG	**bulk**	**1/4,890** (1/2,380-1/10,080)
		bulk	**1/2,220** (1/1,060-1/4,690)
		bulk	**1/1,390** (1/640-1/2,960)
	NS/B2m-/-	**bulk**	**1/9,840** (1/4,910-1/19,700)
	NS	**bulk**	**1/1,600** (1/630-1/4,060)
(Lim et al., 2009)	NSG	$CD45^-CD31^-$ **$CD49f^{hi}EpCAM^{lo}$**	**1/21,500** (1/38,000-1/12,000)
		$CD45^-CD31^-$ **$EpCAM^{mid\text{-}hi}CD49f^{lo\text{-}mid}$**	not detected
		$CD45^-CD31^-$ **$EpCAM^-CD49f^-$**	

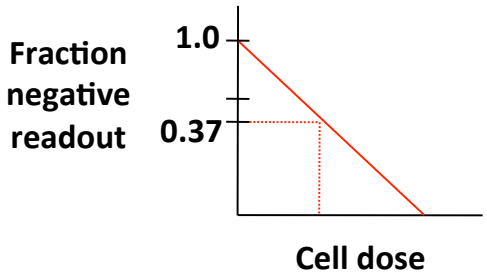

Figure 1.6 Limiting dilution analysis

A correlation between the cell dose assayed and the fraction negative readout to determine the dose at limit for a single cell of interest is shown. Assuming a Poisson distribution, the frequency of the cell of interest (in this case stem cell activity) is determined by the relationship $F_0 = e^{(-m)}$ where F_0 is the fraction negative readout at a particular cell dose transplanted, and m is the average number of cells per culture. Thus, to calculate the cell dose where on average one stem cell is transplanted, m = 1, and F_0 = 0.368 (or ~37% fraction negative readout)(Eirew et al., 2010).

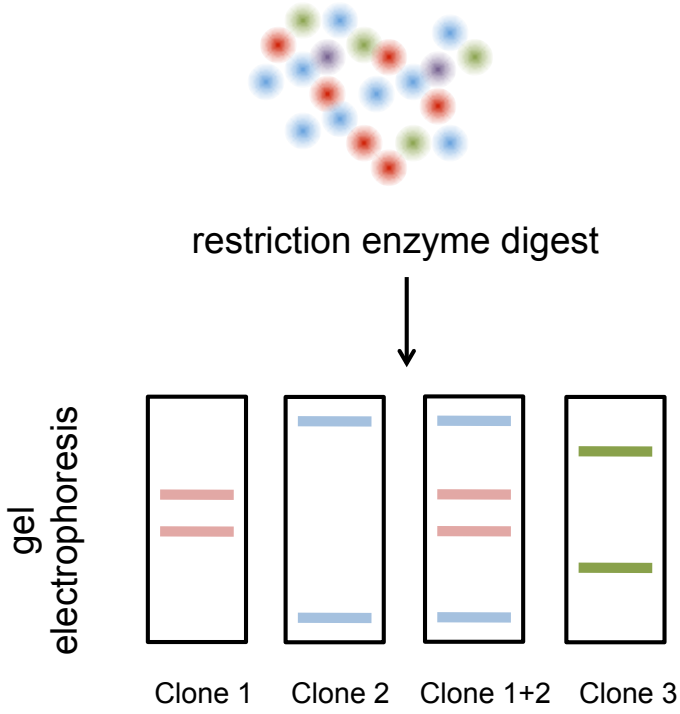

Figure 1.7 Clones analyzed by site integration analysis are resolved by gel electrophoresis

Shown is a schematic of how a single polyclonal sample would be analyzed by gel electrophoresis after restriction enzyme digest, using an enzyme with a restriction site within the vector backbone.

CHAPTER 2: DEVELOPMENT AND VALIDATION OF A SINGLE-CELL GENOMIC BARCODING APPROACH FOR TRACKING *IN VIVO* REGENERATED CLONAL POPULATIONS

2.1 Introduction

Cellular barcoding is a method whereby short non-coding DNA-based sequences are integrated into the genome of individual cells, typically by retroviral or lentiviral transduction, to serve as permanent unique identifiers of their clonal progeny. Several high-complexity retroviral-based or lentiviral-based barcode libraries have been generated and successfully used by several groups for this purpose (Gerrits et al., 2010; Grosselin et al., 2013; Lu et al., 2011; Naik et al., 2013; Schepers et al., 2008). However, analytical methods to discriminate true clones from background, to eliminate redundancy of barcode detection (clones with multiple integrations), and to estimate clone size have not been well developed.

Definition of a threshold that can discriminate a true barcode clone from background noise is essential for the correct interpretation of barcode data. This is especially important when sequencing is performed on a MPS platform after PCR amplification, as both of these may introduce low levels of erroneous barcodes. Previous methods have set arbitrary thresholds based on an inflection point (Lu et al., 2011; Naik et al., 2013), on the assumption that false barcodes will contain a low number of sequence reads whereas true barcodes contain a high number of sequence reads. However, such methods have not been rigorously tested, and as such no measures of sensitivity, specificity and reproducibility of true barcode detection using these methods have been provided.

Another issue that may confound the interpretation of clonal data is the frequency of multiple integrations. Previous methods have attempted to reduce the number of multiple integrants by maintaining a multiplicity of infection (MOI) of 1 or lower, and/or by applying algorithms that reduce the number of barcode clones in each dataset, from an estimated MOI (Gerrits et al., 2010; Lu et al., 2011; Naik et al., 2013). However, these methods lack the ability to identify which barcode clones may be redundant to specifically eliminate only those barcodes.

Lastly, all previous clonal tracking methods have been semi-quantitative, reliant on relative degrees of detection on a microarray platform (Schepers et al., 2008), frequency of clone detection when using Sanger sequencing (Gerrits et al., 2010), or relative read abundances when using a MPS platform (Lu et al., 2011; Naik et al., 2013). Use of such relative measures of clone size restricts comparisons of clonal abundance to single samples (even when multiple samples are analyzed using the same platform) because the number of unknown experimental clones as well as their absolute sizes will differ between experimental samples, thus altering the relative representation of each clone. Furthermore, it is assumed that the clonal abundances within a single sample correlate linearly with barcode frequency, which has not been experimentally validated.

To address these issues we first prepared and characterized a diverse library of lentiviral vectors. We then used it to develop an approach for the analysis of clonal tracking data that accurately detects barcode clones with a measurable sensitivity, specificity, reproducibility, identifies redundant barcodes (due to multiple integrants), provides a clone size estimate in terms of absolute cell content, and also calibrates each sample analyzed such that the data obtained can be compared between experiments. We

have also adapted our sequencing protocol for multiplexing of libraries (Wiegand et al., 2010), which improves the time and cost-effectiveness of sequencing relatively low complexity libraries.

2.2 Materials and methods

2.2.1 MNDU3-PGK-GFP lentiviral vector

The MNDU3-PGK-GFP (MPG) lentiviral vector (Logan et al., 2004) is 7.9 kb, encoding a GFP fluorescence reporter gene downstream of the PGK promoter. To express a gene of interest, a multiple cloning site was located downsteam of the MNDU3 promoter, but left empty here for the purposes of generating a biologically neutral barcode library. The vector was designed with a self-inactivating (SIN) region within the U3 region of the 3' long terminal repeat (LTR) such that after reverse transcription and integration into infected cells, the virus cannot self-replicate. These elements are contained between the 5' and 3' LTR, and thus are integrated into the host cell genome upon infection.

2.2.2 Barcode library construction

Barcode oligonucleotides were designed using forward (5'-TCGAGAAGTAANNATCNNGATSSAAANNGGTNNAACNNTGTAAAACGACGGCCAGTGAGC-3') and reverse (5'-CCGGGCTCACTGGCCGTCGTTTTACANNGTTNNACCNNTTTSSATC-NNGATNNTTACTTC-3') oligonucleotide sequences flanked by a 5' *Xho*I restriction site and a 3' *Xma*I restriction site and were ordered (Life Technologies) with a 5' phosphorylation modification, annealed in an equimolar ratio, and directly ligated into the MPG vector downstream of the GFP reporter cDNA. DH10B

bacterial (Life Technologies) were transformed with the ligated vector and plated at clonal density onto ampicillin-terrific broth (Life Technologies) agar plates. An initial 141 bacterial colonies were individually picked and analyzed by Sanger sequencing. The bacterial colonies were pooled, homogenized, and plasmids were purified following the manufacturer's recommendations (MaxiPrep, Qiagen) and the purified plasmids used for production of lentiviral packaging as described below.

2.2.3 Preparation of high-titer lentiviral supernatants

Lentiviral particles were packaged using the human embryonic kidney 293T cell line, transiently transfected with the barcoded plasmid DNA in addition to vectors encoding (1) VSV-G, envelope protein, (2) POL, reverse transcriptase and integrase enzymes, and (3) GAG, capsid protein as described (Imren et al., 2004). Expression of VSV-G, POL and GAG allow for production of lentiviral particles containing the barcoded vector. Following an initial media change at 24 hours post-transfection, the media containing active lentiviral particles was harvested at 48 and 72 hours post-transfection, filtered through a 0.45 μm mesh, and stored at -70°C.

To calculate the infectious viral titer, 10-fold serial dilutions of an aliquot of the supernatant were used to transduce HeLa cells (from 10 to 1,000-fold). Seventy-two hours after transduction, the cells were harvested and analyzed for percent GFP^+ cells. The viral titer was then calculated as the total cells transduced x the percent GFP^+ cells (e.g. 0.1 for 10% positive) x the dilution factor (e.g. 1,000 for a 1,000-fold dilution). Here the dilution resulting in <30% GFP^+ cells was used for this calculation viral titer, because

it represents the best estimate for single integrations. The titer of the barcoded lentiviral virus was thus determined to be 10^9 infectious units/ml.

2.2.4 Preparation of spiked-in control cells

Bacterial colonies containing the plasmids with barcodes for the individual spiked-in controls were sub-cloned, purified by MaxiPrep (Qiagen), and used to make a high titer lentivirus. Both FACS-purified $CD10^+CD90^+CD49f^+$ human basal mammary cells (see below, 2.2.5) and 184-hTERT immortalized human mammary epithelial cells (Raouf et al., 2005) were transduced and after 3 days of culture, 10, 20, 100, 250, 500 and 1,000 GFP^+ cells were sorted per well into 96 well plates. These were combined with samples of mouse or human mammary cells from which barcode amplicon libraries were constructed for MPS.

2.2.5 Dissociation of human mammary epithelial cells

Tissue from reduction mammoplasty surgeries were collected with informed consent, as approved by the University of British Columbia Research Ethics Board, and dissociated as previously described (Eirew et al., 2008). Briefly, the tissue was minced with a scalpel, and dissociated for 18 hours at 37°C in Dulbecco's Minimal Essential Media (DMEM) / Ham's F12 media (1:1, STEMCELL Technologies) supplemented with 2% bovine serum albumin (BSA, Gibco), 300 U/ml collagenase (Sigma) and 100 U/ml hyaluronidase (Sigma). "A" pellets, rich in mammary epithelial organoids, were obtained by an initial centrifugation at 80 g for 4 minutes. "A" pellets were cryopreserved at -156°C in

DMEM/F12 containing 50% fetal bovine serum (FBS, STEMCELL Technologies) and 6% dimethylsulfoxide (DMSO).

Prior to use, a vial of cryopreserved "A" pellet was thawed, rinsed with Hank's Balanced Salt Solution supplemented with 2% FBS (referred to as "HF"), and then the cells were enzymatically dissociated in 2.5 mg/ml trypsin with 1 mM EDTA (STEMCELL Technologies) and 5 mg/ml dispase (STEMCELL Technologies) with 100 μg/ml DnaseI (Sigma), washing with HF between each step. The resulting cell suspension was passed through a 40 μm mesh to obtain a single cell suspension.

2.2.6 Dissociation of mouse mammary epithelial cells

Mouse mammary glands were isolated from 8-12 week-old normal virgin female C57Bl/6J mice and single cell suspensions generated as previously described (Makarem et al., 2013). Mice were bred and used in the Animal Resource Centre of the BC Cancer Research Centre according to protocols approved by the Animal Care Committee of the University of British Columbia in keeping with Canadian Council of Animal Care guidelines. Briefly, the isolated mammary glands were dissociated for 18 hours at 37°C in DMEM/F12 media (STEMCELL Technologies) containing 1 mg/ml collagenase A (Roche Diagnostics) and 100 U/ml hyaluronidase (Sigma). After 18 hours, the glands were vortexed to disrupt the tissue, and erythrocytes lysed with 0.8% ammonium chloride with 0.1 mM EDTA in water (STEMCELL Technologies) prior to dissociation with trypsin-EDTA and dispase-DNaseI similar to human cells (described above).

2.2.7 Transduction and pre-culture of human and mouse mammary cells

Human mammary cells were transduced in SF-7 media supplemented with 5% FBS (Raouf et al., 2008), which uses DMEM/F12 (STEMCELL Technologies) as the base media supplemented with 0.1% BSA, 0.5 μg/ml hydrocortisone, 1 μg/ml insulin, 10 ng/ml EGF (Sigma), and 10 ng/ml cholera toxin (Sigma). The cells obtained were then cultured at 37°C and 20% O_2 for 3 days on tissue culture plates coated with Matrigel (1:60, BD Biosciences) for primary human mammary epithelial cells, and without Matrigel for 184-hTERT cells. Lentviral transduction was performed in liquid suspension cultures containing cells at a concentration of $\leq 10^6$ cells in 100 μl of SF-7 media supplemented with 5% FBS. Once the lentivirus was added to the cells at a concentration of $\sim 10^6$ infectious units per 100 μl reaction volume, the cells were incubated at 37°C and 20% O_2 for 4 hours, and then washed twice with HF prior to being used for any subsequent experimental procedure.

Mouse mammary cells were similarly transduced but in FAD rather than SF-7 medium. FAD media consists of DMEM/F12 supplemented with 10% FBS, 10 ng/ml EGF, 1.8×10^{-4} M adenine (Sigma), 5 μg/ml insulin, 0.5 μg/ml hydrocortisone, 10^{-10}M cholera toxin, and 10 μM Y-27632 (Reagents Direct).

2.2.8 Construction of barcode amplicon libraries for MPS

Genomic DNA was extracted from samples containing the spiked-in control cells using a PrepGEM DNA extraction kit (ZyGEM). To leverage the capacity of current MPS sequencers to generate $>10^9$ reads in a single run, we also introduced a separate fault-tolerant sequenced-based index to uniquely identify experimental groups using a plate-

based library construction protocol. In this step, barcode amplicons were generated in a 35-cycle PCR reaction using sequence-specific primers with adaptors compatible with Illumina PE1 and PE2 primers (Illumina Inc.). Then, in a second 8-10 cycle PCR reaction, they were pooled at equimolar ratios and loaded into a single lane of a flow cell for paired-end sequencing on an Illumina HiSeq 2000, or a MiSeq platform using a custom index sequencing primer for read 2 (Wiegand et al., 2010). To improve cluster recognition, a control phiX library was spiked into the amplicon libraries prior to sequencing (~40% by mole by HiSeq and 7% by MiSeq). The forward and reverse sequence-specific primers used were: ACACTCTTTCCCTACACGACGCTCTTCCGATCT and CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT, respectively, and the Illumina PE1 and PE2 primer sequences were: AATGATACGGCGACCACCGAGATC-TACACTCTTTCCCTACACGACGCTCTTCCGATCT and CAAGCAGAAGACGGCATACGAGATNNNNNNCGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCT, respectively.

2.2.9 Computational processing of raw sequencing data from barcoded samples

Barcode sequences were extracted from the resulting sequence files using custom scripts. Only barcodes with a minimum base quality of 20 that matched the constant regions in the 27 nucleotide barcode sequence in both the forward and reverse direction (±3 mismatches) were retrieved. Barcode sequences were identified from the raw sequence reads using the flanking known viral vector sequences in the forward (5'-ACAAGTAAAGCGGCCAACTCGAGAAGTAA-3') and reverse (5'-

CGAGCTCGAATTTGATCAGTCGACCCCGGGCTCACTGGCCGTCGTTTTACA-3') directions. From this a list of unique barcode sequences with corresponding read abundance was generated, and the read values corresponding to the spiked-in controls were used for defining a threshold and estimating clone size.

2.2.10 Filtering and thresholding approach

Using the list of unique barcode sequences, all the single, double, and triple mismatches were combined sequentially (as noted above), and the sum of the read abundance taken for those barcodes that were grouped. The spiked-in controls were then retrieved. We subsequently used these spiked-in control values to eliminate outlier amplicon libraries of poor quality that demonstrated poor correlation with the known input cell number.

2.2.11 Southern blot analysis

Southern blot analysis was performed using standard techniques (Sambrook, 1989). Briefly, DNA extracted from individual clones were digested with *EcoR*I and *Kpn*I in separate reactions, the digested DNA electrophoresed on a 1% agarose gel, transferred to a Zetaprobe membrane (Bio-rad), incubated with a ^{32}P-labeled probe recognizing the GFP reporter gene, and imaged using a phospho-imager (Molecular Dynamics).

2.3 Results

2.3.1 Design and construction of a high-complexity lentiviral-based barcode library compatible with MPS platforms

A non-coding DNA 'barcode' oligonucleotide was adapted from Gerrits et al. and custom ordered from Invitrogen. The barcode sequence (Figure 2.1) was 27 bp long and contained 5 pairs of degenerate nucleotides (N or S) separated by 4 sets of 3 constant nucleotides. The 27 bp barcode was flanked by a short stretch of constant nucleotides and 5' *Xho*I and 3' *Xma*I restriction sites. The entire oligonucleotide sequence was designed to be 54 bp, which can be covered within a single forward, or reverse read, on either an Illumina HiSeq or MiSeq platform.

The 5' ends of the forward and reverse oligonucleotides were phosphorylated to allow for direct orientation-specific ligation into a linearized vector. These oligonucleotides were dissolved separately in annealing buffer to reach a concentration of 50 μM, and then combined in an equimolar ratio. The resulting mixture was heated to 95°C for 5 minutes and slowly cooled to room temperature. Annealed barcode oligonucleotides were resolved from unannealed strands (54 bp), or longer concatamers (>200 bp) using a polyacrylamide gel (Figure 2.2). The appropriate band, approximately 108 bp in size was then isolated.

Insertion of the barcode oligonucleotide sequences into the MPG lentiviral vector (Figure 2.3) required that the vector be digested efficiently with both *Xho*I and *Xma*I enzymes, as incomplete digestion would have resulted in a re-circularized vector without a barcode insert. To determine the efficiency of enzyme activity, the MPG vector was first digested with each enzyme separately (Figure 2.4). Digestion with *Xho*I produced a

single 7.5 kb band consistent with the length of the vector. However, *Xma*I produced both the 7.5 kb band and a high molecular weight (>40,000 kb) band, consistent with a circular vector, indicating this enzyme was not as efficient. Thus, the MPG vector was digested sequentially, first with *Xma*I such that the linearized vector could be gel purified from the undigested vector. Then the purified product was secondarily digested with *Xho*I.

Prior to preparing a ligation reaction for insertion of the PAGE-purified annealed barcode oligonucleotides into the double digested MPG vector, the purity of both samples was analyzed using the Agilent Bioanalyzer (Figure 2.5). The digitally constructed electropherogram results indicated high purity of both samples, with no detectable impurities (as would have been evidenced by molecular weights not consistent with the desired products). The double digested vector was then ligated with the annealed barcode oligonucleotides purified previously. In addition, two ligation controls were included. The first control contained only double digested vector without the barcode oligonucleotides and no T4 DNA ligase, to enable any undigested/circular vector to be detected. The second control contained the double digested vector without the barcode oligonucleotides, but with the addition of T4 DNA ligase, to enable any single digested/linear vector to be detected.

The ligation reactions contained 10 ng of the double digested vector with the PAGE-purified barcode oligonucleotides in a 1:3 stoichiometric ratio, for a final reaction volume of 10 μl. From each of these ligation reactions, 1 μl was used for bacterial transformation by electroporation of electro-competent E. coli cells to examine the efficiency of ligation and the presence of background single or undigested MPG vector. Following a one-hour incubation at 37^{o}C in 500 μl SOC media, the transformed cultures

were plated at various densities and colony forming units (CFUs) per μl of barcode vector library were calculated (Table 2.1).

The CFU results suggested that non-recombinant clones were present in 10% of the CFU population within the 9 μl barcode vector library. Also, the remaining 9 μl of the ligation reaction containing the barcode oligonucleotide insert was estimated to contain approximately 125,000 unique barcodes from 125,000 individual CFUs, assuming a barcode library diversity close to 100%. The remaining 9 μl were then used to transform E.coli from which approximately 10^6 bacterial colonies were obtained, suggesting that alterations to the plating density improved the total CFU yield.

Construction of the barcoded plasmid library was thus designed to simultaneously maximize the number of unique barcodes as well as minimize the number of non-recombinant barcode clones. These characteristics of the barcode library were then validated by sequencing as described below.

2.3.2 Characterization of library complexity

To determine the library complexity and the background frequency of non-recombinant clones, Sanger sequencing was performed on 141 individually picked transformed bacterial colonies of the original 10^6. Each of these 141 colonies was found to contain a single unique barcode insert in the correct orientation. This suggests that the CFU estimate from Table 2.1 provided an over-estimation of the frequency of non-recombinant clones, and that in the ligation reaction containing the barcode insert, recombination to incorporate the barcode insert was highly efficient.

To obtain a more comprehensive estimate of the library complexity, MPS was performed on the barcoded plasmid library in triplicate. The plasmid library was obtained from the large-scale bacterial transformation that resulted in approximately 10^6 bacterial colonies that were then pooled together and homogenized prior to plasmid purification. From the raw MPS data obtained on the Illumina HiSeq, the data was processed without mismatch groupings. This revealed a diversity of ~$2x10^5$ unique barcodes (Figure 2.6). The average number of sequence reads for each unique barcode was 13 with a standard deviation of 4, indicating that there was no systemic bias in read abundance such that a subset of barcodes were substantially over or under-represented.

To further test for bias in barcode representation in the constructed library, the barcode plasmid library was incorporated into the MPG lentiviral vector from which an infectious virus supernatant was created. MPS of genomic DNA extracted from FACS-purified transduced (GFP^+) normal primary human mammary basal epithelial cells revealed that the read abundances for the barcodes identified were contained within a 2-fold range, with an average of 624 reads and a standard deviation of 157 reads (Figure 2.7).

Another potential source of bias that may influence PCR efficiency due to differences in melting and annealing temperatures is the G-C content of the variable regions within the barcode sequence. Of the 12 variable nucleotides, 2 are defined as "S" (IUPAC nomenclature) meaning they can be either a G or C. However, the remaining 10 defined as "N" can be occupied by any of the 4 nucleotides. Three different samples were analyzed for their G-C content prior to being analyzed by MPS: (i) the barcoded plasmid library (containing ~$2x10^5$ unique barcodes), (ii) transduced test cells (shown in Figure

2.7), and (iii) cells produced in the cleared fat pad of a recipient mouse that had been transplanted 7 weeks previously with mouse mammary epithelial cells transduced with the barcoded lentiviral library. The G-C content of all three samples was similar, with the mode of the distribution at 4 G-C (4 of the possible 10 variable nucleotides were occupied by either a G or C). Moreover, from the transduced test cells, the G-C content was further analyzed as it related to the read abundance for the unique barcodes within this sample, and the mode of the distribution was consistently at 4 G-C, regardless of read abundance. This suggests that the G-C content is slightly skewed toward more A or T nucleotides, which was then corrected once the 2 "S" nucleotide positions were also considered.

Therefore, the barcode library was shown to contain ~$2x10^5$ unique barcodes without substantial bias of barcode representation in the plasmid library, nor after being packaged into a library of lentiviral vectors. Furthermore, the G-C ratio within the variable regions of the barcode design is consistent with random nucleotide incorporation.

2.3.3 The "spiked-in" method for assessing the cell content of barcoded clones

To normalize for variables inherent in extracting genomic DNA and/or in the construction of molecular libraries for sequencing, we added a set of control samples to each experimental sample to serve as an internal calibration (Figure 2.9). These "spiked-in" controls consisted of defined numbers of cells containing known barcodes that covered the expected range of experimental clone sizes (Table 2.3). 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 the spiked-in controls in its respective indexed library (Table 2.4). For a total of 80 sets of spiked-in controls from 3 experimental datasets, 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 2.10). The first two datasets were acquired on the Illumina HiSeq platform and these yielded a polynomial regression with a standard deviation (SD) of <3 cells across all groups of spiked in controls. Therefore, we used a single-cell threshold (the fractional representation after normalization corresponding to a single cell) to discriminate between true barcode clones and those attributable to background/noise from sequencing or library construction. In the third MPS run conducted on the MiSeq platform, an increased sensitivity and reduced specificity was observed (revealed by a larger SD, <44 cells, and wider 95% CI across all groups of spiked-in controls, Table 2.5).

Application of these thresholds to virtual datasets from each of the 3 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, Table 2.6), the 100-cell clones were detected with 55-100% sensitivity, and the 20-cell clones with 18-67% sensitivity (Table 2.6). The specificity of detecting "true" clones was shown to be >99% (Table 2.7), based on finding only one of the 2,687 false positive barcodes in the 28 control libraries to be above the defined thresholds.

Reproducibility of clone detection was determined from an analysis of 7 pairs of replicate datasets obtained from the DNA of transduced mouse mammary epithelial cells that had undergone clonal expansion *in vivo* followed by one week of further

amplification of the cells *in vitro* (described in Chapter 3). Paired datasets were generated by splitting each of the 7 DNA extracts into 2 equal fractions that were then individually subjected to library construction and sequencing. Of the 113 clones identified in these 14 datasets, 88 were detected in both replicates of each pair (Table 2.8) and yielded clone size values of 26 to 12,235 cells. Clones detected in only one replicate were generally smaller (mean clone size of 34 cells and SD of 31 cells) than clones detected in both (mean clone size of 1091 cells and SD of 2,091 cells). These results are consistent with an increased likelihood of uneven sampling for clones containing <100 cells (Figure 2.11).

Previously described vector-based genomic barcoding strategies have relied on statistical modeling to determine conditions for lentiviral transduction that minimize the incidence of multiple barcode integrations by reducing the transduction efficiency to ≤30%. We mimicked this approach using a slightly different transduction protocol and then examined this issue directly by determining the 95% CI of the sizes estimated for each of our known spiked-in control samples. From the correlations obtained (R^2 = 0.94, R^2 = 0.86, and R^2 = 0.70 for the 3 MPS runs, respectively; Table 2.5), we calculated the 95% CI associated with each individual value, and grouped clones to eliminate redundant barcodes (likely derived from a single cell) where the 95% CI overlapped (Table 2.9). To measure directly the frequency of cells that would contain more than one integrated barcode under the conditions used for the primary cells, we analyzed clonally expanded single transduced 184-hTERT cells and then examined the number of viral integration sites in each clone by Southern blotting. Analysis of 23 such clones showed that 21 (~90%) contained a single integrant (Figure 2.12). This represents a lower frequency than

identified by overlapping 95% CI and indicates that a CI-based methodology is effective at removing multiple barcode integration events, but increases the frequency of true unique barcode clones being identified as a false negative.

Therefore, the use of spiked-in controls for calibration of sequence read abundances between samples allows for detection of barcode clones with a high sensitivity, specificity and reproducibility (for clones containing >100 cells). Furthermore, application of a 95% confidence interval following calculation of absolute clone size allows for rigorous elimination of likely redundant barcode clones.

2.4 Discussion

Cellular barcoding provides a high-resolution and high-throughput approach to analyzing the clonal outputs of single cells with varying growth and differentiation potentials (Lu et al., 2011; Naik et al., 2013). However, the reliability of this method is dependent on the complexity of the barcode library being used. The barcode library must be sufficiently complex and the representation of unique barcodes must be without significant bias, such that the probability of two clones integrated with the same barcode is reasonably low. Here, we initially validated the complexity of our barcode library by analyzing 141 bacterial colonies by Sanger sequencing. The finding that all 141 colonies contained a unique barcode indicated that the library complexity was indeed high. Furthermore, we did not detect a single non-recombinant clone, indicating that cellular clones marked with a lentiviral vector without a barcode sequence would be <1%. However, a more definitive estimate of the barcode library complexity was accomplished with MPS. The MPS results suggested that at least $2x10^5$ unique barcodes were present in the library, without significant over or under-representation of any subset. The probability of two clones

being marked with the same barcode could thus be estimated to be approximately $5x10^{-4}$ (0.05%) when simultaneously tracking 100 clones, and $5x10^{-3}$ (0.5%) when tracking 1,000 clones, which is reasonable for most clonal tracking experiments on normal and transformed cell populations.

The three sets of control libraries sequenced on either the HiSeq or MiSeq platforms provided specific measurements of clone detection sensitivity as a function of clone size, and confirmed the expectation that larger clones (approximately 100 cells or larger) would be consistently detected, but smaller clones not. Interestingly, there were several significant differences observed between the experiments analyzed on the HiSeq as compared to the MiSeq. Although the MiSeq does not have the capability of producing as many sequence reads as does the HiSeq (several billion paired-end reads compared to tens of millions), the sequencing chemistry differs such that the amount of phiX library required to spike-in prior to sequencing in order to increase the read complexity in regions where the barcode sequence is constant, was less than for the HiSeq (7% compared to 40% by mole). This increased the efficiency of read coverage for analysis of the intended barcode sequences. Although the reported sensitivity and specificity of clone detection was similar between the MPS runs on the HiSeq compared to MiSeq, these were actually different if clone detection was considered prior to application of a threshold to discriminate between true and false barcode clones. In the HiSeq runs, we observed a greater separation between the spiked-in controls compared to the noise (false barcodes), such that the single-cell threshold applied reliably separated the true barcode clones with the false ones, resulting in a specificity of >99%. However, for the MiSeq run, all of the spiked-in controls, even those consisting of only 10 cells, were detected in

the raw sequencing data generated. This suggested that the MiSeq had an increased sensitivity of clone detection. However, these were detected in the same range of read abundances where there were many false barcodes. In order to discriminate between true and false barcodes a threshold was defined at the read abundance equivalent of 20 cells, such that the specificity was 100% but with a significantly reduced sensitivity for detecting the 10 and 20 cell spiked-in controls.

The issue of how to discriminate between true and false barcode clones remains a very important issue in the field, especially when clones may remain initially "quiescent" for extended periods. As a result they are initially undetectable or may contain such few cells they may be in the range of noise. Our approach to the analysis of cellular barcoding data provides a conservative approach to detecting barcode clones, and thus has a very high specificity of barcode detection (>99%, and low frequency of detecting false barcodes). Furthermore, by calculating sensitivity and reproducibility of clone detection, we can provide a measure of confidence in the results derived important to interpreting the underlying biology. The use of spiked-in controls also serves as a standard by which different sequencing platforms, experiments, and methodologies can be compared so that variations in the methods used in the field can be objectively assessed.

2.5 Figures & tables

5’-Phos-**TCGAG AAGTAA NN ATC NN GAT SS AAA NN GGT NN AAC NN TGTAAAACGACGGCCAGTGAG C**-3’
3’-**C TTCATT NN TAG NN CTA SS TTT NN CCA NN TTG NN ACATTTTGCTGCCGGTCACTC GGGCC**-Phos-5’

Figure 2.1 Barcode oligonucleotide sequence

The 27 bp barcode sequence, shown in blue, was designed similarly to that reported by Gerrits et al. The 5’ and 3’ ends shown in red were designed with overhangs compatible with the *XhoI* and *XmaI* restriction sites, respectively, and were phosphorylated for direct ligation into the MNDU3-PGK-GFP lentiviral vector containing the same restriction sites downstream of the *GFP* reporter gene.

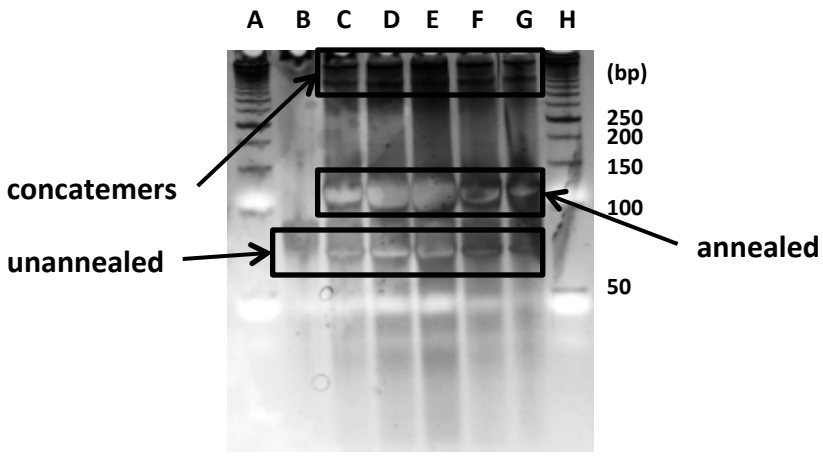

Figure 2.2 PAGE purification of the barcode oligonucleotides

A 10% polyacrylamide gel was used to purify annealed (~108 bp) barcode oligonucleotides from unannealed (~54 bp) barcode oligonucleotides and concatamers (>200 bp). The lanes were loaded with a 50 bp DNA ladder (A and H), unannealed oligonucleotides as a reference (B), and annealed oligonucleotides (C to G).

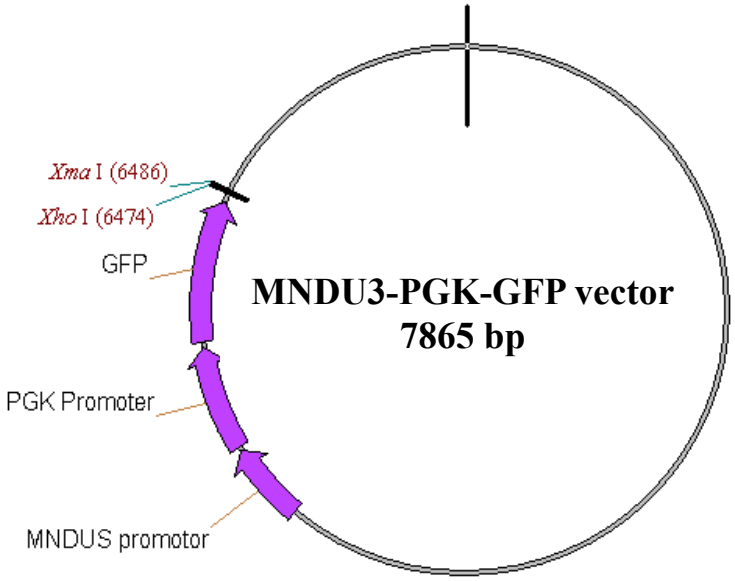

Figure 2.3 Map of the MNDU3-PGK-GFP lentiviral vector (generated from Vector NTI [Life Technologies])

The *XhoI* and *XmaI* restriction sites used for direct ligation of the barcode oligonucleotide sequence are located downstream of the *GFP* reporter gene and upstream of the 3' LTR.

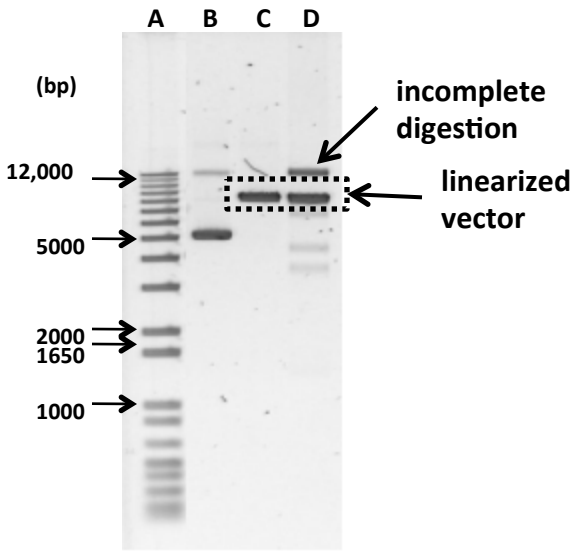

Figure 2.4 Enzyme digestion efficiency of the MPG vector

A 0.7% agarose-TBE gel was used to examine the length of products after MPG vector digestion with restriction enzymes *Xho*I and *Xma*I separately and together. The lanes were loaded with a 1 kb plus DNA ladder (A), undigested circular MPG vector as a reference (B), and *Xho*I and *Xma*I-digested MPG vector (C and D, respectively).

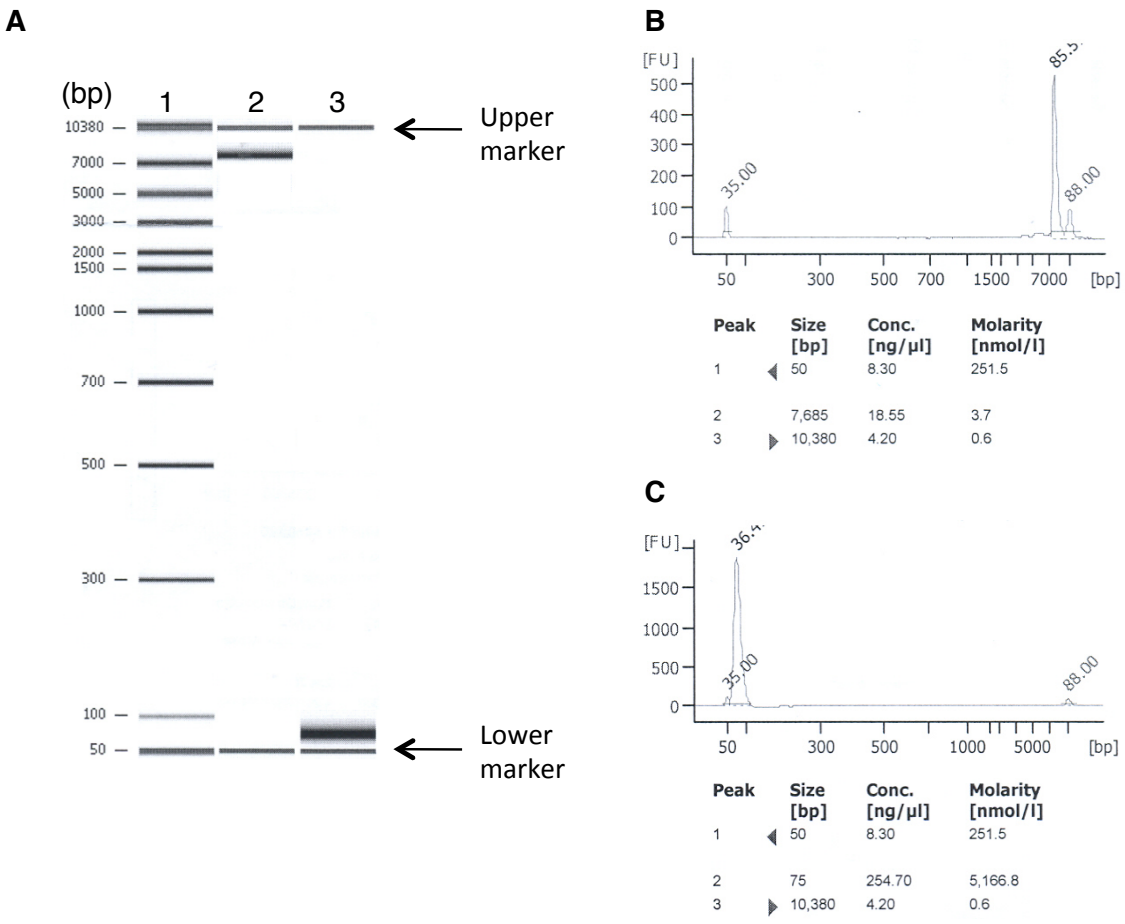

Figure 2.5 Barcode oligonucleotide and double-digested plasmid purity analysis

Prior to direct ligation and bacterial transformation, the annealed and PAGE-purified barcode oligonucleotides and the sequentially double-digested plasmid were analyzed for sample purity using a Bioanalyzer (Agilent). A digitally-constructed DNA gel (A) and electropherogram results are shown individually for lanes 2 and 3 (B and C, respectively). The units in B and C specify units of time (seconds, indicated for each peak) for migration of the band, and also the estimated size and concentration of the peaks observed.

Table 2.1 Calculation of CFUs from the constructed barcode plasmid library

Sample	Volume from 500 µL culture (µL)	No. colonies /plate	CFU/µL of ligation reaction	Total CFU in remaining 9 µL pool	Frequency of total CFUs
Vector + barcode	10	276	13800	124200	0.908
Control: undigested vector	50	54	540	4860	0.036
Control: singly digested vector	50	85	850	7650	0.056

Table 2.2 Confirmed barcodes by Sanger sequencing

Colony	Sequence
1	CTCGAGAAGAACAATCCCGATTGGAAACGGGTACAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
2	CTCGAGAAGTAAAAATCAAGATGGAAAGTGGTTAAACAATGTAAAACGACGGCCAGTGAGCCCGGG
3	CTCGAGAAGTAAAAATCACGATGGAAAGAGGTGTAACCCTGTAAAACGACGGCCAGTGAGCCCGGG
4	CTCGAGAAGTAAAAATCCAGATCCAAATTGGTCTAACCATGTAAAACGACGGCCAGTGAGCCCGGG
5	CTCGAGAAGTAAAAATCTAGATCCAAATTGGTGAAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
6	CTCGAGAAGTAAAAATCTAGATGCAAATGGGTCGAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
7	CTCGAGAAGTAAAAATCTGGATCGAAAGAGGTAAAACGCTGTAAAACGACGGCCAGTGAGCCCGGG
8	CTCGAGAAGTAAAAATCTTGATCGAAATCGGTCCAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
9	CTCGAGAAGTAAACATCAAGATCCAAAAAGGTCAAACTATGTAAAACGACGGCCAGTGAGCCCGGG
10	CTCGAGAAGTAAACATCAGGATCCAAACAGGTTCAACCATGTAAAACGACGGCCAGTGAGCCCGGG
11	CTCGAGAAGTAAACATCCCGATCCAAAGCGGTCCAACACTGTAAAACGACGGCCAGTGAGCCCGGG
12	CTCGAGAAGTAAACATCCCGATCGAAAGCGGTCCAACGATGTAAAACGACGGCCAGTGAGCCCGGG
13	CTCGAGAAGTAAACATCCGGATGGAAATGGGTTAAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
14	CTCGAGAAGTAAACATCCTGATCCAAAATGGTCTAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
15	CTCGAGAAGTAAACATCCTGATCCAAAATGGTGCAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
16	CTCGAGAAGTAAACATCTAGATGCAAAGAGGTAAAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
17	CTCGAGAAGTAAACATCTGGATGCAAATGGGTCTAACGCTGTAAAACGACGGCCAGTGAGCCCGGG
18	CTCGAGAAGTAAACATCTTGATCCAAAAGGGTATAACACTGTAAAACGACGGCCAGTGAGCCCGGG
19	CTCGAGAAGTAAAGATCATGATCCAAAATGGTTGAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
20	CTCGAGAAGTAAAGATCCAGATGCAAATTGGTGTAACATTGTAAAACGACGGCCAGTGAGCCCGGG
21	CTCGAGAAGTAAAGATCGCGATCCAAAGTGGTATAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
22	CTCGAGAAGTAAAGATCGGGATCGAAACAGGTGGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
23	CTCGAGAAGTAAAGATCTTGATCGAAAGTGGTATAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
24	CTCGAGAAGTAAATATCAAGATCGAAACTGGTTGAACGATGTAAAACGACGGCCAGTGAGCCCGGG
25	CTCGAGAAGTAAATATCAAGATCGAAATGGGTGTAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
26	CTCGAGAAGTAAATATCAAGATGCAAATTGGTATAACATTGTAAAACGACGGCCAGTGAGCCCGGG
27	CTCGAGAAGTAAATATCAGGATGCAAACAGGTCGAACTATGTAAAACGACGGCCAGTGAGCCCGGG
28	CTCGAGAAGTAAATATCCGGATGGAAAGGGGTCGAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
29	CTCGAGAAGTAAATATCCTGATCGAAATGGGTCGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
30	CTCGAGAAGTAAATATCGAGATGGAAATCGGTAGAACGCTGTAAAACGACGGCCAGTGAGCCCGGG
31	CTCGAGAAGTAAATATCTTGATGCAAAAGGGTAGAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
32	CTCGAGAAGTAACAATCATGATGCAAAGCGGTCAAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
33	CTCGAGAAGTAACAATCATGATGCAAATTGGTAGAACTATGTAAAACGACGGCCAGTGAGCCCGGG
34	CTCGAGAAGTAACAATCCAGATCGAAATTGGTACAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
35	CTCGAGAAGTAACAATCCAGATCGAAATTGGTACAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
36	CTCGAGAAGTAACAATCCCGATCCAAAGAGGTTTAACCATGTAAAACGACGGCCAGTGAGCCCGGG
37	CTCGAGAAGTAACAATCCCGATCGAAACCGGTCTAACCCTGTAAAACGACGGCCAGTGAGCCCGGG
38	CTCGAGAAGTAACAATCCGGATCCAAACGGGTTTAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
39	CTCGAGAAGTAACAATCCTGATCCAAAAAGGTATAACTATGTAAAACGACGGCCAGTGAGCCCGGG

(Table continued on subsequent page…)

Colony	Sequence
40	CTCGAGAAGTAACAATCCTGATGCAAAGTGGTGTAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
41	CTCGAGAAGTAACAATCGAGATCCAAACCGGTGGAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
42	CTCGAGAAGTAACAATCGCGATCCAAATCGGTACAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
43	CTCGAGAAGTAACAATCGGGATGGAAAGGGGTGGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
44	CTCGAGAAGTAACCATCACGATCGAAAACGGTAGAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
45	CTCGAGAAGTAACCATCAGGATCCAAAACGGTGGAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
46	CTCGAGAAGTAACCATCCCGATCCAAACCGGTCCAACCCTGTAAAACGACGGCCAGTGAGCCCGGG
47	CTCGAGAAGTAACCATCCCGATCCAAACCGGTTCAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
48	CTCGAGAAGTAACCATCCCGATCGAAAAAGGTACAACAATGTAAAACGACGGCCAGTGAGCCCGGG
49	CTCGAGAAGTAACCATCCCGATCGAAACCGGTCCAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
50	CTCGAGAAGTAACCATCCGGATGGAAAGGGGTGGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
51	CTCGAGAAGTAACCATCCGGATGGAAAGGGGTGGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
52	CTCGAGAAGTAACCATCCTGATCCAAACCGGTACAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
53	CTCGAGAAGTAACCATCCTGATCGAAACTGGTCCCACACTGTAAAACGACGGCCAGAGAGCCCGGG
54	CTCGAGAAGTAACCATCCTGATCGAAAGTGGTGCAACCCTGTAAAACGACGGCCAGTGAGCCCGGG
55	CTCGAGAAGTAACCATCCTGATCGAAATAGGTTTAACCATGTAAAACGACGGCCAGTGAGCCCGGG
56	CTCGAGAAGTAACCATCGAGATGGAAACCGGTGCAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
57	CTCGAGAAGTAACCATCGCGATCGAAAGGGGTGAAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
58	CTCGAGAAGTAACCATCGGGATCGAAAATGGTGTAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
59	CTCGAGAAGTAACCATCGGGATGGAAAGGGGTGAAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
60	CTCGAGAAGTAACCATCTAGATGGAAATTGGTTCAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
61	CTCGAGAAGTAACCATCTTGATCCAAAACGGTGCAACTATGTAAAACGATGGCCAGTGAGCCCGGG
62	CTCGAGAAGTAACGATCATGATCGAAAAGGGTAAAACTATGTAAAACGACGGCCAGTGAGCCCGGG
63	CTCGAGAAGTAACGATCATGATGCAAACTGGTTAAACAATGTAAAACGACGGCCAGTGAGCCCGGG
64	CTCGAGAAGTAACGATCCGGATGGAAACCGGTGCAACAATGTAAAACGACGGCCAGTGAGCCCGGG
65	CTCGAGAAGTAACGATCGCGATCCAAACCGGTTGAACAATGTAAAACGACGGCCAGTGAGCCCGGG
66	CTCGAGAAGTAACGATCTAGATCGAAATAGGTCAAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
67	CTCGAGAAGTAACGATCTAGATGGAAATTGGTATAACTATGTAAAACGACGGCCAGTGAGCCCGGG
68	CTCGAGAAGTAACTATCAAGATGGAAATGGGTATAACATTGTAAAACGACGGCCAGTGAGCCCGGG
69	CTCGAGAAGTAACTATCACGATGGAAAATGGTTCAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
70	CTCGAGAAGTAACTATCCCGATCCAAACTGGTTAAACACTGTAAAACGACGGCCAGTGAGCCCGGG
71	CTCGAGAAGTAACTATCCCGATGGAAAGGGGTTAAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
72	CTCGAGAAGTAACTATCGGGATGCAAAAAGGTTTAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
73	CTCGAGAAGTAACTATCGTGATCGAAATTGGTCTAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
74	CTCGAGAAGTAACTATCTGGATCGAAAACGGTGGAACACTGTAAAACGACGGCCAGTGAGCCCGGG
75	CTCGAGAAGTAAGAATCATGATGCAAAGAGGTCTAACTATGTAAAACGACGGCCAGTGAGCCCGGG
76	CTCGAGAAGTAAGAATCGTGATGGAAACTGCTTCAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
77	CTCGAGAAGTAAGAATCTAGATCGAAAGCGGTACAACAATGTAAAACGACGGCCAGTGAGCCCGGG
78	CTCGAGAAGTAAGAATCTTGATGGAAAGGGGTGCAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
79	CTCGAGAAGTAAGCATCATGATCCAAATAGGTAAAACATTGTAAAACGACGGCCAGTGAGCCCGGG
80	CTCGAGAAGTAAGCATCATGATCGAAAATGGTTTAACTATGTAAAACGACGGCCAGTGAGCCCGGG

(Table continued on subsequent page…)

Colony	Sequence
81	CTCGAGAAGTAAGCATCATGATCGAAAATGGTTTAACTATGTAAAACGACGGCCAGTGAGCCCGGG
82	CTCGAGAAGTAAGCATCCCGATGCAAACCGGTCAAACATTGTAAAACGACGGCCAGTGAGCCCGGG
83	CTCGAGAAGTAAGCATCCGGATGCAAAATGGTTAAACGATGTAAAACGACGGCCAGTGAGCCCGGG
84	CTCGAGAAGTAAGCATCGGGATGGAAACGGGTGCAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
85	CTCGAGAAGTAAGCATCGTGATGCAAAGGGGTGCAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
86	CTCGAGAAGTAAGCATCTGGATCCAAATTGGTATAACAATGTAAAACGACGGCCAGTGAGCCCGGG
87	CTCGAGAAGTAAGCATCTTGATGCAAAAGGGTGTAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
88	CTCGAGAAGTAAGCATCTTGATGCAAAATGGTCAAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
89	CTCGAGAAGTAAGGATCCCGATGGAAAGCGGTGAAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
90	CTCGAGAAGTAAGGATCCGGATGCAAAGGGGTAGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
91	CTCGAGAAGTAAGGATCGCGATGGAAACGGGTGGAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
92	CTCGAGAAGTAAGGATCGGGATCGAAACTGGTCGAACGATGTAAAACGACGGCCAGTGAGCCCGGG
93	CTCGAGAAGTAAGGATCGGGATGGAAACGGGTGGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
94	CTCGAGAAGTAAGGATCGGGATGGAAAGGGGTGGAACGGTGTAAAACGACGGCCAGTGAGCCCGGG
95	CTCGAGAAGTAAGGATCGTGATCGAAATAGGTTTAACAATGTAAAACGACGGCCAGTGAGCCCGGG
96	CTCGAGAAGTAAGGATCTAGATCCAAATGGGTTAAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
97	CTCGAGAAGTAAGTATCAGGATCCAAAAAGGTCCAACAATGTAAAACGACGGCCAGTGAGCCCGGG
98	CTCGAGAAGTAAGTATCGAGATGGAAAGAGGTATAACGATGTAAAACGACGGCCAGTGAGCCCGGG
99	CTCGAGAAGTAAGTATCTTGATCGAAAGTGGTCAAACATTGTAAAACGACGGCCAGTGAGCCCGGG
100	CTCGAGAAGTAATAATCAAGATGCAAAGTGGTGAAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
101	CTCGAGAAGTAATAATCACGATGCAAACAGGTCAAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
102	CTCGAGAAGTAATAATCCAGATCCAAAAGGGTGCAACCATGTAAAACGACGGCCAGTGAGCCCGGG
103	CTCGAGAAGTAATAATCCCGATGGAAACGGGTCAAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
104	CTCGAGAAGTAATAATCCGGATCGAAACTGGTTGAACAATGTAAAACGACGGCCAGTGAGCCCGGG
105	CTCGAGAAGTAATAATCGCGATCCAAACCGGTCTAACATTGTAAAACGACGGCCAGTGAGCCCGGG
106	CTCGAGAAGTAATAATCGGGATGCAAACGGCTGAAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
107	CTCGAGAAGTAATAATCGTGATCGAAACTGGTACAACTATGTAAAACGACGGCCAGTGAGCCCGGG
108	CTCGAGAAGTAATAATCTGGATGCAAAGGGGTATAACACTGTAAAACGACGGCCAGTGAGCCCGGG
109	CTCGAGAAGTAATAATCTGGATGCAAATAGGTTTAACAATGTAAAACGACGGCCAGTGAGCCCGGG
110	CTCGAGAAGTAATAATCTTGATCGAAACCGGTCAAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
111	CTCGAGAAGTAATAATCTTGATCGAAAGTGGTTGAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
112	CTCGAGAAGTAATCATCACGATGCAAACCGGTGCAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
113	CTCGAGAAGTAATCATCCCGATCCAAATGGGTCCAACGCTGTAAAACGACGGCCAGTGAGCCCGGG
114	CTCGAGAAGTAATCATCCCGATGCAAAGGGGTGTAACATTGTAAAACGACGGCCAGTGAGCCCGGG
115	CTCGAGAAGTAATCATCCGGATGGAAAGTGGTCCAACAGTGTAAAACGACGGCCAGTGAGCCCGGG
116	CTCGAGAAGTAATCATCGAGATGGAAAAAGGTATAACACTGTAAAACGACGGCCAGTGAGCCCGGG
117	CTCGAGAAGTAATCATCGTGATCGAAAAGGGTTTAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
118	CTCGAGAAGTAATCATCGTGATCGAAATAGGTGGAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
119	CTCGAGAAGTAATCATCTAGATGCAAAGCGGTTAAACATTGTAAAACGACGGCCAGTGAGCCCGGG
120	CTCGAGAAGTAATCATCTCGATCCAAACCGGTTTAACCCTGTAAAACGACGGCCAGTGAGCCCGGG
121	CTCGAGAAGTAATCATCTCGATCCAAACCGGTTTAACCCTGTAAAACGACGGCCAGTGAGCCCGGG

(Table continued on subsequent page…)

Colony	Sequence
122	CTCGAGAAGTAATCATCTCGATCCAAAGAGGTACAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
123	CTCGAGAAGTAATCATCTCGATGCAAATAGGTTAAACAATGTAAAACGACGGCCAGTGAGCCCGGG
124	CTCGAGAAGTAATCATCTCGATGCAAATGGGTGAAACGCTGTAAAACGACGGCCAGTGAGCCCGGG
125	CTCGAGAAGTAATGATCAAGATCGAAATAGGTGTAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
126	CTCGAGAAGTAATGATCAAGATGCAAAGTGGTTTAACGTTGTAAAACGACGGCCAGTGAGCCCGGG
127	CTCGAGAAGTAATGATCCAGATCCAAACAGGTCGAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
128	CTCGAGAAGTAATGATCCCGATGCAAAGGGGTCAAACACTGTAAAACGACGGCCAGTGAGCCCGGG
129	CTCGAGAAGTAATGATCCCGATGGAAACAGGTACAACTCTGTAAAACGACGGCCAGTGAGCCCGGG
130	CTCGAGAAGTAATGATCGTGATCCAAAGTGGTATAACGATGTAAAACGACGGCCAGTGAGCCCGGG
131	CTCGAGAAGTAATGATCTAGATGCAAAATGGTTTAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
132	CTCGAGAAGTAATGATCTCGATCGAAAAGGGTGAAACGATGTAAAACGACGGCCAGTGAGCCCGGG
133	CTCGAGAAGTAATTATCAAGATGCAAAGAGGTTAAACATTGTAAAACGACGGCCAGTGAGCCCGGG
134	CTCGAGAAGTAATTATCAGGATCGAAATGGGTTTAACCTTGTAAAACGACGGCCAGTGAGCCCGGG
135	CTCGAGAAGTAATTATCAGGATGGAAATAGGTTTAACTTTGTAAAACGACGGCCAGTGAGCCCGGG
136	CTCGAGAAGTAATTATCATGATGGAAAATGGTATAACATTGTAAAACGACGGCCAGTGAGCCCGGG
137	CTCGAGAAGTAATTATCCGGATCCAAATAGGTTGAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
138	CTCGAGAAGTAATTATCCTGATGGAAAGAGGTAAAACCGTGTAAAACGACGGCCAGTGAGCCCGGG
139	CTCGAGAAGTAATTATCGCGATGCAAAGGGGTATAACTATGTAAAACGACGGCCAGTGAGCCCGGG
140	CTCGAGAAGTAATTATCTAGATGCAAAATGGTTAAACTGTGTAAAACGACGGCCAGTGAGCCCGGG
141	CTCGAGAAGTAATTATCTAGATGGAAAAAGGTTGAACGCTGTAAAACGACGGCCAGTGAGCCCGGG

The above table includes a list of the barcode identified from 141 hand-picked bacterial colonies. Afterward, however, the same barcode library was sequenced by MPS (Figure 2.6).

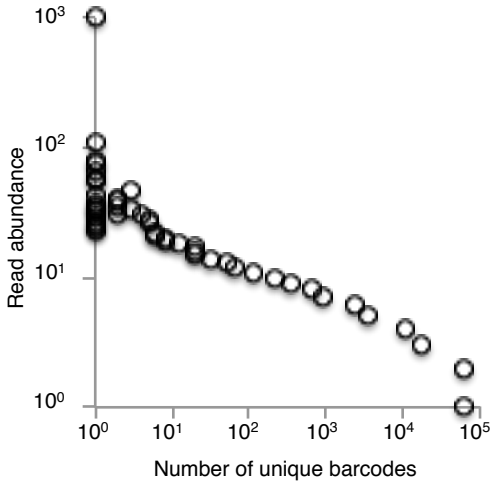

Figure 2.6 Barcode plasmid library diversity determined by MPS

Three experimental replicates of plasmid purified from a MaxiPrep of the barcode-containing MPG vector were sequenced and analyzed. Average read abundances were calculated where the same barcodes were identified in more than one replicate. From this list, the plasmid diversity is shown as the abundance in number of reads obtained from MPS (y-axis) as a function of the number of unique barcodes (x-axis). The total number of unique barcodes is ~$2x10^5$, and the mean number of reads and SD for each unique barcode is 13 ± 4.

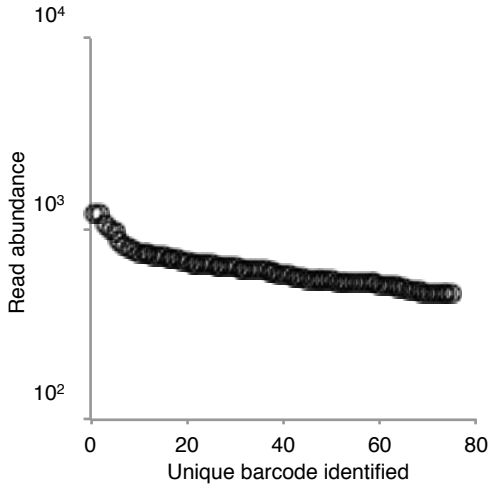

Figure 2.7 Analysis of barcode-transduced primary human mammary basal epithelial cells reveals no systematic biases by MPS

Normal primary human mammary basal epithelial cells were transduced with the barcoded lentiviral library and the transduced cells analyzed for their barcodes by MPS. The results show that all clones identified in the test population contained a narrow range of reads (459 to 1,189). The average number of sequence reads was 624 with a SD of 157 reads.

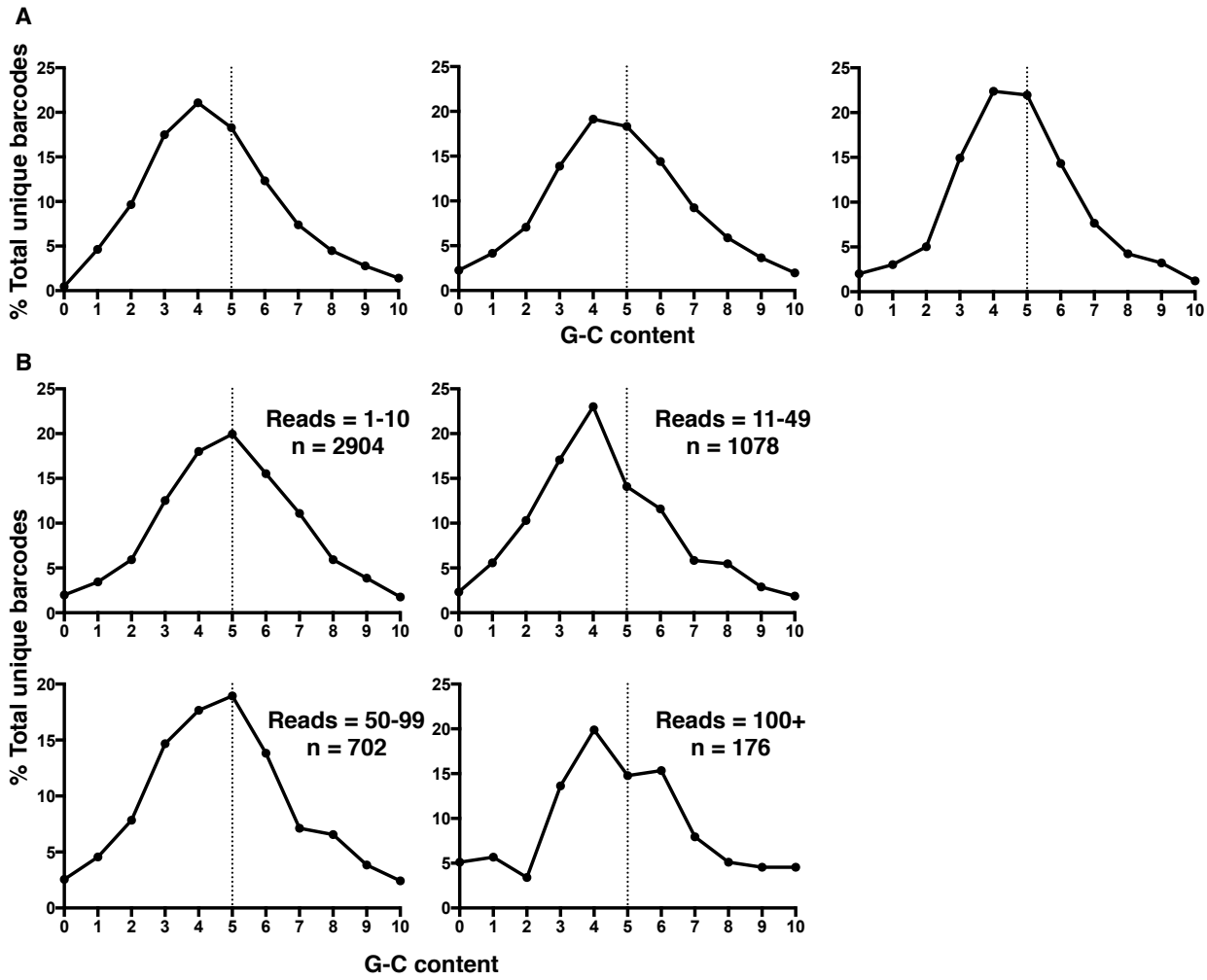

Figure 2.8 Analysis of G-C content in the barcoded plasmid library and infected cells

The G-C content of the variable regions ('N') in the barcode oligonucleotide is shown for the barcoded plasmid library (A, left plot), test transduced cells sequenced after 3 days (A, middle plot) and barcoded mouse mammary epithelial cells after 7 weeks *in vivo* (A, right plot), where 0 is no G or C nucleotides in any of the variable regions, and 10 is G or C nucleotides in all the variable regions. The y-axis represents the percentage of total unique barcodes with the corresponding G-C content shown on the x-axis. Also shown

(B) is a breakdown of the G-C content according to the read abundances for the unique barcodes in the test transduced cells (of which the middle plot of A represents the sum of the 4 plots in B).

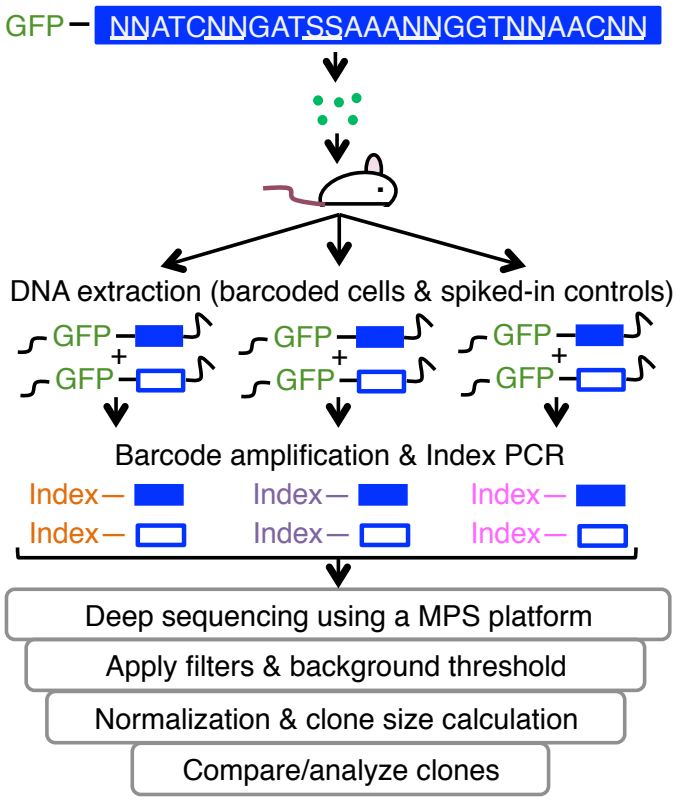

Figure 2.9 Experimental workflow for analysis of barcoded samples

Degenerate nucleotides for the variable regions in the barcode sequence are indicated with an N or S. At the end of the experiment, spiked-in control cells were added to each library, sequenced on a MPS platform, and custom scripts used to filter and analyze the data. The spiked-in controls were then used for normalization and clone size calculations.

Table 2.3 Spiked-in controls used for clone size calibrations

MPS run	Controls	Barcode sequence
1	500 cells	**TCATCCTGATGCAAATTGGTGTAACCC**
	100 cells	**AGATCTAGATGGAAAGGGGTCCAACCA**
	20 cells	**GCATCACGATGGAAATAGGTTGAACTT**
2	500 cells	**AGATCCTGATGCAAACGGGTCAAACAC**
	100 cells	**TAATCTCGATCGAAAATGGTAGAACTT**
	20 cells	**AGATCTAGATGGAAAGGGGTCCAACCA**
3	1000 cells	**AGATCCTGATGCAAACGGGTCAAACAC**
	500 cells	**GCATCACGATGGAAATAGGTTGAACTT**
	250 cells	**CTATCCTGATGCAAAAAGGTGAAACAG**
	100 cells	**AGATCTAGATGGAAAGGGGTCCAACCA**
	20 cells	**TCATCCTGATGCAAATTGGTGTAACCC**
	10 cells	**TAATCTCGATCGAAAATGGTAGAACTT**

Table 2.4 Fraction read representation of spiked-in control datasets

The number of sequence reads obtained for each of the clones within each of the libraries is shown for MPS run #1, 2 and 3. The fractional read value was used to normalize data between libraries, and remove variability in read representation due to differences in number of reads per library, and the number and size of unknown experimental clones in each library. The fractional read value is calculated as the number of reads divided by the sum of reads from the 500, 100, and 20 cell controls within that library.

Table 2.4

MPS #1:

Library	500 cell spike-in		100 cell spike-in		20 cell spike-in	
	Reads	Fractional read value	Reads	Fractional read value	Reads	Fractional read value
10126	328008	0.9338	20421	0.0581	2820	0.008
10127	397459	0.9271	25099	0.0585	6155	0.0144
10128	455992	0.9292	28861	0.0588	5881	0.012
10135	512180	0.9228	38795	0.0699	4080	0.0074
10136	627401	0.9402	33182	0.0497	6725	0.0101
10137	614739	0.92	46933	0.0702	6525	0.0098
10138	167401	0.9352	10645	0.0595	950	0.0053
10139	149489	0.8882	14663	0.0871	4154	0.0247
10140	14428	0.9585	482	0.032	142	0.0094
10141	989	0.9687	29	0.0284	3	0.0029
10143	131412	0.9381	7477	0.0534	1199	0.0086
10144	260166	0.9136	19822	0.0696	4775	0.0168
10145	29150	0.9311	1589	0.0508	568	0.0181
10146	3795	0.9652	105	0.0267	32	0.0081
10204	3265	0.9049	232	0.0643	111	0.0308
10205	35926	0.8949	3726	0.0928	494	0.0123
10206	6145	0.8862	600	0.0865	189	0.0273
10207	1135	0.9818	19	0.0164	2	0.0017
10208	2241	0.9693	53	0.0229	18	0.0078
10209	11007	0.9176	764	0.0637	224	0.0187
10210	8098	0.922	534	0.0608	151	0.0172
10211	4751	0.9798	89	0.0184	9	0.0019
10212	4085	0.9698	104	0.0247	23	0.0055
10213	13004	0.9255	851	0.0606	196	0.0139
10214	21433	0.8988	1985	0.0832	429	0.018
10215	4086	0.9801	74	0.0178	9	0.0022
10216	2740	0.9835	40	0.0144	6	0.0022

MPS #2:

Library	500 cell spike-in		100 cell spike-in		20 cell spike-in	
	Reads	Fractional read value	Reads	Fractional read value	Reads	Fractional read value
23779	208913	0.9186	9565	0.0421	8943	0.0393
23780	214487	0.9332	8576	0.0373	6770	0.0295
23781	249936	0.9017	14024	0.0506	13220	0.0477
23782	193583	0.9196	11838	0.0562	5093	0.0242
23783	72082	0.9123	3612	0.0457	3313	0.0419
23785	184038	0.9185	9321	0.0465	7005	0.0350
23786	214828	0.9115	10432	0.0443	10416	0.0442
23820	9343	0.9459	270	0.0273	264	0.0267
23822	2570	0.9561	84	0.0313	34	0.0126
23828	28595	0.9309	1113	0.0362	1010	0.0329
23833	2244	0.9606	54	0.0231	38	0.0163
23834	2106	0.9674	41	0.0188	30	0.0138
23835	1094	0.9622	24	0.0211	19	0.0167
23836	703	0.9262	34	0.0448	22	0.0290
23838	2454	0.9388	85	0.0325	75	0.0287
23839	693	0.9023	50	0.0651	25	0.0326
23845	837	0.8710	67	0.0697	57	0.0593
23847	835	0.9288	41	0.0456	23	0.0256

MPS #3:

Library	500 cell spike-in		100 cell spike-in		20 cell spike-in	
	Reads	Fractional read value	Reads	Fractional read value	Reads	Fractional read value
A1	40017	0.8442	6646	0.1402	737	0.0155
B1	45308	0.8402	8150	0.1511	466	0.0086
B2	42104	0.8351	7216	0.1431	1096	0.0217
C1	37855	0.8749	4959	0.1146	455	0.0105
C2	34892	0.8502	5946	0.1449	201	0.0049
D1	43376	0.8756	5841	0.1179	321	0.0065
D2	31176	0.7491	10440	0.2508	3	0.0001
E1	43583	0.8233	8765	0.1656	591	0.0112
F1	36647	0.8675	5232	0.1238	366	0.0087
G1	39081	0.8741	4525	0.1012	1105	0.0247
H1	45978	0.8432	7656	0.1404	891	0.0163
A3	3159	0.8566	467	0.1266	62	0.0168
A4	2793	0.8110	560	0.1626	91	0.0264
A5	3620	0.7051	1259	0.2452	255	0.0497
B3	1200	0.8202	246	0.1681	17	0.0116
B4	5505	0.8159	1186	0.1758	56	0.0083
B5	8315	0.8697	1171	0.1225	75	0.0078
C3	5105	0.8493	800	0.1331	106	0.0176
C4	3245	0.8278	661	0.1686	14	0.0036
C5	7587	0.8899	856	0.1004	83	0.0097
D3	9888	0.7602	2871	0.2207	248	0.0191
D4	2227	0.8912	264	0.1056	8	0.0032
D5	11646	0.8506	1748	0.1277	297	0.0217
E2	93	0.7815	23	0.1933	3	0.0252
E3	2101	0.7351	590	0.2064	167	0.0584
E4	2281	0.7662	489	0.1643	207	0.0695
F2	20	0.7407	5	0.1852	2	0.0741
F3	397	0.8340	65	0.1366	14	0.0294
F4	4894	0.8075	1082	0.1785	85	0.0140
G2	134	0.7976	34	0.2024	0	0
G3	397	0.8069	82	0.1667	13	0.0264
G4	57914	0.7790	16429	0.2210	0	0
H2	2348	0.8735	340	0.1265	0	0
H3	604	0.8666	93	0.1334	0	0
H4	22110	0.8420	4069	0.1550	79	0.0030

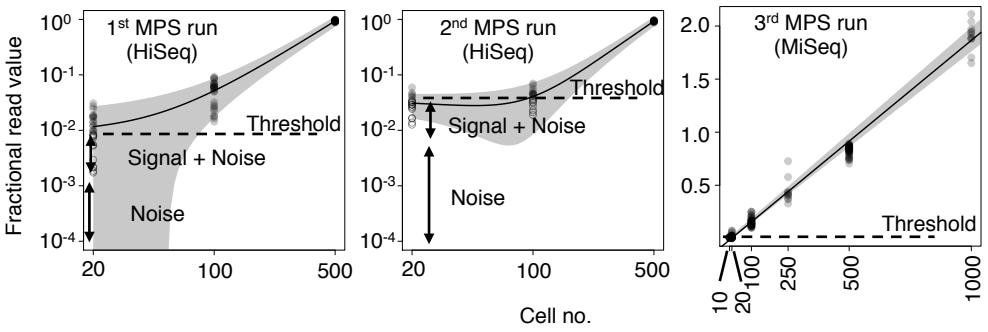

Figure 2.10 Regression analysis of fractional read representation versus cell number

The relationship between absolute cell number and the fractional read value (the normalized barcode read abundance value) for the MPS data from the three MPS runs performed are shown. The solid black line in each 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 a single cell and approximately 20 cells, for the first two MPS runs and the third MPS run, respectively.

Table 2.5 SD and 95% CI for three MPS runs

(A) MPS run #1

Control	Average fractional read value	Standard deviation	Lower 95% CI	Upper 95% CI	95% CI (+/-)
500 cells	0.9365	0.0300	0.9246	0.9484	0.0119
100 cells	0.0518	0.0239	0.0424	0.0613	0.0095
20 cells	0.0117	0.0078	0.0086	0.0148	0.0031

(B) MPS run #2

Control	Average fractional read value	Standard deviation	Lower 95% CI	Upper 95% CI	95% CI (+/-)
500 cells	0.9332	0.0286	0.9151	0.9514	0.0182
100 cells	0.0381	0.0163	0.0278	0.0485	0.0104
20 cells	0.0286	0.0141	0.0197	0.0376	0.0090

(C) MPS run #3

Control	Average fractional read value	Standard deviation	Lower 95% CI	Upper 95% CI	95% CI (+/-)
1000 cells	1.9038	0.2657	1.815	1.993	0.178
500 cells	0.8244	0.0476	0.8081	0.8408	0.0327
250 cells	0.4442	0.2253	0.3685	0.5199	0.1514
100 cells	0.1577	0.0393	0.1442	0.1712	0.027
20 cells	0.0178	0.0189	0.0114	0.0243	0.0130
10 cells	0.0078	0.0111	0.0040	0.0115	0.0074

From the above calculations, a correlation was derived relating the clone size (cell number) to the 95% CI. (*A*) For the 1st MPS dataset, y = 0.0027 ln (x) – 0.0045, where y is the 95% CI in units of fractional read value, and x is the clone size in cell number. (*B*)

For the 2nd MPS dataset, y = -0.009 ln (x) + 0.018. (*C*) For the 3rd MPS dataset, y = -0.1252x + 39.961.

Table 2.6 Sensitivity of barcode clone detection

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%)

For each of the 28 datasets containing on the spiked-in controls, the corresponding thresholds were applied and the sensitivity of barcode clone detection using the defined threshold calculated.

Table 2.7 Specificity of barcode clone detection

The calculations for specificity were performed using a set of 6, 11 and 11 controls for the 1st, 2nd and 3rd MPS runs, respectively. Each set of controls contain a 500, 100 and 20 cell control.

Table 2.7

MPS run	Control	False-positive clones	True negative clones	Specificity
1	1	0	134	100
	2	0	177	100
	3	0	87	100
	4	1	128	99
	5	0	177	100
	6	0	260	100
	Average			99.9
2	1	0	56	100
	2	0	181	100
	3	0	169	100
	4	0	111	100
	5	0	91	100
	6	0	136	100
	7	0	118	100
	8	0	98	100
	9	0	55	100
	10	0	234	100
	11	0	125	100
	Average			100
3	1	0	37	100
	2	0	19	100
	3	0	21	100
	4	0	22	100
	5	0	17	100
	6	0	29	100
	7	0	27	100
	8	0	16	100
	9	0	28	100
	10	0	31	100
	11	0	103	100
	Average			100

Table 2.8 Reproducibility of barcode clone detection

Paired replicate library #1

Replicate 1	Replicate 2
1226	1252
436	425
376	388
374	348
137	141
108	153
93	64
47	140
29	57

Paired replicate library #2

Replicate 1	Replicate 2
2931	534
1987	2357
1302	1512
1193	1497
749	850
667	826
612	853
589	743
425	486
391	461
300	362
272	351
268	385
249	263
202	249
89	167
ND	131
ND	62
ND	43
ND	30

Paired replicate library #3

Replicate 1	Replicate 2
3128	6266
2152	4191
913	1642
819	1453
556	1014
415	760
392	771
361	629
358	765
351	676
299	576
252	477
147	301
106	299
75	210
68	209
66	53
2	96
ND	151
ND	63
ND	48
ND	16
ND	13
ND	4

Paired replicate library #4

Replicate 1	Replicate 2
9313	9399
2933	2913
340	347
213	243
82	10
68	165
30	ND
20	31
ND	170
ND	103
ND	20
ND	5

Paired replicate library #5

Replicate 1	Replicate 2
12734	9873
3035	2046
439	305
436	473
117	45
89	495
38	ND
23	ND
ND	195
ND	113
ND	7

Paired replicate #6

Replicate 1	Replicate 2
5230	5928
2061	2429
1049	1233
1011	1135
712	769
641	753
595	814
578	740
511	799
489	484
456	607
432	452
407	493
393	459
204	257
96	209
72	ND
ND	40

Paired replicate #7

Replicate 1	Replicate 2
11000	13469
1155	1517
1147	1391
1115	1547
769	1025
717	857
588	631
488	754
471	615
435	587
418	558
243	426
224	ND
117	168
110	92
99	41
58	ND
35	51
ND	55

Each row indicates a unique clone identified in either one or both paired replicate libraries. ND indicates a clone that was not detected.

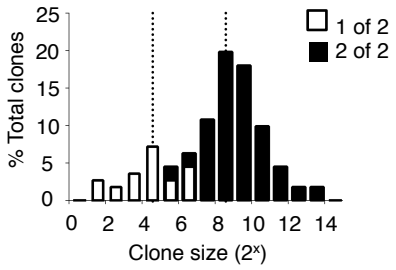

Figure 2.11 Size distribution plot of reproducibly detected clones

Reproducibility of clone detection is shown with respect to clone size in 7 replicate libraries (Table 2.8), the average of which is depicted here. Solid bars indicate clones detected in both libraries and open bars indicate clones detected in only 1 of the 2. Dotted lines indicate the mode of each size distribution.

Table 2.9 Unique barcodes identified and number of barcodes merged due to overlapping 95% CIs

Sample	Unique barcodes identified	No. barcodes merged
Fatpad 1 (1° and 2°)	78	11
Fatpad 2 (1° and 2°)	62	22
Fatpad 3 (1° and 2°)	40	11
Fatpad 4 (1° and 2°)	54	9
Fatpad 5	422	224
Fatpad 6	274	98
Fatpad 7	299	122
Fatpad 8	127	66
Human xenograft 1	44	24
Human xenograft 2 (1° and 2°)	46	10
Human xenograft 3	303	215

The number of barcodes merged due to overlapping 95% CI indicated in the table above correspond to experiments described later in Chapters 3 and 4 for transplants into the mouse fatpad and human xenografts, respectively.

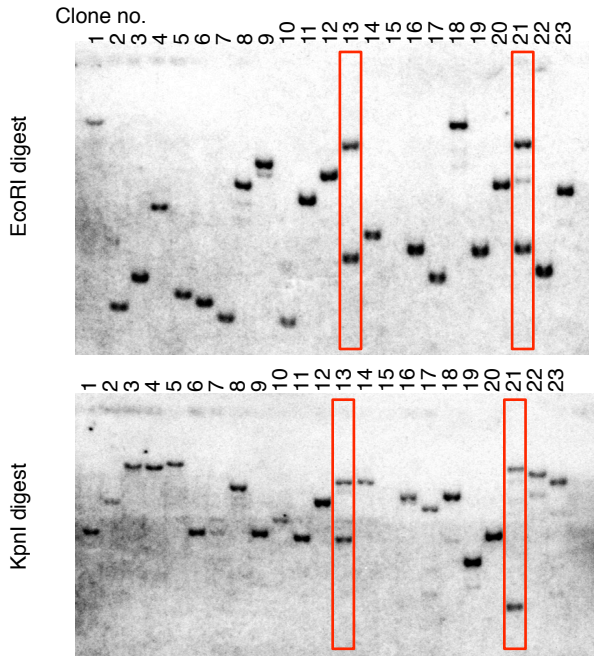

Figure 2.12 Southern blot revealing frequency of multiple integrations

Twenty-three clones were analyzed to determine the number of unique barcode integrations in each. DNA extracts from each clone were individually digested with restriction endonuclease *EcoR*I (top) or *Kpn*I (bottom), and the blot was incubated with a ^{32}P-labeled probe recognizing the *GFP* reporter gene. Each band indicates a single integration, and the clones with >1 integration are boxed in red. The blot was probed with radioactive ^{32}P, and imaged with a phospho-analyzer.

CHAPTER 3: INVESTIGATING THE HETEROGENEITY OF MOUSE MAMMARY CELL REGENERATIVE ACTIVITY ASSESSED IN A SYNGENEIC TRANSPLANT MODEL

3.1 Introduction

The overall goal of this chapter was to investigate the heterogeneity of regenerative activity that is displayed by normal adult mouse BCs and LCs when assessed in non-limiting transplants. At the time the experiments in this thesis were being designed, MRU activity had been found to be a property exclusively associated with mammary BCs. When limiting doses of cells are transplanted, a single basal mammary stem cell can demonstrate the ability to reconstitute an entire mammary gland (Shackleton et al., 2006; Stingl et al., 2006a). Evidence of a long-term lineage-restricted LP cell that can sustain the luminal lineage through multiple cycles of pregnancy and lactation was reported (Booth et al., 2007; Wagner et al., 2002). However, cells with a luminal phenotype have not been found to be able, on their own, to directly regenerate a full mammary gland when injected into a cleared fat pad.(Shackleton et al., 2006; Stingl et al., 2006a).

Subsequent *in situ* lineage tracing studies suggested that the lineage differentiation potential of BCs may be restricted to the myoepithelial lineage, when its potential for generating LCs is out-competed by surrounding luminal-restricted progenitors within the same mammary epithelium (van Amerongen et al., 2012; Van Keymeulen et al., 2011). However, the *in vivo* regenerative activity of BCs when transplanted at non-limiting dilution, and how the number of cells co-transplanted affects their growth and differentiation behaviour is not known.

Accordingly, it was of interest to examine the clonal diversity of cells derived from *non-limiting* transplants of normal adult virgin mouse basal and luminal mammary epithelial cells, in terms of their *in vivo* growth and differentiation activity. To examine this diversity, the barcoded lentiviral library and clone size analysis method described in Chapter 2 was adopted. ***I hypothesized that transplants of non-limiting numbers of mouse BCs would result in the appearance of clones with both bi-lineage and lineage-restricted differentiation, and that transplants of mouse LCs would result in a similar diversity of growth and differentiation, but that these clones would be less prevalent than their BC counterparts.***

3.2 Materials and methods

3.2.1 Preparation and transduction of mouse mammary epithelial cell suspensions

These methods have been described in Chapter 2 (Sections 2.2.6 and 2.2.7).

3.2.2 Flow cytometry

To purify specific subsets of mouse cells, non-specific antibody binding was blocked with rat serum (Sigma) and anti-mouse CD16/32 Fc-gamma III/II Receptor antibody. Mouse mammary cells were then depleted of hematopoietic, and endothelial cells using biotinylated antibodies to mouse CD45, TER-119 and CD31, respectively, and stromal cells using biotinylated antibodies to mouse BP-1, followed by streptavidin-phycoerythrin (PE) or streptavidin-Brilliant violet 421. Allophycocyanin (APC)-conjugated antibody to mouse CD49f, PerCP-Cy5.5-conjugated anti-mouse CD326

(EpCAM), and PE/Cyanin (Cy)-7-conjugated anti-mouse Sca1 were used to isolate the fractions desired. Table 3.1 provides details of the antibodies used and their sources.

3.2.3 Transplantation of mouse mammary cells

Mouse mammary cells were transplanted into mammary fat pads "cleared" of their content of endogenous mammary cells in pre-pubertal (21-24 day post-natal) virgin female C57Bl/6J mice as previously described (Makarem et al., 2013; Shackleton et al., 2006; Stingl et al., 2006a). Briefly, the removal of the endogenous mammary cells involved first making an abdominal midline incision to expose the inguinal fat pads and then the endogenous mammary epithelial tissue within the proximal end of the fat pad, closest to the ventral mid-line of the mouse was excised using the central lymph node as a landmark to define the growing edge of the mammary gland (Figure 1.1). A volume of 10 μL containing 1-6x10^4 lentivirally barcoded mouse mammary cells suspended in 25% Matrigel (BD Biosciences) and sterile trypan blue dye (to visualize the injection) was injected into each fat pad thus cleared. After 4 weeks, a slow-release silicone elastomer pellet containing 2 mg 17ß-estradiol and 4 mg progesterone (Sigma) was implanted subcutaneously, and left there for the remaining 2-4 weeks of the *in vivo* assay. After a total of 6-8 weeks, the fat pads were dissected, and dissociated as described above.

3.2.4 CFC assays

CFC assays were performed by culturing cells in tissue culture dishes with added irradiated 3T3 fibroblasts at 5% O_2 and 37°C for 7 days in the same growth media used for lentiviral transduction (described in Chapter 2).

3.2.5 Immunohistochemistry

Fat pads containing regenerated mouse mammary tissue were fixed in 10% buffered formalin (Fisher), washed in 70% ethanol, and embedded in paraffin. 4 μm sections were treated first with Target Retrieval solution (DAKO) and then Cleanvision solution (Immunologic) for mouse tissues. Sections were stained with an anti-CK14 antibody, an anti-SMA antibody, an anti-CK8 antibody, or an anti-CK18 antibody. A secondary mouse or rabbit antibody conjugated to alkaline phosphatase and treated with permanent red (DAKO) was used to obtain a positive pink staining. Table 3.1 provides details of the antibodies used and their sources.

3.2.6 Barcode sample processing and analysis

This was performed using "spiked-in" controls, as described in Chapter 2 (Sections 2.2.4, 2.2.8, 2.2.9 and 2.2.10).

3.3 Results

3.3.1 Experimental design

$CD45^{-}Ter119^{-}BP\text{-}1^{-}CD31^{-}EpCAM^{+}CD49f^{++}$ (operationally referred to hereafter as "basal") cells, or BCs, were obtained by FACS at a purity of >97% (Figure 3.1A) from 8-12 week-old virgin female C57Bl/6J mice. The sort gates used have been previously shown to enrich for both MRUs (Shackleton et al., 2006; Stingl et al., 2006a) and CFCs (Makarem et al., 2013; Smalley et al., 1998). We then transduced these cells using a 4-

hour infection protocol expected to deliver a single integrated viral gene into >90% of the transduced cells (as demonstrated in Chapter 2).

In a first experiment, the transduced cells were then cultured for 7 days to allow expression of the GFP reporter with minimal cell expansion, and the GFP^+ cells (~30%) present at the end of that period were then isolated by FACS and transplanted (~$6x10^4$/fat pad, grafts #1-4, Figure 3.2). Seven weeks later, the regenerated glands were removed, and the total $GFP^+EpCAM^+CD49f^{++}$ (basal) and $GFP^+EpCAM^{++}CD49f^+$ (operationally referred to hereafter as "luminal") cell, or LC, populations were isolated from each fat pad by FACS (>97% purity, Figure 3.1B). DNA extracts were then prepared directly from ~40% and ~90% of these GFP^+ cells, respectively, to determine their barcode composition by MPS. Another 10% of each sorted fraction was first expanded *in vitro* (under CFC assay conditions) and then DNA extracts obtained and sequenced to extend the primary clone data. The remaining ~50% of the original regenerated BCs harvested from each primary fat pad were transplanted into a cleared fat pad in a secondary mouse and their outputs assessed another 12 weeks later (Figure 3.2).

In a second experiment, the transduced cells were transplanted immediately following the 4-hour exposure to virus at cell doses of $5x10^4$ and 10^4 per fat pad (2 fat pads/cell dose, grafts #5, #6, and #7, #8, respectively, Figure 3.3). FACS analysis of a separate aliquot of these cells cultured for another 48 hours demonstrated 36% of the cells to be GFP^+, thus the transplant doses of $5x10^4$ and 10^4 correspond to ~$1.8x10^4$ and $3.6x10^3$ GFP^+ cells, respectively. Eight weeks post-transplant, histological, immunochemical, and *in vitro* CFC content analysis of the regenerated mammary structures confirmed their normal organization and composition (Figure 3.4). This

included the presence of the recently described $Sca1^+$ (non-CFC-containing) and $Sca1^-$ (CFC-containing) subsets of LCs (Figure 3.1B). Approximately 90% of the total barcoded (GFP^+Sca1^+ and GFP^+Sca1^-) LCs and ~50% of the matching GFP^+ BCs harvested from the remaining fat pads in this second experiment were then analyzed directly for their barcode content by MPS.

To investigate the potential of LCs to regenerate mammary cells *in vivo*, $EpCAM^{++}CD49f^+$ LCs, previously demonstrated to contain a high (~20%) frequency of CFCs but minimal to no MRUs, were isolated by FACS (at >99.9% purity, Figure 3.1A), similarly barcoded, and transplanted into the cleared fat pads of syngeneic recipient mice (2 fat pads at ~$1.3x10^5$ cells/ fat pad, Figure 3.5). A separate aliquot of the transduced cells was analyzed 48 hours later and found to contain 35% GFP^+ cells (indicating that each fat pad had been transplanted with ~ $4.6x10^4$ GFP^+ LCs). Eight weeks later, the fat pads into which these cells were transplanted were imaged by whole-mount fluorescence microscopy which revealed the presence of elaborate, branched mammary structures typical of a regenerated mammary gland (Figure 3.6). A single cell suspension was then prepared from these fat pads and FACS used to isolate the GFP^+ BCs (50% of the total), and the GFP^+ (90% of the total) $Sca1^+$ and $Sca1^-$ luminal subsets (all at >97% purities, Figure 3.1B). Each of these was then analyzed directly for its barcode content by MPS.

Barcode analysis was performed using the spiked-in method described in Chapter 2 (the primary and secondary transplants in grafts #1-4 corresponding to HiSeq MPS run #1 and 2, respectively, grafts #5-8 and #9-10 corresponding to MPS run #3 on the MiSeq). The number of LCs and BCs retrieved from each graft and analyzed by MPS is shown in Table 3.2.

3.3.2 Mouse basal mammary epithelial cells commonly display restricted as well as bi-lineage differentiation patterns in primary recipients

Analysis of the barcode sequences in the cells obtained from the primary mice transplanted with cells cultured for 7 days following transduction identified a total of 144 clones (112 from the directly analyzed cells and another 32 from the cells that were expanded *in vitro* prior to DNA extraction, Figure 3.7, see primary mouse data in Table 3.3) with a maximum clone size of 8,660 cells (Figure 3.8A). Approximately 40% (45/112) of the regenerated clones and all of the 27 largest clones (>400 cells/clone) were bi-lineage (Figure 3.8B) containing basal and luminal progeny in approximately equivalent numbers over a large range of clone sizes (Figure 3.8C). In the other 67 primary clones in this experiment, only LCs (56/112), or only BCs (11/112), were detected (Figure 3.8B). Although the average size of these single lineage clones was smaller than that of the bi-lineage clones (Figure 3.9), some contained as many as 380 cells, thus arguing against the detection of a restricted lineage content being explained by a smaller clone size. However, a number of clones of <100 cells must have been missed since not all the GFP^+ cells detected could be confidently assigned to a particular clone (Table 3.2).

In the mice transplanted with cells directly following transduction, we identified a total of 611 primary clones (374 in the fat pads injected with $5x10^4$ cells each and 237 in the fat pads injected with 10^4 cells each, Figure 3.10, Table 3.4, Figure 3.11A). The same 3 different types of clones were observed as in the first experiment, and their overall representation remained similar; i.e., the clones containing only LCs were the most

prevalent and those containing only BCs were the least prevalent (Figure 3.11B). As in the previous experiment, many bi-lineage clones contained equivalent numbers of BCs and LCs, although examples of greater bias were noted in the second experiment (Figure 3.11C). On average, all 3 different types of clones were larger than in the first experiment (some single lineage clones containing up to 6,000 cells). Similar to the first experiment, clones detected with only LCs were the most prevalent, whereas BC-only clones were the least prevalent. It is interesting, however, that the proportion of bi-lineage clones increased slightly (from 33% to 43% in the transplants of $5x10^4$ and 10^4 BCs, respectively) concomitant with a decrease in the proportion of basal-restricted clones (18% to 6%, respectively), whereas the proportion of luminal-restricted clones remained relatively constant (49% and 51%, respectively, Figure 3.11B). This suggests that under conditions where fewer BCs are transplanted, some basal-restricted clones have the potential to produce cells of the luminal lineage. Further analysis of the differences between the BC transplants performed at high versus low doses showed that the distribution of bi-lineage, luminal-restricted and basal-restricted clone sizes was similar for both input transplant doses. However, at the lower transplant dose, the basal-restricted clones produced substantially larger single-lineage clones (up to 5,187 BCs compared to 167 BCs at the higher transplant dose, Figure 3.12, Table 3.4). This suggests a mechanism whereby the expansion of basal-restricted clones is affected by their prevalence.

More detailed analysis of the $Sca1^+$ and $Sca1^-$ fractions within the luminal populations present in both the bi-lineage and luminal-restricted clones showed that both included some in which only $Sca1^+$, or only $Sca1^-$ cells were detected, but these were

generally small, close to the limit of detection. In contrast, those in which both Sca1$^+$ and Sca1$^-$ cells were detected were a common feature of the largest clones containing LCs, regardless of whether BCs were also detectable in them (Figure 3.13). However, the presence of the more primitive Sca1$^-$ LCs was a significantly more common feature of the luminal-restricted clones compared to the bi-lineage clones (47% vs 18%, $p = 0.023$).

These results reveal a previously unanticipated diversity in the growth and differentiation activity of basal mammary cells when transplanted under non-limiting conditions. Further, the consistent detection of both bi-lineage and lineage-restricted clones contrasts with the results of previous limiting dilution or single-cell transplant studies where the cells with regenerative activity universally display bi-lineage differentiation activity .

3.3.3 Serially transplanted clones derived from mouse BCs display unexpected patterns of growth and differentiation

Secondary recipients of barcoded mouse mammary cells were also set up in the first experiment. For these, 50% of the BCs harvested from each of the 4 primary fat pads were injected into 2 secondary fat pads, and the pairs pooled 12 weeks later for FACS separation of BCs and LCs and analyzed by MPS (Figure 3.2). From the analyses of these cells, we identified a total of 63 clones in the secondary mice (Figure 3.7), and found that they spanned the same range of sizes (up to 8,160 cells) and types as the clones obtained in the primary mice from which they had been derived (Figure 3.8). Strikingly, 48% (30/63) of the clones detected in the secondary mice had not attained a detectable size in the primary mice (Figure 3.14A). Moreover, the features of the corresponding primary

clones predicted neither the differentiation pattern nor the size of the other 33 clones evident in the secondary mice (Figure 3.14B). In fact, very few primary clones that had appeared exclusively basal were perpetuated when retransplanted and most clones detected in the secondary mice were small (~100 cells) and only contained a single lineage. On the other hand, 15 clones that initially appeared to be luminal-restricted and small (<200 cells) became robustly bi-lineage (300-8,000 cells) in secondary hosts (a 300-fold expansion in clone size in some instances, Figure 3.14B). Conversely, 4 initially bi-lineage clones produced only a single lineage in secondary hosts, with 11 continuing to display robust bi-lineage activity. These results confirm that a display of bi-lineage differentiation activity in primary mice is not a universal feature of cells with serially transplantable regenerative activity, nor does the differentiation activity displayed after 7 weeks by BCs transplanted in non-limiting numbers necessarily reflect their full differentiation potential.

Interestingly, most of the 30 clones that first become evident in the secondary mice were either bi-lineage (53%) or basal-restricted (37%) and the 3 clones (10%) that appeared luminal-restricted were small (Figure 3.14B). This made it difficult to exclude a potential content of BCs. Thus, delayed clonal growth, at least in transplants of non-limiting numbers of cells with regenerative potential, may be associated with predominantly basal or bi-lineage differentiation ability.

In summary, many BCs can demonstrate delayed growth, and delayed differentiation potential when transplanted at non-limiting doses. Furthermore, the growth and differentiation patterns displayed by a clone in a primary mouse may not be predictive of its activity in a subsequent transplant.

3.3.4 LCs display unanticipated developmental plasticity in primary recipients

Analysis of the barcode sequences in the cells obtained from two fat pads each transplanted with $1.3x10^5$ LCs identified 14 clones in one fat pad (fat pad #9) and 9 clones in the other (fat pad #10, Figure 3.15, Table 3.5). Compared to transplanted BCs, which display *in vivo* regenerative activity at a frequency of 1 in ~450 cells, LCs with a similar capacity for *in vivo* regeneration were detected at 1 in ~11,000. Although the number of clones contributing to regeneration of a mammary gland was fewer when only LCs were transplanted, the regenerated glands visualized by whole-mount fluorescence microscopy showed a similar extent of growth and branching of GFP^+ ducts. Furthermore, the size distribution of the clones regenerated from transplanted LCs and BCs were strikingly similar (Figure 3.16A). In the first transplant (fat pad #9), all three types of previously described clones were detected (bi-lineage, luminal-restricted and basal-restricted), whereas in the second transplant (fat pad #10) only bi-lineage and luminal-restricted clones were detected. The proportions of clone types was also dissimilar to the grafts from transplanted BCs, since from these transplants, bi-lineage clones were predominant in one fat pad, and luminal-restricted clones were predominant in the other (Figure 3.16B). Interestingly, in the bi-lineage clones, there was an equal proportion of basal and luminal cells, as obtained in the bi-lineage clones derived from BC transplants (Figure 3.16C). Similarly, the bi-lineage clones derived from LCs were larger in size than their single-lineage clones, with a similar clone size distribution (Figure 3.17).

The majority of bi-lineage and luminal-restricted clones derived from the transplanted LCs also contained Sca1$^-$ clonogenic LCs. Moreover, clones containing only Sca1$^+$ (non-clonogenic) cells represented only a minor subset of all clones detected, and these were smaller in size. This is consistent with the notion that these clones contain primitive cells that contribute to *in vivo* regeneration, and are thus not fully differentiated.

These results indicate that although rare, some cells with a luminal phenotype have extensive *in vivo* regenerative activity, similar to that typical of BCs.

3.4 Discussion

The barcoding methodology applied here detected the generation of 792 clones in cleared mammary fat pads transplanted with purified mouse BCs, and 23 clones from purified transplanted LCs. The frequencies of cells with this *in vivo* clonogenic activity were determined to be 1 in 450 BCs as compared to 1 in 11,000 LCs. Analysis of the time of appearance, longevity, size and lineage content of the clones produced showed some, but not all, aspects of the functional properties of MRUs, thus offering evidence of different regenerative behaviours of basal and luminal mammary cells in non-limiting transplants. For example, the prevalence of BC-derived bi-lineage clones containing cells that could proliferate extensively *in vitro* is consistent with the expected reconstruction of a fully reconstituted normal mammary gland from single basal mammary stem cells. However, for the first time, it was possible to quantify the frequency of such mouse bi-lineage clones that contain progeny with the same bi-lineage activity demonstrable in secondary recipients (~30%) – the classical functional test of self-renewal. We also showed that mouse bi-lineage clones have a consistently reduced content of LCs (50% as compared to

the 75% value for the normal resting adult mammary gland), as previously documented for human MRU transplants (Eirew et al., 2008) with many of the regenerated LCs belonging to highly prevalent smaller luminal-restricted clones (at the clone detection limit).

The most novel findings encountered here were the high frequency of clones that were not detected until their transfer to secondary mice and the frequency of primary clones that contained exclusively LCs in spite of their origin from a cell with a basal phenotype. The latter could identify a previously unrecognized type of committed progenitor with extensive but limited proliferative ability. On the other hand, such an outcome could also be explained by either cell non-autonomous or intrinsic stochastic mechanisms affecting commitment or execution of basal versus luminal programs. Under conditions where many clones are stimulated to grow and differentiate simultaneously (Figure 3.19), complex interactions may suppress the expression of particular lineage growth and differentiation programs, and thus contribute to a diversity of clone types not necessarily reflective of the potential of the cell of origin. This was examined directly by transplanting BCs at two different cell doses which then showed that fewer but larger basal-restricted clones were obtained when the number of cells transplanted per fat pad was reduced. Concomitantly, a greater abundance of bi-lineage clones was generated. This finding, suggests that some of the basal-restricted clones obtained when a higher transplant dose was used had a suppressed capacity for luminal differentiation, as found in recent *in situ* lineage-tracing studies (van Amerongen et al., 2012; Van Keymeulen et al., 2011). However, even with the lower transplant dose, we found some basal-restricted clones. Notably, these clones were not detected among the progeny of serially

transplanted cells, suggesting these may represent a previously unrecognized basal-restricted progenitor with limited *in vivo* growth capacity.

Our studies also demonstrate that a rare proportion of LCs can be stimulated to generate bi-lineage or even basal-restricted clones *in vivo*. This finding suggests a level of developmental plasticity consistent with recent studies showing LCs can be activated to display MRU activity upon exposure to Matrigel *in vitro* (Makarem et al., 2013) or *in vivo* (Shehata et al., 2012). On the other hand, these may represent rare cells that are in the process of LC commitment, but have not yet lost the functional potential associated with basal MRUs.

In summary, these results confirm the extensive *in vivo* regenerative activity possessed by normal mouse mammary BCs and a rarer subset of cells with a luminal phenotype. The lability of expression of the growth and differentiation they display suggest complex interactive regulatory mechanisms that now await definition. In addition, they suggest potential relevance to the intrinsic propensity different mammary cell types may have to malignant transformation.

3.5 Figures & tables

Table 3.1 Antibodies used for FACS sorting and immunohistochemistry

Antibodies used for FACS sorting:

Antibody	Fluorophore	Clone	Company
Rat anti-mouse CD16/32 Fc-gamma III/II Receptor	N/A	2.4G2	STEMCELL Technologies
Rat anti-mouse CD45	biotin	30-F11	Biolegend
Rat anti-mouse CD31	biotin	MEC 13.1	BD Pharmingen
Rat anti-mouse TER-119	biotin	N/A	Biolegend
Rat anti-mouse BP-1	biotin	6C3	eBioscience
Streptavidin	phycoerythrin	N/A	eBioscience
Rat anti-mouse CD49f	allophycocyanin	GoH3	R&D Systems
Rat anti-mouse CD326	PerCP-Cy5.5	G8.8	Biolegend
Rat anti-mouse Sca1	PE/Cy7	D7	Biolegend
Streptavidin	Brilliant violent 421	N/A	Biolegend

Antibodies used for immunohistochemistry:

Antibody	Species reactivity	Concentration	Clone	Company
Mouse anti-CK14	mouse and human	1 in 50	NCL-LL002	Novacastra
Mouse anti-SMA	mouse and human	1 in 100	1A4	DAKO
Rabbit anti-CK8	mouse	1 in 50	ab59400	Abcam
Rabbit anti-CK18	mouse	1 in 50	E431-1	Millipore

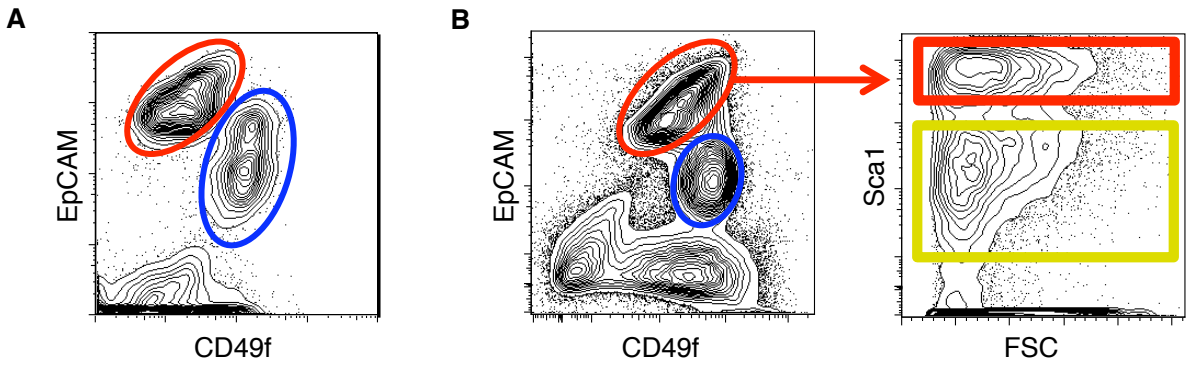

Figure 3.1 Strategy for isolating mouse mammary basal and luminal epithelial cells by FACS

(A) Viable ($DAPI^-$) mouse BCs and LCs were isolated by FACS from freshly dissociated mammary glands of normal adult virgin mice according to their $EpCAM^+CD49f^{++}$ (blue oval gate) or $EpCAM^{++}CD49f^+$ (red oval gate), respectively, after depleting hematopoietic, endothelial, and stromal cells as described previously (Makarem et al., 2013). (B) Regenerated mouse BCs and LCs were similarly isolated using the same strategy as in (A). For grafts #5-8, the regenerated LCs were further separated into $Sca1^+$ (red rectangular gate) and $Sca1^-$ (yellow rectangular gate) subsets as described previously (Shehata et al., 2012).

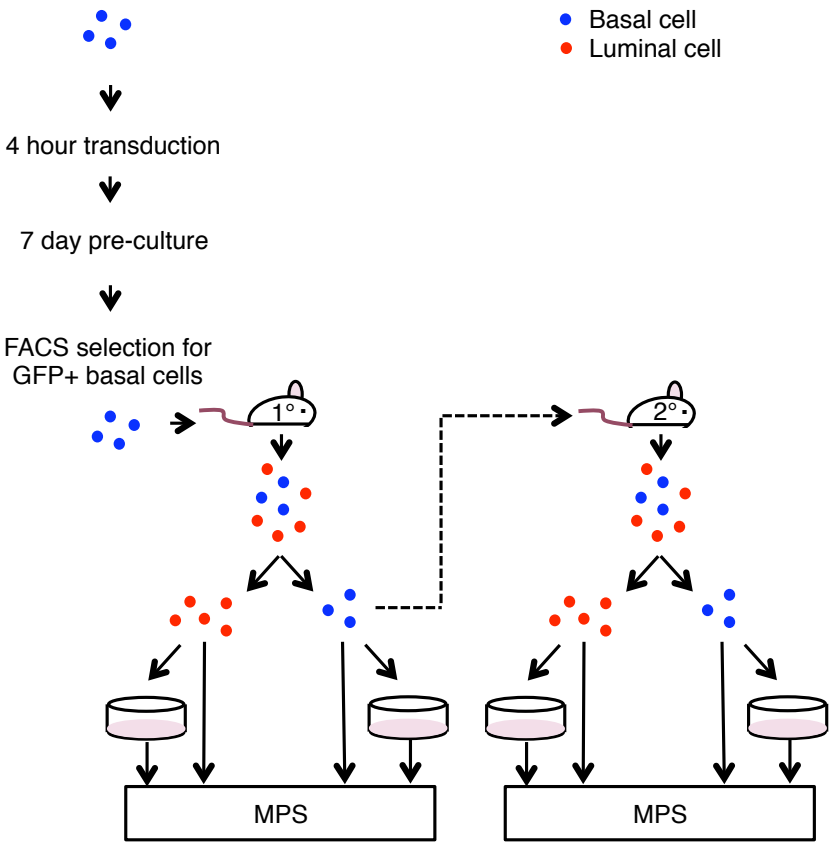

Figure 3.2 Experimental design for tracking the progeny of barcoded basal mammary cells after an initial 7 days *in vitro* prior to transplant

Basal mouse mammary cells were barcoded, cultured for 7 days, and GFP^+ BCs then selected by FACS and transplanted into the cleared fat pads of syngeneic recipient mice. Seven weeks later, the fat pads were removed, and the regenerated mammary cells isolated and sorted into GFP^+ luminal and basal fractions, 90% and 40%, respectively, for direct analysis by MPS. The remaining 10% of the LCs and 10% of the BCs were cultured for 7 days under CFC assay conditions prior to MPS. The remaining ~50% of the BCs were transplanted into the cleared fat pads of a single secondary mouse, and the regenerated cells similarly analyzed another 7 weeks later.

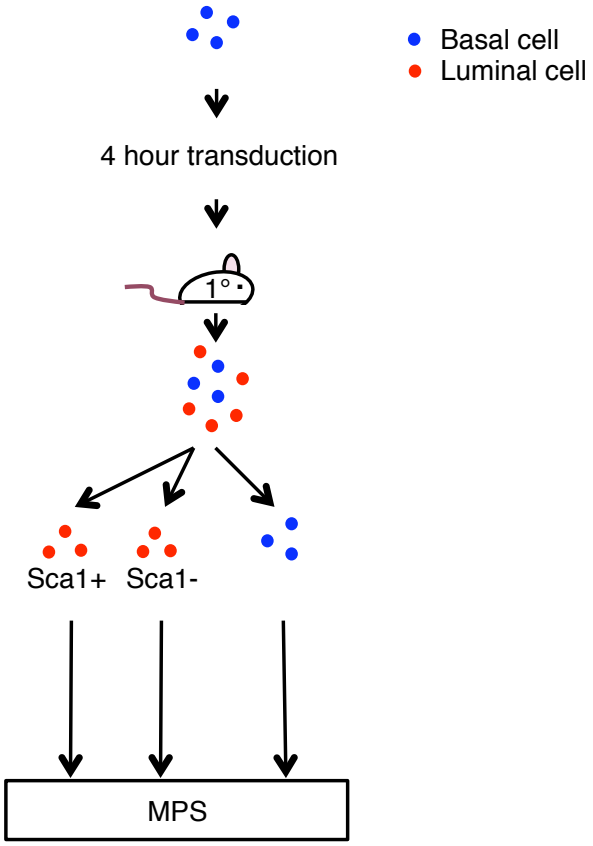

Figure 3.3 Experimental design for tracking the progeny of barcoded basal cells transplanted directly post-transduction

Basal mouse mammary cells were barcoded and immediately transplanted into the cleared fat pads of syngeneic recipient mice. Eight weeks later, the glands were removed, the regenerated cells sorted into GFP^{+} $Sca1^{+}$ and $Sca1^{-}$ luminal, and basal fractions, and 90%, 90% and 40%, respectively, taken for direct analysis by MPS.

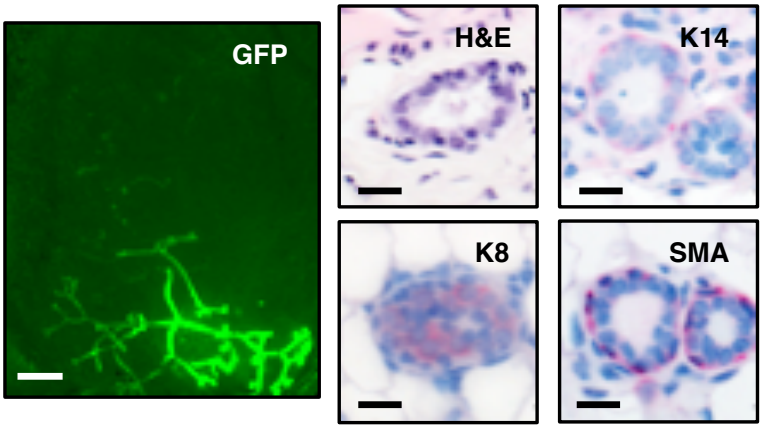

Figure 3.4 Whole-mount and immunochemical analysis of mammary structures regenerated from transplanted barcoded BCs

Left image is a fluorescence photomicrograph showing the GFP^+ cell contribution to a gland structure generated in a fat pad of a mouse transplanted with barcoded (GFP^+) basal mammary epithelial cells. The four panels on the right show a hematoxylin and eosin (H&E) stained paraffin section of another such preparation and the other three show examples of immunohistochemical staining for cytokeratin 8 (K8), a marker of luminal cells, cytokeratin 14 (K14) and smooth muscle actin (SMA), the latter two being markers of BCs. Positive immunohistochemical staining is pink and all slides are counterstained with hematoxylin. White and black scale bars indicate 1 mm and 20 μm, respectively.

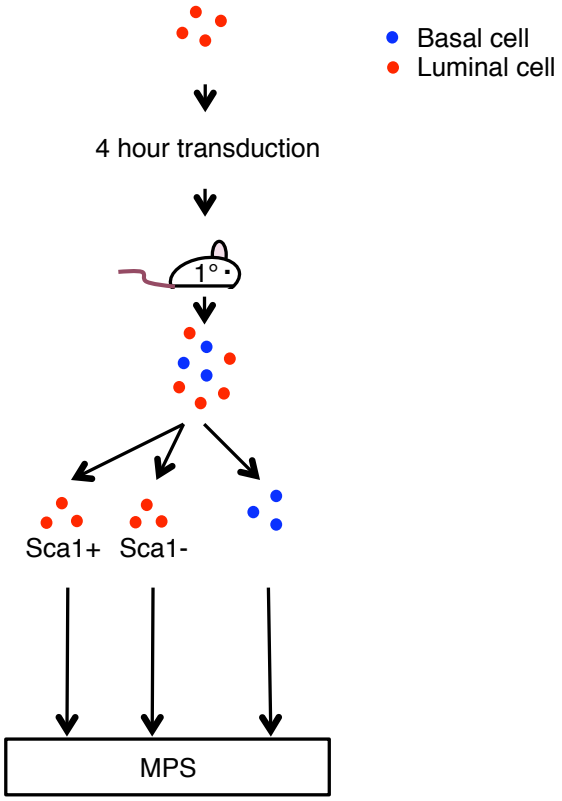

Figure 3.5 Experimental design for tracking the barcoded progeny of LCs transplanted directly after transduction

Luminal mouse mammary cells were barcoded and immediately transplanted into the cleared fat pads of syngeneic recipient mice. Eight weeks later, the glands were removed, the regenerated cells sorted into GFP^+ $Sca1^+$ and $Sca1^-$ luminal, and basal fractions, and 90%, 90% and 40%, respectively, taken for direct analysis by MPS.

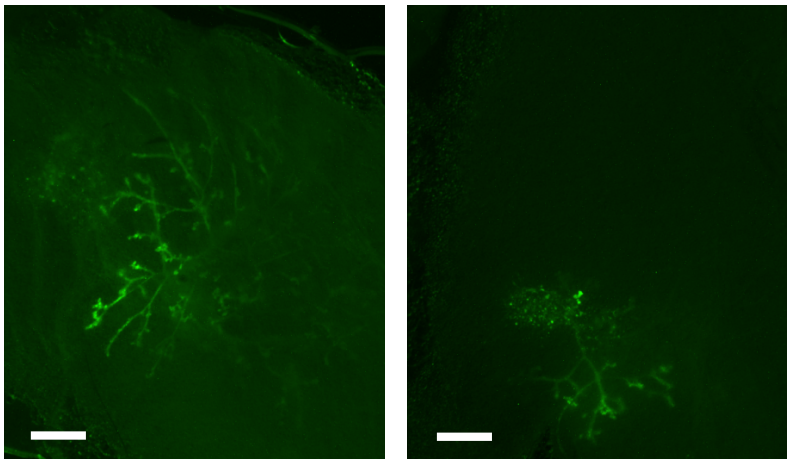

Figure 3.6 Whole-mount fluorescent images of regenerated mammary structures from barcoded LC transplants

Images showing the GFP^+ cell contribution to a gland structure generated in a fat pad from barcoded mouse luminal mammary epithelial cells. The image on the left and right correspond to grafts #9 and 10, respectively (Table 3.2). White scale bars indicate 1 mm.

Table 3.2 Clonal data from transplanted BCs and LCs

Clones from $6x10^4$ purified BCs transplanted after an initial 7 days in culture:

Fat pad	1° BCs	1° LCs	1° BCs re-transplanted	2° BCs	2° LCs
1	11,270	27,739	4,335	20,550	8,686
2	36,554	17,868	16,977	26,490	10,000
3	12,394	7,415	4,897	12,900	5,400
4	53,911	28,824	25,655	3,000	3,650

% of cells analyzed cells that were not detected as barcoded (below designated threshold) in primary mice:

Fatpad	Fraction	# cells detected	# cells sampled	% undetected	Minimum no. of clones undetected (if assuming 100 cells)
1	Basal	5,299	11,270	53	60
	Luminal	9,942	27,739	64	178
2	Basal	8,764	36,554	76	278
	Luminal	4,369	17,868	76	135
3	Basal	8,431	12,394	32	40
	Luminal	7,674	7,415	0	0
4	Basal	18,861	53,911	65	351
	Luminal	13,654	28,824	53	152

Phenotypes of cells produced from $5x10^4$ purified BCs:

Mouse	Regenerated BCs		Regenerated Sca1- LCs		Regenerated Sca1+ LCs	
	Isolated	Sampled	Isolated	Sampled	Isolated	Sampled
5	13,826	6,713	13,311	12,111	6,266	5,066
6	15,678	7,639	10,604	9,404	4,994	3,794

Phenotypes of cells produced from 10^4 purified BCs:

Mouse	Regenerated BCs		Regenerated Sca1- LCs		Regenerated Sca1+ LCs	
	Isolated	Sampled	Isolated	Sampled	Isolated	Sampled
7	14,410	7,005	5,634	4,434	5,753	4,553
8	8,911	4,255	6,119	4,919	3,998	2,798

Phenotypes of cells produced from $1.3x10^5$ purified LCs:

Mouse	Regenerated BCs		Regenerated Sca1- LCs		Regenerated Sca1+ LCs	
	Isolated	Sampled	Isolated	Sampled	Isolated	Sampled
9	8,928	4,289	6,492	6,142	4,688	4,338
10	956	303	692	342	982	632

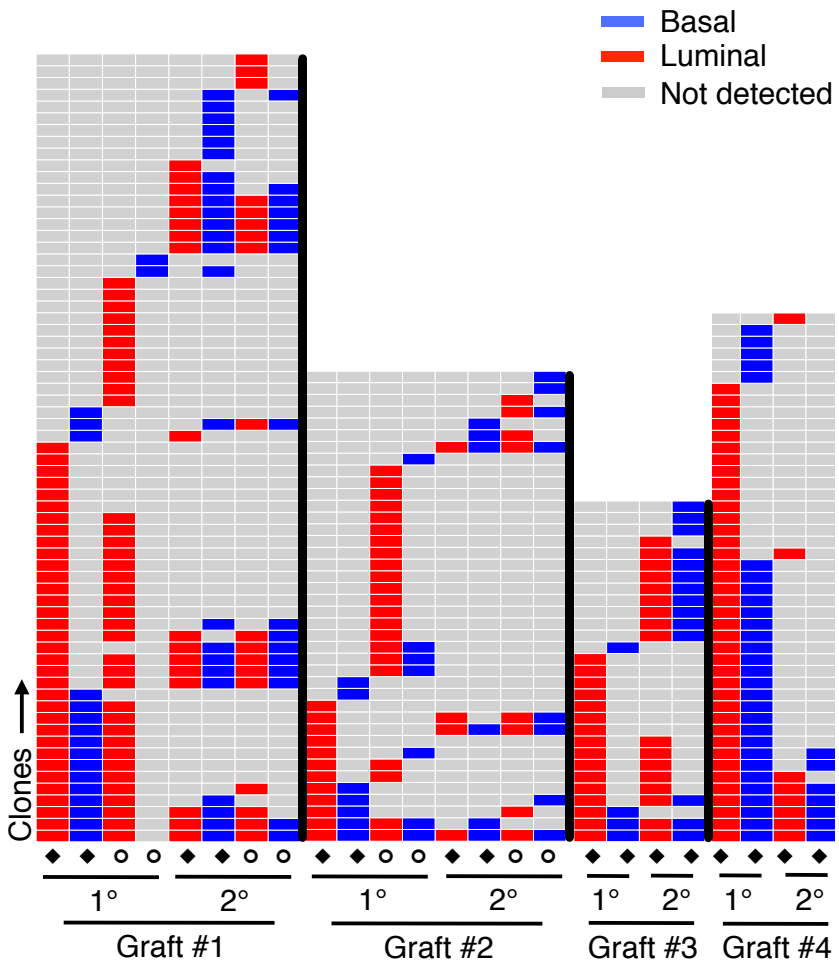

Figure 3.7 Barcoded clones produced from BCs transplanted after 7 days in culture post-transduction

A depiction of all clones detected in the 4 fat pads (#1 to #4) of the primary (1°) mice and, where relevant, in the transplanted fat pads of secondary (2°) recipients from the first such experiment. The y-axis shows sequential clones detected within each fat pad analyzed (the vertical black line separates the clones found in each fat pad). The columns refer to the clones detected in the cells isolated directly from the *in vivo* assay (◆) or after a further 1-week expansion *in vitro* (O).

Table 3.3. Clones detected in primary and secondary mice transplanted with barcoded BCs after 7 days in culture post-transduction

Graft #1:

Clone	Primary				Secondary			
	BCs	LCs	Basal CFC	Luminal CFC	BCs	LCs	Basal CFC	Luminal CFC
1	1854	3308	0	10910	135	415	0	851
2	1011	2071	0	22190	199	0	0	0
3	512	631	0	7665	0	0	0	164
4	407	667	0	17800	0	0	0	0
5	287	243	0	111	0	0	0	0
6	236	203	0	29	0	0	0	0
7	232	186	0	210	0	0	0	0
8	185	261	0	598	0	0	0	0
9	155	266	0	672	3291	4869	30118	6792
10	135	192	0	0	0	0	0	0
11	121	226	0	363	2251	3360	20417	4672
12	85	218	0	4294	0	0	0	0
13	34	0	0	0	0	141	0	0
14	23	0	0	0	87	0	6290	320
15	13	0	0	0	0	0	0	0
16	9	75	0	321	0	0	0	0
17	0	191	0	2974	220	373	13373	1207
18	0	125	0	6766	0	0	0	0
19	0	108	0	1812	0	0	0	0
20	0	101	0	0	0	0	0	0
21	0	96	0	2536	0	0	0	0
22	0	78	0	567	0	0	0	0
23	0	77	0	177	0	0	0	0
24	0	66	0	691	119	0	4371	0
25	0	65	0	4605	0	0	0	0
26	0	64	0	299	0	0	0	0
27	0	59	0	0	0	0	0	0
28	0	54	0	2683	0	0	0	0
29	0	53	0	0	0	0	0	0
30	0	52	0	0	0	0	0	0

(Table continued on subsequent page…)

Clone	Primary				Secondary			
	BCs	LCs	Basal CFC	Luminal CFC	BCs	LCs	Basal CFC	Luminal CFC
31	0	50	0	717	0	48	4018	761
32	0	45	0	0	536	275	2080	147
33	0	39	0	0	0	0	0	0
34	0	27	0	1055	0	0	0	0
35	0	26	0	540	40	450	6849	901
36	0	18	0	0	0	0	0	0
37	0	2	0	306	126	163	2562	784
38	0	0	510	0	0	0	0	0
39	0	0	144	0	72	0	0	0
40	0	0	0	2119	0	0	0	0
41	0	0	0	1911	0	0	0	0
42	0	0	0	1708	0	0	0	0
43	0	0	0	1399	0	0	0	0
44	0	0	0	1181	0	0	0	0
45	0	0	0	1115	0	0	0	0
46	0	0	0	1034	0	0	0	0
47	0	0	0	867	0	0	0	0
48	0	0	0	756	0	0	0	0
49	0	0	0	515	0	0	0	0
50	0	0	0	168	0	0	0	0
51	0	0	0	0	403	896	910	777
52	0	0	0	0	370	0	0	0
53	0	0	0	0	237	520	7698	649
54	0	0	0	0	230	55	6049	0
55	0	0	0	0	197	144	12259	547
56	0	0	0	0	175	193	2791	1779
57	0	0	0	0	136	0	0	0
58	0	0	0	0	99	0	0	0
59	0	0	0	0	92	0	0	0
60	0	0	0	0	69	376	3085	1982
61	0	0	0	0	56	0	0	0
62	0	0	0	0	27	0	2756	0
63	0	0	0	0	17	78	0	0
64	0	0	0	0	0	109	0	0
65	0	0	0	0	0	0	0	229
66	0	0	0	0	0	0	0	144
67	0	0	0	0	0	0	0	5

Graft #2:

Clone	Primary				Secondary			
	BCs	LCs	Basal CFC	Luminal CFC	BCs	LCs	Basal CFC	Luminal CFC
68	5529	3131	3146	838	1219	1379	38846	7588
69	2490	684	392	99	9	0	0	0
70	351	95	0	0	0	0	0	1097
71	152	93	0	0	0	0	0	0
72	117	0	0	0	0	0	0	0
73	75	0	0	0	0	0	0	0
74	50	82	0	0	0	0	4497	0
75	0	61	0	205	0	0	0	0
76	0	53	0	0	0	0	0	0
77	0	79	0	250	0	0	0	0
78	0	37	1433	0	0	0	0	0
79	0	27	0	0	3469	4493	123356	31834
80	0	15	0	0	0	0	0	0
81	0	12	0	0	0	31	2824	1090
82	0	0	2184	205	0	0	0	0
83	0	0	1580	292	0	0	0	0
84	0	0	300	236	0	0	0	0
85	0	0	241	0	0	0	0	0
86	0	0	0	266	0	0	0	0
87	0	0	0	207	0	0	0	0
88	0	0	0	197	0	0	0	0
89	0	0	0	174	0	0	0	0
90	0	0	0	155	0	0	0	0
91	0	0	0	149	0	0	0	0
92	0	0	0	145	0	0	0	0
93	0	0	0	116	0	0	0	0
94	0	0	0	107	0	0	0	0
95	0	0	0	107	0	0	0	0
96	0	0	0	99	0	0	0	0
97	0	0	0	79	0	0	0	0
98	0	0	0	62	0	0	0	0
99	0	0	0	49	0	0	0	0
100	0	0	0	16	0	0	0	0
101	0	0	0	0	150	0	0	292

(Table continued on subsequent page…)

Clone	Primary				Secondary			
	BCs	LCs	Basal CFC	Luminal CFC	BCs	LCs	Basal CFC	Luminal CFC
102	0	0	0	0	90	27	260	59
103	0	0	0	0	0	0	1084	0
104	0	0	0	0	0	0	898	222
105	0	0	0	0	0	0	397	0
106	0	0	0	0	37	0	0	0
107	0	0	0	0	0	0	0	94

Graft #3:

Clone	Primary		Secondary	
	BCs	LCs	BCs	LCs
108	4155	3090	965	927
109	3989	2878	0	0
110	173	337	200	917
111	114	0	0	0
112	0	187	0	87
113	0	161	0	56
114	0	156	0	0
115	0	146	0	36
116	0	127	0	0
117	0	111	0	106
118	0	109	0	17
119	0	102	0	0
120	0	93	0	0
121	0	63	0	0
122	0	46	0	0
123	0	43	547	623
124	0	25	0	0
125	0	0	1660	2770
126	0	0	1062	1062
127	0	0	235	459
128	0	0	219	73
129	0	0	180	225
130	0	0	165	0
131	0	0	152	187
132	0	0	97	0
133	0	0	87	269

(Table continued on subsequent page…)

Clone	Primary		Secondary	
	BCs	LCs	BCs	LCs
134	0	0	81	0
135	0	0	75	379
136	0	0	0	39

Graft #4:

Clone	Primary		Secondary	
	BCs	LCs	BCs	LCs
137	3733	3608	1007	1370
138	3155	1329	514	0
139	2594	1564	1070	2173
140	2489	1938	0	0
141	1179	555	0	0
142	1127	795	672	0
143	753	389	0	0
144	708	502	624	635
145	702	73	0	0
146	373	219	0	0
147	290	136	0	0
148	223	181	0	1417
149	218	137	0	0
150	214	0	0	0
151	188	123	6546	15554
152	160	22	0	0
153	153	30	0	0
154	148	106	0	0
155	121	113	108	540
156	89	0	0	0
157	57	0	0	0
158	43	0	0	0
159	35	109	0	0
160	30	124	0	0
161	26	0	0	0
162	19	32	0	0
163	19	16	0	0
164	11	33	0	0
165	5	107	0	0

(Table continued on subsequent page…)

Clone	Primary		Secondary	
	BCs	LCs	BCs	LCs
166	0	383	0	0
167	0	168	0	0
168	0	144	0	0
169	0	100	0	0
170	0	92	0	0
171	0	76	0	0
172	0	71	0	0
173	0	68	0	0
174	0	67	0	0
175	0	60	0	440
176	0	52	0	0
177	0	51	0	0
178	0	41	0	0
179	0	25	0	0
180	0	15	0	0
181	0	0	0	46

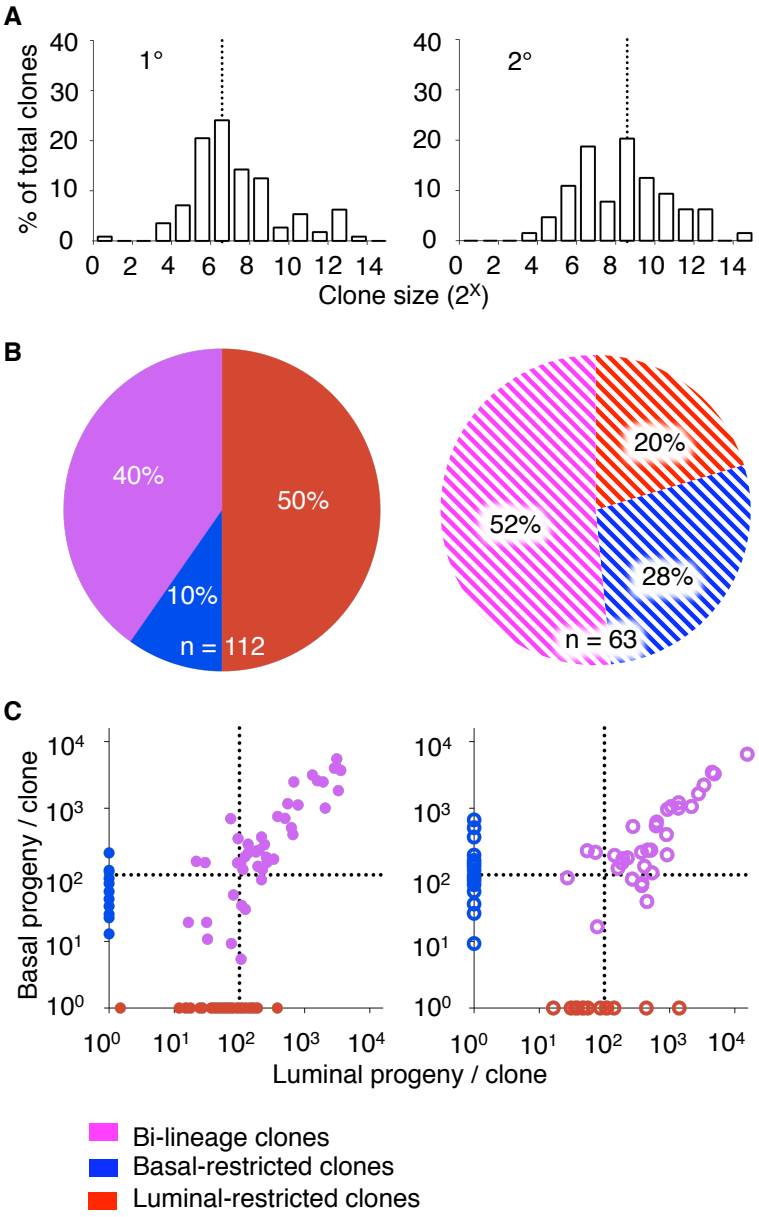
A
% of total clones
1°
2°
Clone size (2^X)
B
40%
50%
10%
n = 112
20%
52%
28%
n = 63
C
Basal progeny / clone
Luminal progeny / clone
Bi-lineage clones
Basal-restricted clones
Luminal-restricted clones

Figure 3.8

Figure 3.8 Size and composition of clones detected in primary and secondary mice transplanted with barcoded BCs after 7 days in culture post-transduction

(A) Distributions of total clone size (binned by $\log_2$ increments) for clones detected in primary (left) and secondary mice (right). The dotted lines indicate the mode of each size distribution plot. (B) Pie charts indicate the proportion of bi-lineage (magenta), luminal-restricted (red), and basal-restricted (blue) clones for either primary (left, solid colours) or secondary (right, colours with white stripes) transplants. (C) The scatter plots show the numbers of basal and/or luminal progeny in each of these clones (solid circles for primary transplants, and open circles for secondary transplants). In these, the dotted line indicates the minimum threshold of reproducible clone detection (100 cells).

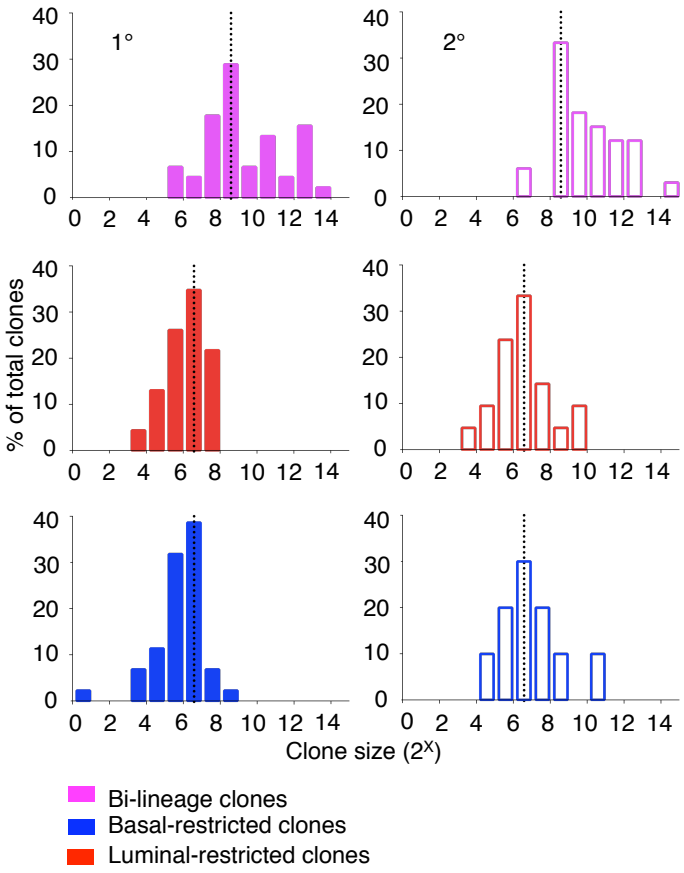

Figure 3.9 Size distribution plots for three clone types in primary and secondary transplants

Size distributions for the 3 primary and secondary clone types indicated in Figure 3.8B (i.e., defined according to their lineage content) are shown. Bi-lineage, luminal-restricted and basal-restricted clones are shown as magenta, red and blue, respectively. Solid and open bars represent primary and secondary clones, respectively. Dotted lines represent the mode of each size distribution plot.

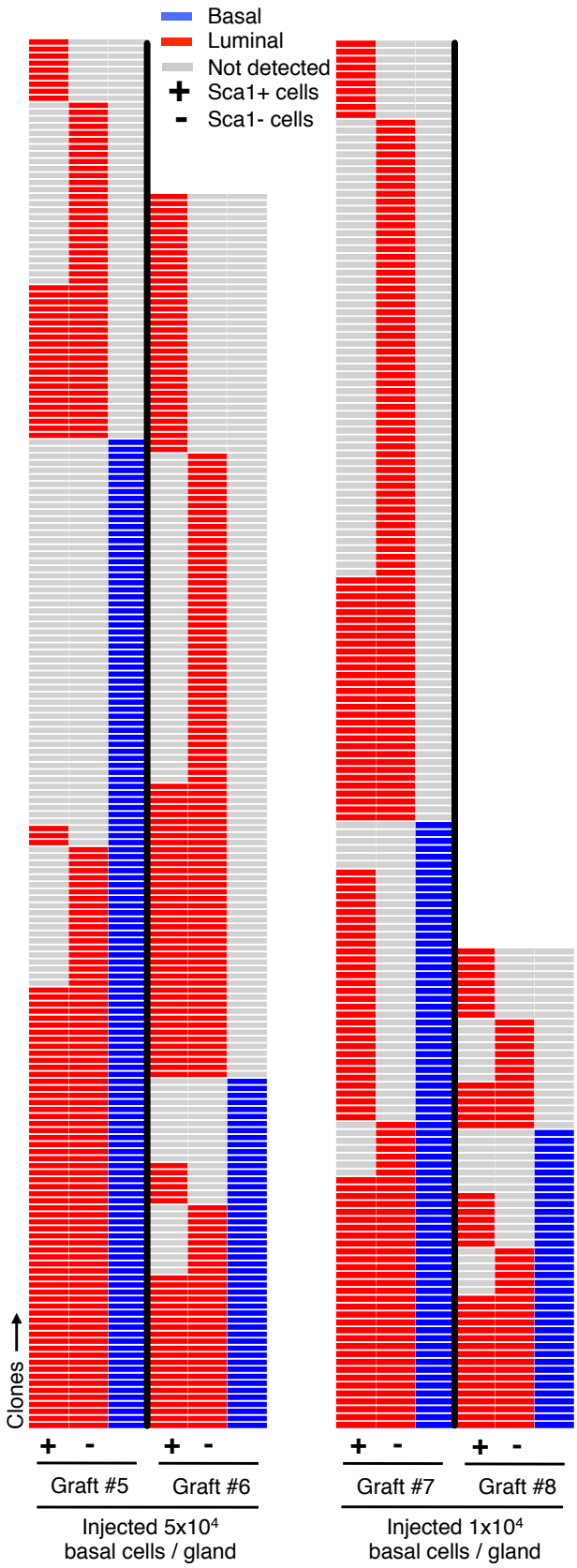

Figure 3.10 Barcoded clones produced from BCs transplanted immediately post-transduction

All clones detected in the 4 fat pads (#5 to #8) of the mice transplanted in the second experiment are shown. The y-axis is the clone number ID within each individual fat pad analyzed. The columns refer to the clones detected in the cells isolated directly from the *in vivo* assay.

Table 3.4 Clones detected in mice transplanted with barcoded BCs immediately post-transduction

Graft #5:

Clone	BCs	$Sca1^+$ LCs	$Sca1^-$ LCs
1	8809	4274	12830
2	7898	224	679
3	7105	174	617
4	2179	417	300
5	1597	3285	11768
6	946	2150	6475
7	352	2652	7897
8	340	133	43
9	253	196	40
10	251	1727	6937
11	243	140	36
12	225	128	547
13	222	170	64
14	206	52	205
15	190	64	216
16	173	249	520
17	167	0	0
18	157	114	299
19	148	53	204
20	148	0	0
21	142	45	0
22	140	0	40
23	134	0	0
24	132	0	0
25	128	114	119
26	124	38	114
27	117	0	92
28	117	0	0
29	111	93	154
30	111	0	91
31	109	0	111
32	107	37	76

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
33	105	100	403
34	105	47	96
35	105	0	0
36	103	0	0
37	103	125	435
38	103	114	197
39	103	67	50
40	102	0	0
41	101	136	435
42	101	109	396
43	101	72	234
44	101	0	35
45	99	0	0
46	99	0	0
47	97	107	438
48	97	37	40
49	93	52	53
50	93	41	40
51	93	0	35
52	91	0	0
53	89	0	0
54	89	120	264
55	89	94	41
56	89	84	227
57	89	0	78
58	88	0	0
59	87	0	0
60	84	40	40
61	84	0	0
62	82	86	0
63	82	45	64
64	82	0	76
65	82	0	0
66	81	0	0
67	81	0	0
68	80	161	288
69	80	0	0

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
70	80	0	0
71	78	133	188
72	78	124	278
73	78	81	350
74	78	0	0
75	76	107	250
76	76	91	267
77	76	50	50
78	76	0	36
79	76	0	0
80	76	0	0
81	76	0	0
82	75	0	0
83	74	37	83
84	74	36	44
85	74	0	121
86	74	0	61
87	74	0	44
88	74	0	34
89	74	0	0
90	74	0	0
91	72	46	55
92	72	0	42
93	72	0	41
94	72	0	0
95	70	0	0
96	70	0	0
97	70	0	0
98	68	169	230
99	68	153	349
100	68	0	39
101	68	0	0
102	68	0	0
103	68	0	0
104	66	138	312
105	66	77	188
106	66	0	0

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
107	66	0	0
108	66	0	0
109	66	0	0
110	65	0	0
111	64	42	34
112	64	37	54
113	64	0	33
114	64	0	0
115	64	0	0
116	64	0	0
117	63	0	0
118	62	107	227
119	62	100	230
120	62	60	220
121	62	51	102
122	62	46	0
123	62	41	86
124	62	0	47
125	62	0	39
126	62	0	0
127	62	0	0
128	62	0	0
129	62	0	0
130	62	0	0
131	62	0	0
132	62	0	0
133	62	0	0
134	61	0	0
135	60	143	229
136	60	126	229
137	60	92	87
138	60	0	98
139	60	0	0
140	60	0	0
141	60	0	0
142	0	238	552
143	0	201	257

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
144	0	140	234
145	0	131	264
146	0	103	320
147	0	93	250
148	0	76	0
149	0	74	234
150	0	72	0
151	0	72	0
152	0	69	210
153	0	69	32
154	0	64	0
155	0	61	218
156	0	61	46
157	0	60	184
158	0	60	70
159	0	60	0
160	0	58	164
161	0	57	73
162	0	57	0
163	0	56	282
164	0	55	0
165	0	53	0
166	0	47	83
167	0	47	0
168	0	46	219
169	0	45	111
170	0	40	79
171	0	38	46
172	0	36	81
173	0	0	88
174	0	0	78
175	0	0	77
176	0	0	72
177	0	0	69
178	0	0	66
179	0	0	64
180	0	0	63

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
181	0	0	56
182	0	0	52
183	0	0	50
184	0	0	50
185	0	0	46
186	0	0	45
187	0	0	45
188	0	0	45
189	0	0	44
190	0	0	43
191	0	0	40
192	0	0	40
193	0	0	39
194	0	0	37
195	0	0	36
196	0	0	36
197	0	0	34
198	0	0	34

Graft #6:

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
1	8115	0	75
2	899	41	0
3	711	246	312
4	588	0	57
5	471	21	86
6	412	42	154
7	353	165	266
8	325	41	64
9	312	196	303
10	246	156	0
11	178	193	0
12	155	0	0
13	135	48	303
14	120	0	0
15	117	45	186
16	114	46	160
17	111	48	252
18	102	0	0
19	100	0	0
20	99	0	0
21	95	0	0
22	92	0	112
23	91	38	94
24	91	48	179
25	88	51	134
26	88	0	0
27	86	0	92
28	82	62	90
29	77	0	80
30	77	63	110
31	76	48	163
32	74	0	62
33	73	57	116
34	71	0	0
35	71	0	88
36	70	53	146

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
37	70	0	75
38	70	0	0
39	68	50	177
40	68	0	0
41	66	71	262
42	66	44	79
43	66	0	77
44	66	0	0
45	66	0	0
46	64	40	0
47	62	44	0
48	62	0	113
49	59	55	112
50	59	46	0
51	0	223	791
52	0	136	322
53	0	117	298
54	0	86	153
55	0	78	0
56	0	77	108
57	0	70	79
58	0	70	126
59	0	69	75
60	0	66	112
61	0	65	0
62	0	64	132
63	0	64	0
64	0	64	95
65	0	63	179
66	0	62	96
67	0	58	114
68	0	57	116
69	0	57	59
70	0	55	142
71	0	55	0
72	0	54	95
73	0	50	0

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
74	0	49	89
75	0	49	0
76	0	48	108
77	0	48	86
78	0	48	66
79	0	48	0
80	0	48	0
81	0	46	113
82	0	46	96
83	0	46	0
84	0	46	0
85	0	45	0
86	0	45	0
87	0	45	0
88	0	44	0
89	0	44	147
90	0	44	0
91	0	44	113
92	0	44	138
93	0	44	107
94	0	44	87
95	0	44	0
96	0	44	0
97	0	44	0
98	0	44	0
99	0	42	129
100	0	42	85
101	0	42	78
102	0	42	0
103	0	42	0
104	0	42	0
105	0	42	0
106	0	41	93
107	0	41	90
108	0	41	69
109	0	41	62
110	0	41	0

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
111	0	41	0
112	0	41	0
113	0	41	0
114	0	40	88
115	0	40	55
116	0	40	0
117	0	40	0
118	0	40	0
119	0	40	0
120	0	40	0
121	0	38	108
122	0	38	82
123	0	38	79
124	0	38	63
125	0	38	0
126	0	38	0
127	0	38	0
128	0	38	0
129	0	38	0
130	0	0	156
131	0	0	154
132	0	0	147
133	0	0	145
134	0	0	125
135	0	0	124
136	0	0	123
137	0	0	115
138	0	0	112
139	0	0	111
140	0	0	109
141	0	0	108
142	0	0	106
143	0	0	102
144	0	0	101
145	0	0	99
146	0	0	97
147	0	0	94

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
148	0	0	93
149	0	0	92
150	0	0	90
151	0	0	88
152	0	0	87
153	0	0	86
154	0	0	86
155	0	0	85
156	0	0	85
157	0	0	84
158	0	0	81
159	0	0	80
160	0	0	80
161	0	0	79
162	0	0	78
163	0	0	77
164	0	0	75
165	0	0	73
166	0	0	69
167	0	0	67
168	0	0	66
169	0	0	64
170	0	0	63
171	0	0	63
172	0	0	62
173	0	0	61
174	0	0	60
175	0	0	59
176	0	0	57

Graft #7:

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
1	3430	445	39
2	893	354	1826
3	820	280	592
4	798	320	1405

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
5	758	0	0
6	736	174	145
7	726	280	674
8	707	0	0
9	639	300	560
10	595	0	47
11	368	87	0
12	347	0	37
13	344	117	405
14	303	55	48
15	292	86	253
16	283	0	0
17	283	93	352
18	277	0	0
19	241	57	127
20	210	53	0
21	204	192	135
22	194	78	79
23	183	0	88
24	171	60	39
25	159	57	159
26	152	0	47
27	142	135	108
28	142	54	0
29	134	60	51
30	128	213	881
31	126	44	64
32	119	0	113
33	115	54	41
34	113	48	86
35	109	39	0
36	107	47	108
37	105	0	48
38	103	151	102
39	103	0	104
40	103	0	0
41	99	42	41

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
42	97	194	599
43	97	79	94
44	97	0	0
45	95	247	744
46	95	49	56
47	93	86	83
48	93	53	51
49	91	69	84
50	89	72	0
51	87	78	0
52	84	59	0
53	84	53	0
54	84	45	0
55	84	42	0
56	84	37	0
57	82	47	0
58	82	45	0
59	80	58	0
60	80	47	0
61	80	42	0
62	78	63	0
63	78	57	0
64	78	37	0
65	78	37	0
66	76	73	0
67	76	42	0
68	76	40	0
69	76	38	0
70	74	42	0
71	72	43	0
72	70	42	0
73	70	38	0
74	70	37	0
75	68	50	0
76	68	40	0
77	68	38	0
78	0	199	735

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
79	0	193	456
80	0	178	730
81	0	168	583
82	0	142	424
83	0	136	469
84	0	116	0
85	0	81	75
86	0	79	118
87	0	73	72
88	0	69	80
89	0	62	119
90	0	60	0
91	0	59	154
92	0	58	52
93	0	54	50
94	0	54	46
95	0	52	50
96	0	49	56
97	0	48	0
98	0	47	185
99	0	47	99
100	0	47	0
101	0	47	0
102	0	45	41
103	0	45	39
104	0	44	56
105	0	43	58
106	0	43	43
107	0	42	44
108	0	40	60
109	0	40	51
110	0	40	47
111	0	39	53
112	0	39	0
113	0	39	0
114	0	38	39
115	0	38	0

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
116	0	38	0
117	0	37	43
118	0	37	0
119	0	0	154
120	0	0	91
121	0	0	88
122	0	0	74
123	0	0	74
124	0	0	72
125	0	0	66
126	0	0	65
127	0	0	64
128	0	0	64
129	0	0	61
130	0	0	58
131	0	0	57
132	0	0	56
133	0	0	55
134	0	0	53
135	0	0	53
136	0	0	52
137	0	0	51
138	0	0	50
139	0	0	50
140	0	0	50
141	0	0	48
142	0	0	47
143	0	0	47
144	0	0	46
145	0	0	46
146	0	0	44
147	0	0	44
148	0	0	44
149	0	0	43
150	0	0	43
151	0	0	43
152	0	0	43

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
153	0	0	43
154	0	0	43
155	0	0	42
156	0	0	42
157	0	0	42
158	0	0	41
159	0	0	41
160	0	0	41
161	0	0	41
162	0	0	41
163	0	0	39
164	0	0	39
165	0	0	39
166	0	0	39
167	0	0	38
168	0	0	38
169	0	0	38
170	0	0	38
171	0	0	38
172	0	0	37
173	0	0	37
174	0	0	37
175	0	0	37
176	0	0	37

Graft #8:

Clone	BCs	$Sca1^+$ LCs	$Sca1^-$ LCs
1	5187	0	0
2	2343	0	73
3	626	0	0
4	533	355	165
5	351	164	130
6	297	120	83
7	215	0	0
8	205	116	97
9	196	0	0
10	188	102	68
11	155	87	50
12	148	1461	0
13	146	0	0
14	140	3130	1679
15	134	2887	1447
16	134	2331	1287
17	132	0	41
18	130	295	67
19	117	0	47
20	96	1173	0
21	96	779	0
22	88	223	253
23	88	0	0
24	82	179	0
25	79	1498	813
26	79	1381	609
27	75	225	41
28	75	182	246
29	73	0	71
30	69	269	0
31	65	0	47
32	63	230	0
33	63	222	40
34	63	0	38
35	63	0	0
36	61	255	0

(Table continued on subsequent page…)

Clone	BCs	Sca1^{+} LCs	Sca1^{-} LCs
37	61	159	81
38	61	0	0
39	0	3558	2041
40	0	603	279
41	0	344	0
42	0	253	37
43	0	233	0
44	0	230	0
45	0	223	0
46	0	207	0
47	0	190	227
48	0	177	0
49	0	167	0
50	0	151	0
51	0	138	72
52	0	134	135
53	0	102	0
54	0	0	55
55	0	0	50
56	0	0	43
57	0	0	41
58	0	0	41
59	0	0	40
60	0	0	38
61	0	0	37

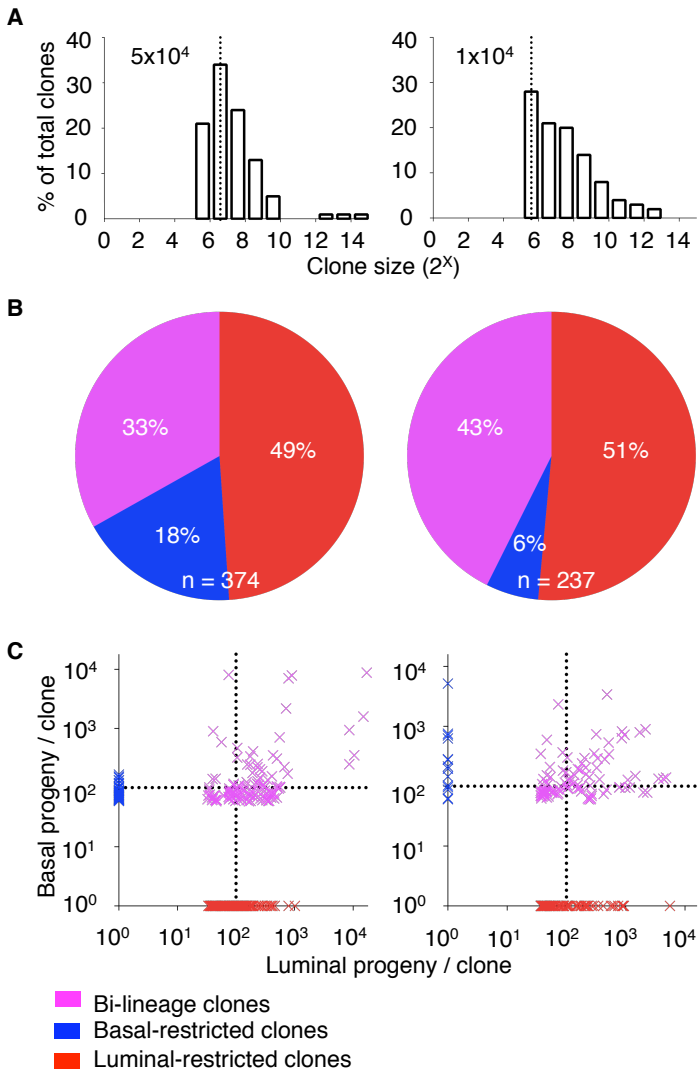

Figure 3.11 Size and composition of clones detected in mice transplanted with barcoded BCs at two different cell doses

(A) Distributions of total clone size (binned by $\log_2$ increments) for clones detected from fat pads transplanted with $5x10^4$ (left) and 10^4 (right) BCs. Dotted lines indicate the mode of each size distribution. (B) Pie charts indicate the proportion of bi-lineage (magenta), luminal-restricted (red), and basal-restricted (blue) clones for either the $5x10^4$ (left) or the 10^4 (right) transplant doses. (C) The scatter plots below show numbers of basal and/or

luminal progeny in each of the clones shown above (in B). In these, the dotted line indicates the minimum threshold of reproducible clone detection (100 cells).

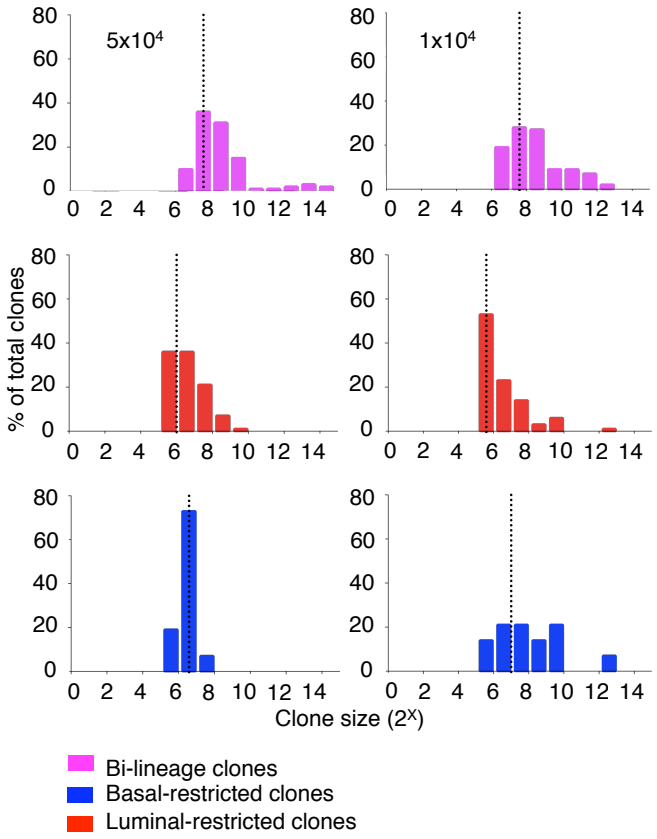

Figure 3.12 Size distribution plots for clone types generated from cells transplanted at two different doses

Size distributions are shown for the three clone types, from transplants of $5x10^4$ (left plots) and 10^4 (right plots) BCs. Bi-lineage, luminal-restricted and basal-restricted clones are shown in magenta, red, and blue, respectively. Dotted lines indicate the mode of each size distribution plot.

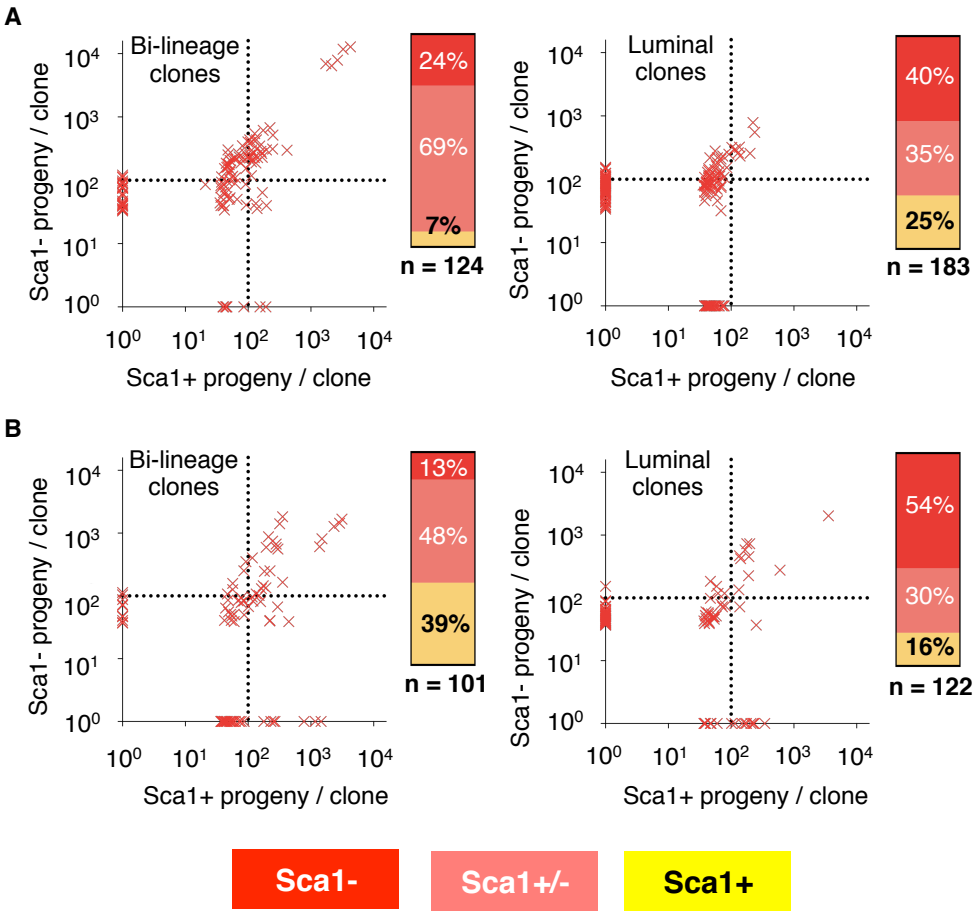

Figure 3.13 Sca1$^+$ and Sca1$^-$ luminal cell content in bi-lineage and luminal-restricted clones derived from transplanted BCs

(A) The scatter plot shows the numbers of Sca1$^-$ (y-axis) and Sca1$^+$ (x-axis) cells from the luminal fraction of bi-lineage (left) and luminal-restricted clones (right), detected in primary mice transplanted with 5x10^4 BCs. The stacked column shows the proportion of either the bi-lineage or luminal-restricted clones that were found to contain exclusively Sca1$^-$ cells (dark red), Sca1$^+$ cells (yellow) or both cell types (light red). (B) Same as (A) for clones in primary mice transplanted with 10^4 BCs.

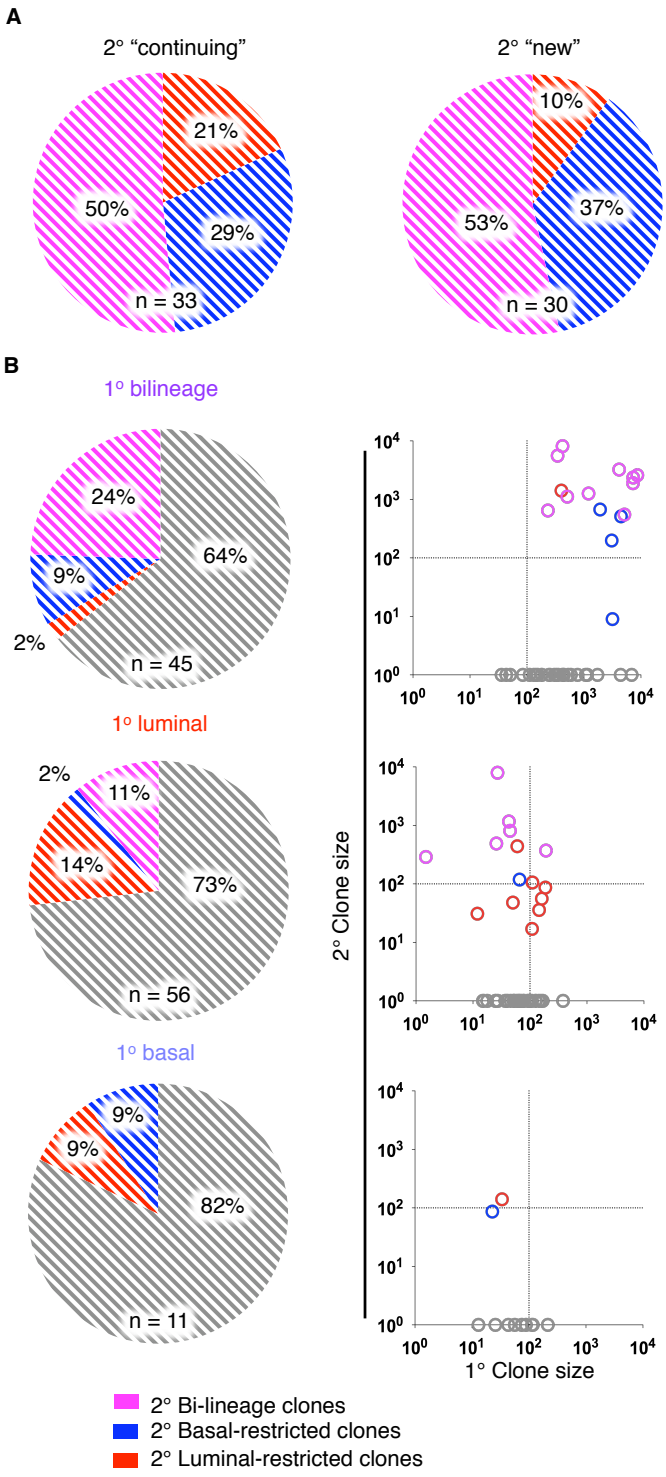
A
2° "continuing"
21%
50%
29%
n = 33
2° "new"
10%
53%
37%
n = 30
B
1° bilineage
24%
64%
9%
2%
n = 45
1° luminal
2%
11%
14%
73%
n = 56
1° basal
9%
9%
82%
n = 11
2° Clone size
1° Clone size
2° Bi-lineage clones
2° Basal-restricted clones
2° Luminal-restricted clones

Figure 3.14

Figure 3.14 Diverse size and lineage composition of mouse mammary clones detected in secondary mice

(A) Pie charts show the proportion of continuing (left) and new (right) clones that were bi-lineage (magenta with white stripes), luminal-restricted (red with white stripes), and basal-restricted (blue with white stripes) in the secondary (2°) mice. (B) Pie charts show the proportions of bi-lineage, luminal-restricted, and basal-restricted clones detected in the primary mice that gave rise to detectable clones in the secondary mice. For each of these, the proportion of clones from the primary transplants that did not re-appear in the secondary transplants are shown in gray with white stripes, and for those that continued, the proportion of clones that were bi-lineage (magenta with white stripes), luminal-restricted (red with white stripes), and basal-restricted (blue with white stripes) in the secondary transplants are also shown. The scatter plots directly beside each pie chart compare the sizes of the matching clones in the primary and secondary transplants. Dotted lines indicate the minimum threshold of reproducible clone detection (100 cells). Each point represents one of the continuing clones, the color of which indicates the composition of each clone in the secondary mouse: bi-lineage (purple open dot), luminal (red open dot), basal (blue open dot), and not detected (gray open dot).

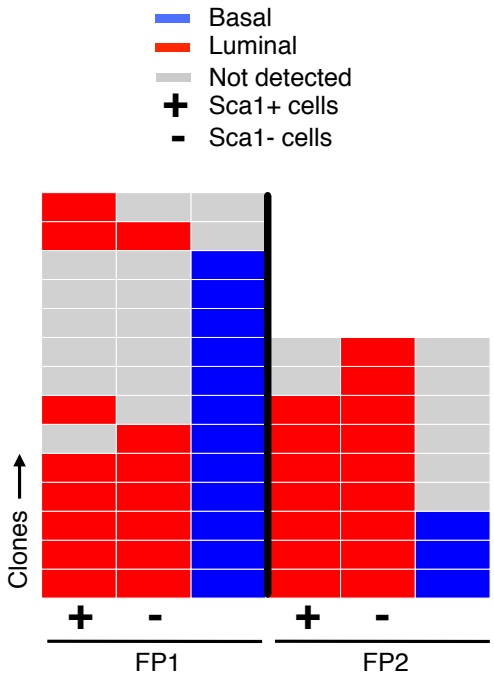

Figure 3.15 Detection of barcoded clones produced from transplanted LCs

All clones detected in the 2 fat pads (#9 to #10) of the mice transplanted with barcoded LCs in the third experiment are shown. The y-axis is the clone number ID within each individual fat pad analyzed. The columns refer to the clones detected in the cells isolated directly from the *in vivo* assay.

Table 3.5 Clones detected in mice transplanted with barcoded LCs

Graft #9:

Clone	BCs	Sca1$^+$ LCs	Sca1$^-$ LCs
1	4646	1532	1367
2	1402	251	184
3	344	1719	918
4	250	1312	682
5	217	1041	628
6	130	0	54
7	231	39	0
8	398	0	0
9	331	0	0
10	280	0	0
11	166	0	0
12	96	0	0
13	0	54	64
14	0	77	0

Graft #10:

Clone	BCs	Sca1$^+$ LCs	Sca1$^-$ LCs
16	78	31	51
17	76	122	628
18	42	38	39
19	0	225	1182
20	0	132	619
21	0	101	484
22	0	21	37
23	0	0	572
24	0	0	40

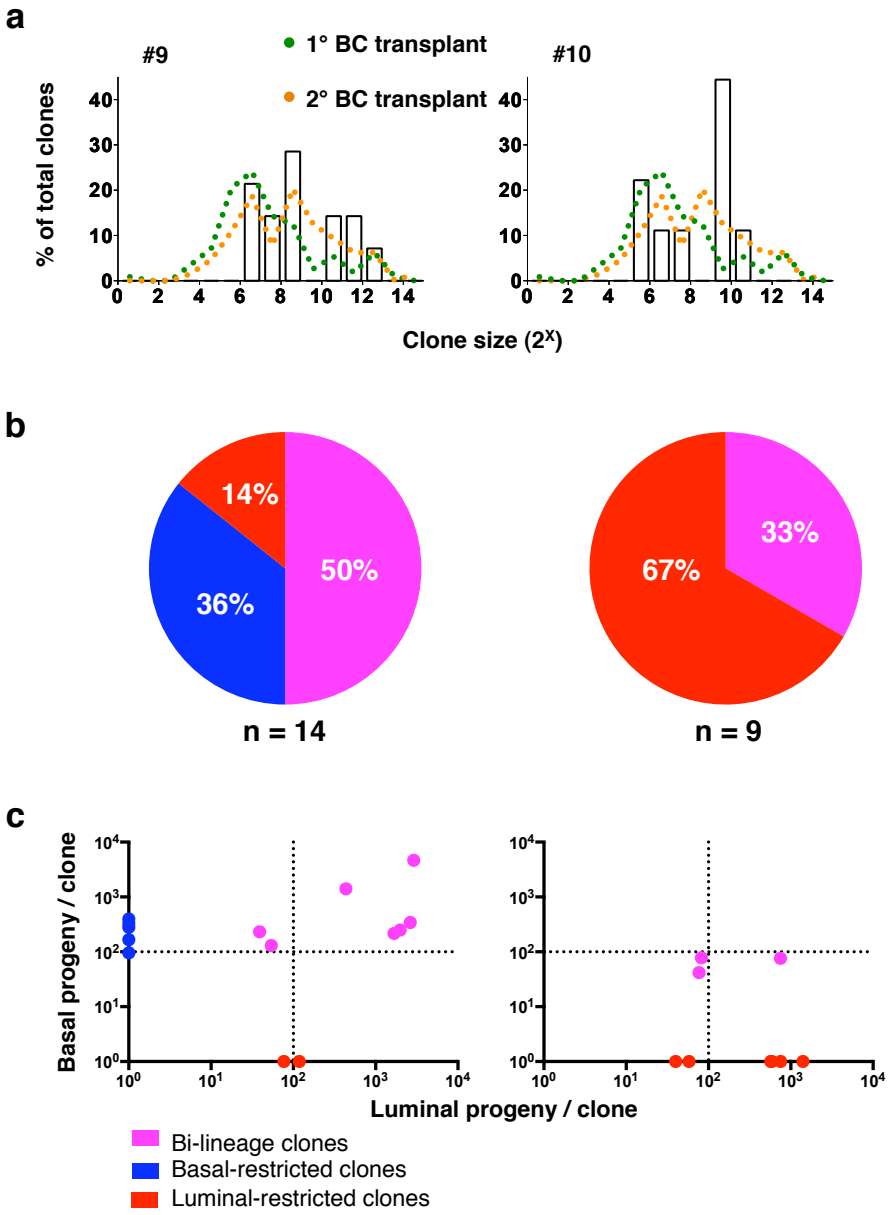
a
#9
#10
1° BC transplant
2° BC transplant
% of total clones
Clone size (2^X)
b
14%
50%
36%
n = 14
33%
67%
n = 9
c
Basal progeny / clone
Luminal progeny / clone
Bi-lineage clones
Basal-restricted clones
Luminal-restricted clones

Figure 3.16

Figure 3.16 Size and composition of clones detected in mice transplanted with barcoded LCs

(A) Distributions of total clone size (binned by $\log_2$ increments) for clones detected from fat pads #9 and #10, each transplanted with $1.3x10^5$ LCs. Dotted lines indicate the mode of each size distribution. Superimposed on these plots are also the size-distribution curves for clones detected from transplanted BCs in 1° (green) and 2° (orange) recipient mice. (B) Pie charts indicate the proportion of bi-lineage (magenta), luminal-restricted (red), and basal-restricted (blue) clones for fat pads #9 (left) and #10 (right). (C) The scatter plots below show numbers of basal and/or luminal progeny in each of the clones shown above (in B). In these, the dotted line indicates the minimum threshold of reproducible clone detection (100 cells).

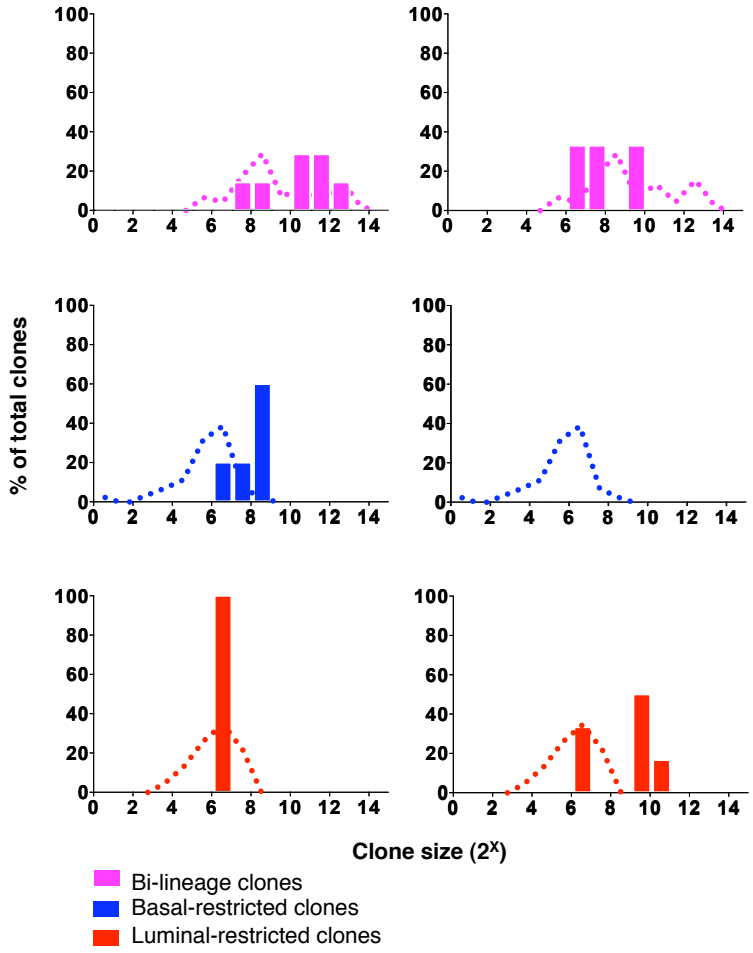

Figure 3.17 Size distribution plots for clone types obtained from transplanted LCs

Size distributions are shown for the three clone types, from fat pads #9 (left plots) and #10 (right plots). Bi-lineage, luminal-restricted and basal-restricted clones are shown in magenta, red, and blue, respectively. Dotted lines indicate the clone size distribution from primary BC transplants (Figure 3.9).

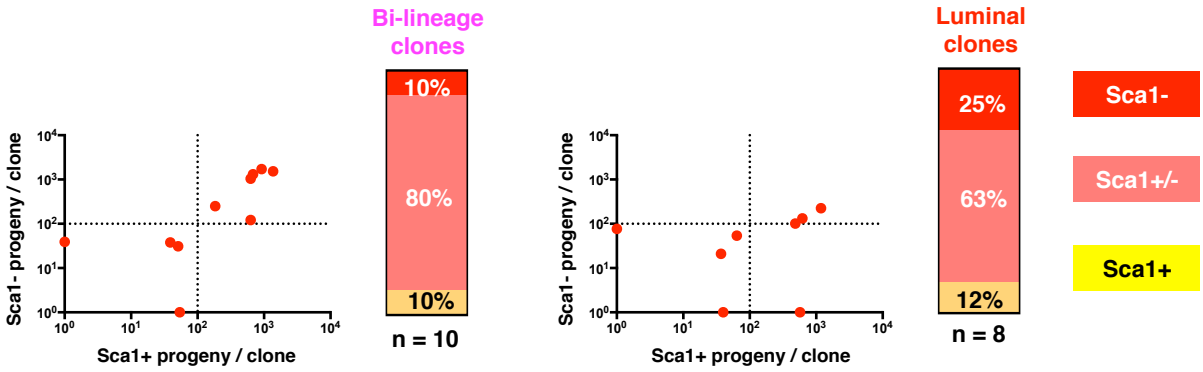

Figure 3.18 Sca1$^+$ and Sca1$^-$ luminal cell content of bi-lineage and luminal-restricted clones generated from transplanted LCs

The scatter plot shows the numbers of Sca1$^-$ (y-axis) and Sca1$^+$ (x-axis) cells from the luminal fraction of bi-lineage (left) and luminal-restricted clones (right), detected in primary mice transplanted with LCs (data pooled from fat pads #9 and #10). The stacked column shows the proportion of either the bi-lineage or luminal-restricted clones that were found to contain exclusively Sca1$^-$ cells (dark red), Sca1$^+$ cells (yellow) or both cell types (light red).

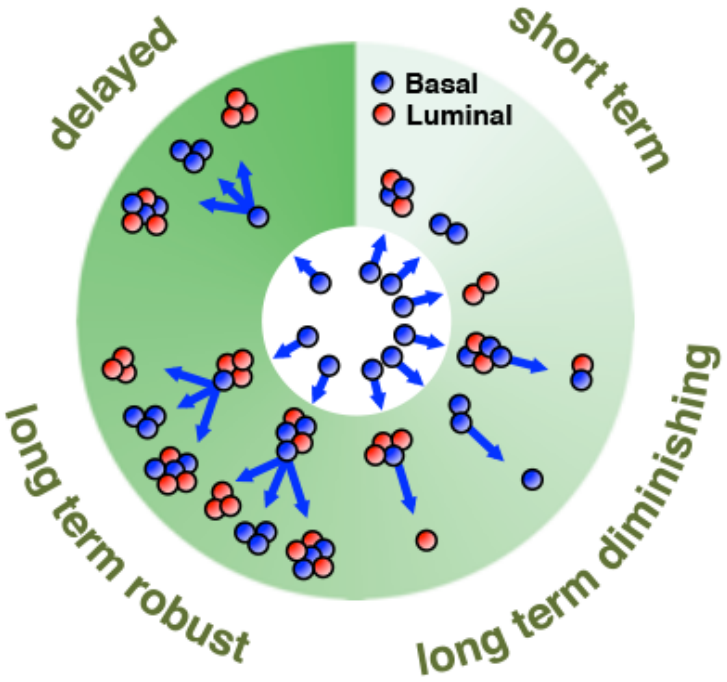

Figure 3.19 Heterogeneity of growth and differentiation potential of transplanted BCs

A model summarizing the different clonal patterns observed when non-limiting numbers of mouse mammary BCs are transplanted.

CHAPTER 4: ANALYSIS OF THE CLONAL GROWTH *IN VIVO* OF NORMAL, SPONTANEOUSLY TRANSFORMED, AND *DE NOVO* TRANSFORMED HUMAN MAMMARY CELLS

4.1 Introduction

The experiments in this chapter were designed to characterize first the individual patterns of growth and differentiation *in vivo* of co-transplanted *normal* human mammary cells, and then to determine how these clonal patterns might be perturbed upon spontaneous or directed transformation. Previous studies had demonstrated that normal human mammary cells with *in vivo* repopulating activity are largely restricted to the basal fraction (Eirew et al., 2008; Lim et al., 2009), although more recent experiments have demonstrated that rare LCs can also generate bi-lineage structures *in vivo*. However, the regenerative potential displayed by these LCs was found to be generally limited and not sustained when their progeny were serially transplanted (Eirew et al., 2008; Lim et al., 2009; Shehata et al., 2012). Therefore, in the present study of the regenerative activity *in vivo* of normal human mammary cells, we focused exclusively on the basal subset.

The cell output endpoint historically used in the subrenal transplant assay to infer the presence of an MRU in the input cells has been the presence of regenerated CFCs, revealed by plating the harvested cells *in vitro*. This endpoint exploits the ability of the secondary clonal assay to expand the direct progeny of the cells generated in the primary *in vivo* assay. This strategy increases the sensitivity of the assay while allowing it to retain objectivity required for an LDA to enable MRUs to be quantified. However, the animal requirements for assessing large numbers of individual MRU clones are prohibitive. The alternative use of a barcoding approach allows this practical barrier to be

circumvented. Coupling this approach with the use of a secondary culture to expand the clones generated *in vivo* would also be expected to increase the sensitivity of their detection. However, the accuracy of determining the sizes of the clones generated *in vivo* will necessarily be diminished, depending on the degree of variation in cell expansion obtained from the secondary CFCs prior to determining the barcode representation in their progeny. On the other hand, inferring clone sizes exclusively from cells harvested directly from the *in vivo* assay may miss clones that fall below the limit of detection. Without prior knowledge of either of these parameters, in initial experiments, we adopted an approach that employed both strategies. This involved deriving barcode sequence data from the progeny of cultured cells (regenerated CFCs), as well as directly from the initial cells harvested.

To investigate how the cell of origin can influence the phenotype and function of human breast tumours, we adopted a reverse genetic approach. Initially, we chose to use a combination of three mutant genes to try to favour the probability of obtaining some tumours from transduced human mammary epithelial BCs and LPs isolated at high purities from normal human reduction mammoplasty samples. The mutant genes used were selected based on their high prevalence in human breast tumours (*TP53* and *PIK3CA*) or RAS/ERK signaling pathway perturbation ($KRAS^{G12D}$), for example, due to mutations in MAP3K1, a serine/threonine kinase that acts upstream of ERK1/2, (Cancer Genome Atlas, 2012). *TP53* and *PIK3CA* mutations have been reported in approximately a third of all human breast tumours, with mutations in *TP53* more commonly seen in basal-like breast tumours (80%) as compared to *PIK3CA* mutations that are more commonly observed in the luminal or HER2-enriched subtypes (29-45%)(Cancer

Genome Atlas, 2012). In the present study, the deleterious missense $TP53^{R273C}$ mutation was used, since a missense mutation at this position within the DNA binding domain has been found to be prevalent in all breast cancer subtypes and acts to prevent its normal binding to p53 binding sites on DNA (Cho et al., 1994). Since the tetramerization domain remains functional, this mutant acts to sequester and render wild-type p53 non-functional thereby acting in a dominant negative fashion. The *H1047R* mutation in the kinase domain of *PIK3CA* was selected because it is the most common point mutation in this gene in all breast cancer subtypes, and its over-expression in normal cells has been found to increase the catalytic activity of the protein kinase and result in enhanced downstream signaling (Cancer Genome Atlas, 2012; Tikoo et al., 2012). $KRAS^{G12D}$ has already been reported to have oncogenic potential in primary normal human mammary cells, although its actual prevalence in human breast cancer appears to be fairly low (~5%)(Cancer Genome Atlas, 2012) compared to other cancers (up to 60% in pancreatic cancer)(Ling et al., 2012). For comparison, we used the same barcoding strategy to analyze the clonal growth behaviour *in vivo* of a spontaneous human breast cancer sample (pleural effusion) that had been previously successfully xenografted (once) and genotyped.

4.2 Materials and methods

4.2.1 Isolation of human mammary epithelial cell subsets

Normal human mammary epithelial cells were obtained from reduction mammoplasty samples that were initially dissociated by a sequence of mechanical and enzymatic procedures, prior to cryopreservation and storage at -156^{o}C, as described in Chapter 2. Also studied were cells isolated from a first xenograft generated from a pleural effusion

that had been obtained from a woman with advanced breast cancer (originally an ER^+ breast tumour). In all cases the dissociated cells were first centrifuged on Ficoll-hypaque and the low-density cells collected and cryopreserved. All samples were obtained with informed consent, as approved by the University of British Columbia Research Ethics Board.

BCs were isolated by FACS according to their $CD45^-CD31^-EpCAM^{lo}CD49f^+$ or $CD45^-CD31^-CD10^+CD90^+CD49f^+$ phenotype. LPs were isolated by FACS according to their $CD45^-CD31^-EpCAM^{hi}CD49f^+$ phenotype. *In vivo* regenerated BCs and LCs were isolated by FACS according to their $CD10^+CD90^+CD49f^+$ and $CD10^-CD90^-EpCAM^+$ phenotypes, respectively. Table 4.1 lists the fluorochrome-labelled antibodies used.

The malignant cells from the pleural effusion sample were thawed, washed with Hank's Balanced Salt Solution supplemented with 5% FBS, and immediately injected into the fat pad of an adult NSG mouse without any period of *in vitro* pre-culture.

4.2.2 Lentiviral transduction of human mammary epithelial cells and cellular barcoding analysis

These methods have been described in Chapter 2 (Section 2.2.5, 2.2.8, 2.2.9 and 2.2.10).

4.2.3 Preparation of oncogene-encoding lentiviral supernatants

Variations of the MNDU3-PGK-GFP lentiviral construct (as described in Chapter 2) were generated to encode for *YFP* or *mCherry* in place of the *GFP* reporter. $KRAS^{G12D}$, $PIK3CA^{H1047R}$ and $TP53^{R273C}$ mutant cDNAs were cloned into the constructs encoding *mCherry*, *YFP* and *GFP*, respectively, using flanking *AscI* and *PacI* restriction sites

downstream of the MNDU3 promoter (Figure 4.1A-C). Human *KRAS* cDNA was cloned from a human cell line, and altered by site-directed mutagenesis to obtain the *G12D* mutant. The $TP53^{R273C}$ mutant was cloned directly from a human cell line already harboring this mutation, and human $PIK3CA^{H1047R}$ cDNA was obtained from Dr. Andrew Weng (Terry Fox Laboratory, BC Cancer Agency). All cDNA clones were sequenced to confirm the presence of the desired point mutations, and absence of other mutations. After ligation into the lentiviral constructs, clones confirmed to contain the mutant genes in the correct orientation were selected for plasmid purification. Lentiviral supernatants were produced as described in Chapter 2, and their titres confirmed to be ~10^9 infectious units/ml.

4.2.4 Transplantation of human mammary cells into mice

An acidic mixture of concentrated rat tail collagen was prepared as previously described (Eirew et al., 2010; Eirew et al., 2008) and neutralized with NaOH prior to use. Normal or transduced human mammary epithelial cells were suspended in the neutralized collagen together with $2x10^5$ irradiated (15 Gy) C3H-10T1/2 mouse embryonic fibroblasts per 20 µl gel and the gels allowed to solidify at 37°C for 30 minutes. The gels were then implanted under the kidney capsule of 5 to 8 week-old virgin female nonobese diabetic severe combined immunodeficiency interleukin-2Rγc-null (NSG) mice, as previously described (Eirew et al., 2008). For transplants of normal human mammary cells, $1.5x10^4$ purified BCs or $3x10^4$ purified LPs were suspended in each gel. Human mammary BCs and LPs transduced with one or more of the mutant genes that were examined after 4 weeks were transplanted in duplicate, with each gel containing $1.5x10^4$

and $3x10^4$ cells per gel, respectively. The cells harvested later from duplicate gels were pooled for subsequent analysis. Transplants of similarly transduced cells examined after 2 weeks (for clonal analysis by barcoding) or after 6-8 weeks (for tumour formation) contained between $1.5x10^4$ to $1.4x10^6$ cells per gel, depending on the yield of BCs and LPs isolated from different mammoplasty samples.

4.2.5 2D *in vitro* CFC assays

Test cells were co-cultured with irradiated NIH-3T3 fibroblasts in serum-free SF-7 media in tissue culture dishes for 8-10 days, as previously described (Eirew et al., 2008).

4.2.6 Immunohistochemistry

Collagen gels or pieces of tumours obtained from mice were fixed in 10% buffered formalin (Fisher), washed in 70% ethanol, embedded in paraffin and 4 μm sections obtained. These were either stained directly with H&E, or first treated with Target Retrieval solution (DAKO) and then a cytomation serum-free protein block (DAKO) followed by staining with either an anti-cytokeratin 14 antibody, an anti-MUC1 antibody, an anti-SMA antibody, an anti-cytokeratin 8/18 antibody, an anti-ER antibody, an anti-Ki-67 antibody, an anti-HER2 antibody, or an anti-EGFR antibody. A secondary mouse antibody conjugated to alkaline phosphatase and treated with permanent red (DAKO) was used to obtain a positive pink staining, whereas a secondary rabbit antibody conjugated to horseradish peroxidase and treated with 3,3'-diaminobenzidine (DAB, DAKO) was used to obtain a positive brown staining. Table 4.1 provides details of the antibodies used and their sources.

4.2.7 Statistics

All reported p-values were calculated using the parametric unpaired Student's t-test.

4.3 Results

4.3.1 Serially co-transplanted human mammary epithelial BCs show diverse and sometimes highly delayed regenerative activities

To investigate the regenerative potential of normal human BCs, we isolated the $CD45^-CD31^-CD10^+CD90^+CD49f^+$ fraction at >97% purity from 3 different normal mammoplasty samples (Figure 4.2A), and then transduced each set with the barcoded lentiviral library described in Chapter 2 using the same 4-hour protocol as for the experiments with mouse mammary cells described in Chapter 3. In the first two experiments, the cells were then cultured for another three days prior to isolating the GFP^+ fraction (~35-38%) of the cells that had retained the original $CD45^-CD31^-CD10^+CD90^+CD49f^+$ phenotype (>95% of the cells under these conditions). $1.5x10^4$ of these FACS-sorted GFP^+ BCs from each experiment were then embedded into two collagen gels that were then implanted under the kidney capsule of two separate NSG mice. In a third experiment, the same number of cells ($1.5x10^4$) was embedded directly into each of two collagen gels directly after being exposed to virus (i.e., without further culture or selection) and both gels were then implanted into a single NSG mouse. In this latter experiment, a separate aliquot of cells was removed when the gels were being made and these cells were then cultured (as for routine colony assays) for 48 hours so that the transduction efficiency could be determined. The result was similar at 40% indicating

that $6x10^3$ GFP^+ cells had been embedded into each implanted gel in the third experiment. All gels were removed 4 weeks post-transplant and both human $GFP^+CD10^+CD90^+CD49f^+$ BCs and $GFP^+CD10^-CD90^-EpCAM^+$ and/or $MUC1^+$ LCs isolated by FACS from the cells obtained from the gels (again at >97% purity, Figure 4.2B, ~$3.5x10^3$ GFP^+ cells in each experiment, Table 4.2). An aliquot of each of these cell suspensions was then analyzed directly for its barcode content, and another after amplifying the initially harvested cells *in vitro* (Figure 4.3). In the first experiment, 40% of the recovered BCs were also set aside and transplanted in collagen gels under the kidney capsule of secondary NSG mice. The cells recovered from these secondary gels another 4 weeks later were then also analyzed for the barcode sequences they contained (Figure 4.3, Figure 4.2B). Histological and immunochemical analysis of gels containing similarly transduced cells confirmed the expected organization and cellular content of the structures regenerated (Figure 4.4).

Barcode analysis was performed using the spiked-in control method described in Chapter 2 (the first two experiments were analyzed on the HiSeq in MPS run #1, and the third experiment on the MiSeq in MPS run #3). In the first 2 experiments, a total of 39 clones were detected (Figure 4.5) with the majority of these represented in the cells that were expanded *in vitro* before being sequenced (37 of the 39 clones). Only 4 clones were detected in the cells analyzed directly, 2 of which were also detected in the *in vitro*-expanded cells. The low frequency of clones detected from the cells analyzed directly, and their calculated sizes, indicate that the primary xenograft was made up of very small clones, below the limit of detection (Table 4.2). In the third experiment, all of the harvested cells were expanded *in vitro* prior to analysis, which allowed a total of 90

clones to be detected. 50% of these were represented in the *in vitro*-expanded BCs and all were represented in the *in vitro*-expanded LCs (Figure 4.5 and Table 4.3). Basal-restricted, luminal-restricted and bi-lineage clones were observed in all three experiments. However, their distributions were highly variant, even when restricted to an analysis of *in vitro*-expanded progeny (5:89:5 for xenograft #1, 22:17:61 for xenograft #2 and 33:37:30 for xenograft #3, respectively, Figure 4.6).

In the first experiment in which secondary transplants were performed, we identified 17 clones. Interestingly, all of these were significantly larger than, and were not detected among, the clones evident in the primary grafts (an average of 246 per secondary clone vs. 35 cells per primary clone, $p = 0.018$; Figure 4.7). Thus, from all three experiments in which an estimated $3.6x10^4$ BCs were tested, a total of 146 clones were detected. This corresponds to a frequency of ~1 clone per 500 purified BCs. This value is just slightly higher than that reported for the frequency of MRUs in bulk mammoplasty cells measured by LDA (1 in 10^3 to 10^4)(Eirew et al., 2008), based on an estimated ~30% content of BCs in bulk mammoplasty cells (Eirew et al., 2008; Lim et al., 2009).

4.3.2 Both normal human mammary BCs and LPs are readily susceptible to transformation

To determine if tumours could be generated *de novo* by transducing normal human BCs and LPs with defined oncogenes, $CD45^-CD31^-EpCAM^{lo}CD49f^+$ BCs and $CD45^-CD31^-EpCAM^{hi}CD49f^+$ LPs were isolated by FACS (Figure 4.2C) and then 0.3 - $14x10^5$ cells of each subset exposed simultaneously to three different MPG lentivirus preparations,

each virus encoding a different mutant gene and a different downstream fluorescent reporter (*KRAS*^{G12D}*-mCherry*, *PIK3CA*^{H1047R}*-YFP*, and *TP53*^{R273C}*-GFP*, Figure 4.1). The cells were then embedded in collagen gels with irradiated fibroblasts, and transplanted as described above for normal mammary basal cells. After 6-8 weeks, 8 (44%) and 12 (67%) of the 18 different paired samples tested produced tumours from transduced BCs and LPs, respectively (Figure 4.8, Table 4.4). Most of the tumours resembled invasive ductal carcinomas, were highly mitotic (and Ki67^{+}, Figure 4.9), and had invasive margins. A more detailed description of their immunohistological features is presented below in Section 4.3.5, and shown in Figure 4.9.

Given the high frequency of tumour formation from both subsets of normal human mammary cells, we next investigated whether all three mutant genes are required for tumorigenesis. Accordingly, the same experiment was undertaken again, but using all possible virus combinations for their transduction. The results showed that appreciable numbers of similar tumours appeared after 6-8 weeks when either BCs and LPs were transduced with *KRAS*^{G12D} alone (2/5 and 6/6, respectively), or *KRAS*^{G12D} in combination with *PIK3CA*^{H1047R} (3/6 and 4/6, respectively) or *TP53*^{R273C} (3/6 and 6/6, respectively, Table 4.4).

4.3.3 Barcoding human mammary BCs and LPs prior to their transformation reveals a high frequency of transformants but different growth behaviours

To investigate and compare the clonal composition of tumours obtained from the various oncogene-transduced BCs and LPs, we added the barcoded lentiviral supernatant to the media in which the cells were being transduced together with the oncogenic viruses in

some of the experiments. From these, we obtained one tumour each from BCs transduced with all 3 genes, $KRAS^{G12D}$ alone, or $KRAS^{G12D}$ in combination with $PIK3CA^{H1047R}$ and/or $TP53^{R273C}$. Another 21 tumours were obtained from similarly transduced LPs (8, 5, 4 and 4 in each of the four conditions, respectively). 20-50% of these tumours, depending on the size of the tumour, were taken for barcode analysis. Analysis of this barcode data was performed similarly to the experiments generated on the Illumina MiSeq, with a new normalization curve generated specifically for this dataset that contained spiked-in controls spanning a range of 20 to 6,250 cells. This produced a linear correlation (R^2 = 0.99) between the fractional read representation (normalized to the sum of the reads from the 20-, 100-, 250- and 500-cell spiked-in controls), and their respective input number of cells (Figure 4.10). From 5 sets of spiked-in controls, the sensitivity of detecting clones consisting of 100 cells or more was 5 of 5 (100%), whereas for 20 cells, it was 2 of 5 (40%). The specificity for discriminating true clones from background was 100% when a threshold of 100 cells was used. However, for these tumour samples, a threshold was not applied, since the majority of the clones they contained consisted of more than 500 cells each. Although many of the smaller spiked-in controls were not detected with the sequencing depth used, the fact that most of the experimental clones detected were considerably larger than the spiked-in controls provides confidence that they were real and accurately quantified.

The barcode data indicates that the number of clones present in the tumours derived from BCs was, on average, 1 per 500 input BCs, when all three oncogenic viruses were used to generate the tumour. The corresponding number of clones present in the tumours derived from LPs was 2-fold lower than for the BC-derived tumours (1 per 10^3

input LPs, Table 4.5). Interestingly, the number of clones present in the BC and LP-derived tumours was higher when *KRAS*^{G12D} was used alone or in combination with either *PIK3CA*^{H1047R} or *TP53*^{R273C}, though the number of clones was still consistently 2-fold lower for the LP compared to the BC-derived tumours. However, the size of the clones present in all of the tumours derived from LPs included many that were larger than the largest clones present in the BC-derived tumours (up to 300-fold larger at $9x10^6$ cells/clone as compared to $3x10^4$ cells/clone, respectively, Table 4.6 and Figure 4.11).

4.3.4 Oncogene-transduced normal human mammary BCs and LCs also show different premalignant growth behaviour

We next sought to investigate whether the oncogene-transduced cells would show differences in their growth trajectories prior to the formation of a palpable tumour. Accordingly, we initiated another series of transduction experiments using the same protocol, but, in this case, sacrificed the mice after 4 weeks and removed the gels for analysis at that time. Specifically, $1.5x10^4$ BCs and $3x10^4$ matching LPs were transduced with all possible combinations of the three oncogenes, and then we determined the total cell and CFC outputs at the 4 week timepoint (Figure 4.12). Controls were also set up with matched *GFP*-transduced normal BCs and LPs (Figure 4.13) adjusted to represent $1.5x10^4$ and $3x10^4$ positively transduced cells, respectively. The transduction efficiency measured on an aliquot of the transduced cells maintained in culture for 72 hours was 89% ± 10% (average ± SD) for a single vector. These average "control" total cell and CFC outputs were then used to calculate the fold-change of the corresponding values for each of the "test" transduction groups (Figure 4.14 for total cells and Figure 4.15 for

CFCs). The transduction efficiency for all single oncogene vectors was the same, and for two or three vectors used in combination was 76% ± 17% (average ± SD) and 62% ± 18%, respectively. Analysis of the cell outputs by flow cytometry confirmed the correct expression of *GFP*, *YFP* and *mCherry* for each corresponding transduction condition (Figure 4.16 and Table 4.7). Noticeably, however, the total cell output from transplanted LPs was only sufficient to detect positively transduced cells under conditions that stimulated the growth of the cells *in vivo* ($KRAS^{G12D}$ alone, $KRAS^{G12D}$+$TP53^{R273C}$, and $KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$, Table 4.7).

Of the three genes expressed in BCs and LPs individually, only $KRAS^{G12D}$ resulted in a statistically significant increase in total cell output from both starting cell types ($p = 0.043$ and 0.048, respectively, Figure 4.14). The same trend was seen in the total CFC outputs, but this did not achieve significance (Figure 4.15). However, when LPs were transduced with $KRAS^{G12D}$+$TP53^{R273C}$, statistically significant expansions in both total cell and CFC outputs were observed ($p = 0.0024$ and $p < 0.0001$, respectively, Figure 4.14 and 4.15). Interestingly, similarly transduced BCs showed no effect on these parameters compared to control transduced cells. Further, when all three mutant genes were used, results for either LPs or BCs were not statistically significant compared to the controls. These results suggest that some of the more potent conditions for transformation of BCs and LPs ($KRAS^{G12D}$ alone for both cell types, or $KRAS^{G12D}$+$TP53^{R273C}$ for LPs) exhibit significant changes in their growth potential that can be detected as early as 4 weeks. However, for other conditions that also produce tumours at an appreciable frequency ($KRAS^{G12D}$+$PIK3CA^{H1047R}$ or $KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$ from either

cell type, and $KRAS^{G12D}$+$TP53^{R273C}$ for BCs only), changes in their growth control must not manifest themselves until later.

We also asked whether changes in the clonal composition of the oncogene-transduced cells might be detectable as early as 2 weeks after their transplantation *in vivo*, again by barcoding the starting cells at the time of exposure to the oncogenic viruses. Consistent with previous findings that MRUs are restricted to the BCs, the frequency of clones obtained *in vivo* from control LPs, even after 2 weeks, was ~4-fold lower than for the BCs (Table 4.8). However, by that time, the number of clones present was already consistently increased (~2-fold) for the LPs transduced with any of the genes alone or in combination. In contrast, the generation of clones from the BCs was not affected (Table 4.8). On the other hand, the barcode data revealed the presence of some significantly larger clones in the gels seeded with either BCs or LPs transduced with $KRAS^{G12D}$ alone (~$3x10^3$ cells/clone) by comparison to matched *mCherry*-transduced normal cells (~100 cells/clone, $p < 0.0001$ Figure 4.17A-D), although this was the only situation where such a perturbation could be detected.

These experiments suggests that the timing and/or extent of changes in growth rate may be different and discernable for different transduction conditions, and cell of origin, even before a tumour is overtly detectable.

4.3.5 Unexpected cell of origin-related differences in tumour phenotype

Immunohistochemical analysis of a first limited series of the *de novo* generated tumours showed that 3 of 5 that originated from BCs transduced with all three genes expressed high levels of ERα, whereas none of six tumours derived from LPs were $ERα^+$ (Figure

4.8). To determine whether this change in phenotype occurs early in transformation, we examined the surface marker profiles of xenografts of oncogene-transduced cells removed after just 4 weeks *in vivo*. The results suggest that $KRAS^{G12D}$ alone increases the outputs from both BCs and LPs of cells with basal and luminal features ($p = 0.014$ and 0.003, and $p = <0.0001$ and 0.0001, respectively, Figure 4.18). The outputs of cells with basal and luminal features from LPs transduced with $KRAS^{G12D}$+$TP53^{R273C}$ and $KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$ were also significantly higher than the corresponding outputs of such cells from transplanted control LPs ($p = 0.002$ and 0.02, and $p = 0.007$ and 0.01, respectively) but not the matching transduced (or control) BCs.

4.3.6 Cellular barcoding of a primary human breast cancer xenograft reveals replicated clonal diversity

As a comparison, we next examined the diversity of clones derived in different mice transplanted with cells obtained from the same patient's breast cancer. For this purpose, we obtained cells from a primary xenograft of tumour cells produced from a pleural effusion that had been obtained from a patient with advanced breast cancer who had originally been diagnosed with an ER^+ breast tumour. The tumour produced in the fat pad of the first NSG recipient of the pleural effusate cells did not appear until ~1 year later, although retrospectively it was found that the actual tumour content of the cells injected was quite low. The resulting primary tumour xenograft that eventually appeared was mechanically dissociated, transduced with the MPG barcoded lentiviral library, and 10^6 cells injected subcutaneously (in 50% matrigel) into each of two secondary NSG mice. After three months, palpable tumours appeared, and 20% of each was analyzed for its

barcode content using the "spiked-in" method described in Chapter 2 (corresponding to MPS run #3 on the MiSeq).

From this analysis, we identified 418 and 454 uniquely marked clones in the two respective "secondary" tumours. These clones ranged in size from ~10^2 to 3x10^5 cells each with a median of ~300 cells per clone (Figure 4.19). Thus, from 10^6 initially transplanted cells (90% transduction efficiency), the tumour-propagating cell frequency determined independently in two mice was very similar at ~1 in 2,000 cells injected. These results are consistent with previous reports that human breast tumour xenografts in mice are propagated by a relatively rare subset of the cells present in the tumour (Al-Hajj et al., 2003).

4.4 Discussion

The experiments performed in this chapter made use of cellular barcoding to compare the regenerative activity of human mammary cells from various normal and transformed sources at clonal resolution. Remarkably, the values obtained for the frequency of cells with *in vivo* growth potential was consistently within the same range. This included results for both BCs and LPs at all stages of transformation, from normal (1/500) to premalignant (~1/200), to *de novo* tumorigenic ~~1/300 to 1/1,000) to a patient's primary xenograft (1/2,000). The lower clonal frequency from the patient xenograft may be due to low tumour cellularity of the transplanted material, since the other two experiments used FACS-purified cells. It is thus inviting to speculate that the clones that contribute to *de novo* tumours upon oncogene transduction, at least in the models used in this study, are those that might normally display regenerative activity upon transplantation. Future

experiments, possibly using inducible vectors would be helpful to pursue this question further. However, we have also demonstrated here that clones detected in secondary transplants of normal BCs are not commonly present in the same clones that achieve a detectable size within 4 weeks in a primary xenograft. It also remains unclear whether these differences reflect the different growth properties of biologically distinct subsets of cells or whether they simply reflect the operation of stochastic responses or reversible behaviours –issues of potential relevance to their ability to be transformed.

Intriguingly, the sizes of the clones identified in the xenografts derived from a patient's malignant breast cells (originally from an ER^+ tumour) were within the same range as the tumours generated *de novo* from BCs in our transduction model (~10^2 to $3x10^5$ cells/clone). This value is smaller than the size of the clones in tumours that we obtained from LPs in the same model (~$2x10^2$ to $9x10^6$ cells/clone). This data is in accordance with our observation that, in this model, ER^+ tumours were also derived from BCs. However, additional primary patient tumours will certainly need to be investigated to determine how closely the features of those generated in our *de novo* model mimic those that appear in patients. The fact that the tumours generated *de novo* arose extremely quickly (within 6-8 weeks), whereas those that develop in patients are thought to require months to years before they become clinically evident (Perou et al., 2000; Sorlie et al., 2001) already indicates at least one major difference. Nevertheless, it is interesting to note that in mouse models, mutations in *PIK3CA* alone produce tumours in all affected mice but with a >12 month latency period (Tikoo et al., 2012) and in our transduction model, only one tumour was obtained within 8 weeks in 10 experiments when either normal human BCs or LPs were transduced with $PIK3CA^{H1047R}$ only; whereas the same

cells transduced with a virus encoding $KRAS^{G12D}$ produced rapidly appearing tumours at a much higher frequency in both BCs and LPs (2/6 and 6/6, respectively).

Whole genome sequencing data was also obtained (at a depth of 50-fold coverage) from the original pleural effusion sample used here, as well as a non-barcoded but similarly passaged secondary xenograft, in order to obtain information about the genotypic complexity of the cells and their changes with sequential passaging (as reported for sample "SA429", by Eirew et al., submitted). This analysis revealed a total of 216 single nucleotide coding variants (SNVs) with varying prevalence in the original pleural effusion compared to the secondary passaged xenograft consistent with the presence of 9 clonal groupings. Comparing this genotypic clonality data to the hundreds of clones detected from the barcoding data reveals that there are significantly more functional (barcoded) clones than were resolved by genomic analysis. This result is not unexpected assuming some genomic "subclones" would contain many tumorigenic cells that would each generate <1% of the whole tumour (Figure 4.19), a depth of detection not feasible by whole genome sequencing. Such genomic sequencing studies define genotypic clones by grouping SNVs that have similar allelic prevalences (Eirew et al., submitted). However, whether these genotypic clones in fact correspond to the observed functional clones requires validation that the grouping of SNVs representing a particular genotypic clone can indeed be found in individual cells. This would require single cell sequencing, a level of investigation that has not yet been reported for human breast tumours. It is thus unclear how the many functional clones shown here relate to the clonal groups identified by genomic analysis. However, the diversity of growth and regenerative activity demonstrated by these barcoded clones suggests that cells within a genotypic

clone can have different functional activities that may be differentially regulated by environmental factors and/or epigenetic modifications.

The model we have established and the use of barcoding to monitor changes in clonal composition will be especially useful for understanding early aspects of breast cancer development and perhaps in other tumours as well. In particular, access to clonal data for tumours being generated from defined human mammary epithelial cell types isolated directly from normal tissue will enable many of the caveats of previous models with immortalized human cell lines or cells from mice to be circumvented.

4.5 Figures & tables

Table 4.1 Antibodies used for FACS sorting and immunohistochemistry

Antibodies used for FACS sorting:

Antibody	Fluorophore	Clone	Company
Mouse anti-human CD45	AlexaFluor 488	HI30	Biolegend
Mouse anti-human CD31	AlexFluor 488	WM59	Biolegend
Mouse anti-human CD10	Phycoerythrin	H110a	BD Pharmingen
Mouse anti-human CD90	Phycoerythrin	eBio5E10	eBioscience
Rat anti-human CD49f	allophycocyanin	GoH3	R&D Systems
Mouse anti-human CD227/MUC1	N/A	214D4	STEMCELL Technologies
Goat anti-mouse secondary	AlexaFluor 700	N/A	Molecular Probes
Mouse anti-human CD326	PerCP-Cy5.5	9C4	Biolegend

Antibodies used for immunohistochemistry:

Antibody	Species reactivity	Concentration	Clone	Company
Mouse anti-CK14	human and mouse	1 in 50	NCL-LL002	Novacastra
Mouse anti-MUC1	human	1 in 50	214D4	STEMCELL Technologies
Mouse anti-SMA	human and mouse	1 in 100	1A4	DAKO
Mouse anti-CK8/18	human	1 in 50	Zym5.2	Invitrogen
Rabbit anti-ERα	human	1 in 50	SP1	ThermoScientific
Rabbit anti-Ki67	human	1 in 300	SP6	ThermoScientific
Rabbit anti-HER2	human	Neat	4B5	Ventana
Rabbit anti-EGFR	human	1 in 50	EP22	Epitomics

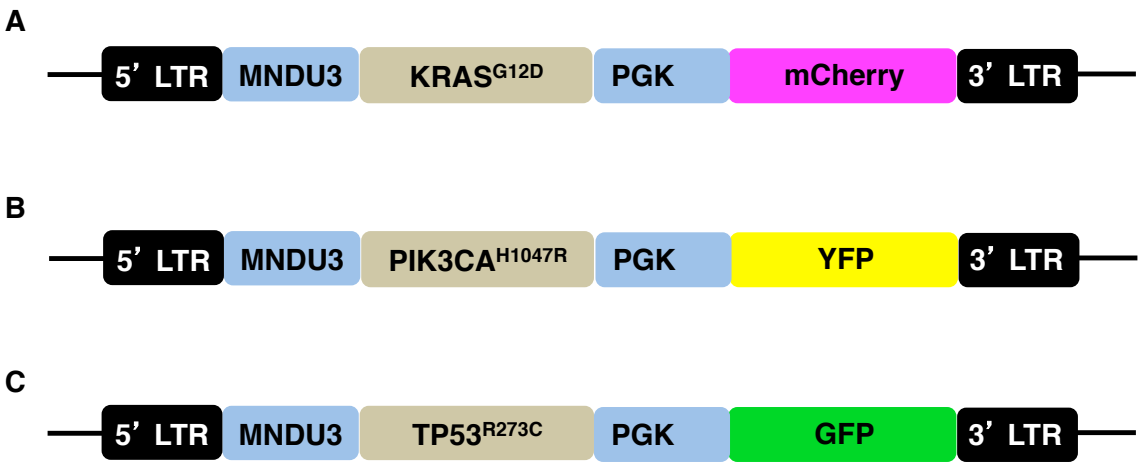

Figure 4.1 Lentiviral constructs encoding *KRAS*^{G12D}, *PIK3CA*^{H1047R} and *TP53*^{R273C}

The lentiviral constructs shown here were derived from the original MPG lentiviral vector described in Chapter 2. Expression of genes *KRAS*^{G12D} (A), *PIK3CA*^{H1047R} (B) and *TP53*^{R273C} (C) are driven by the *MNDU3* promoter whereas their corresponding fluorescent reporters are driven by the *PGK* promoter.

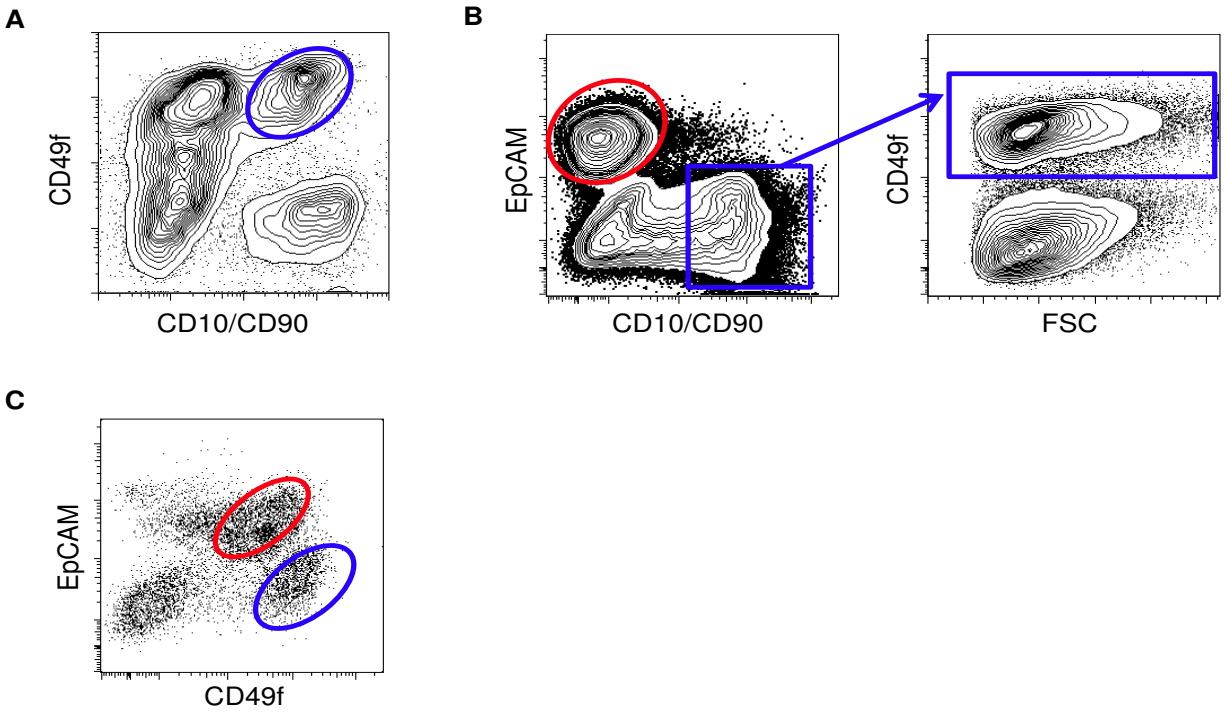

Figure 4.2 Strategy for isolating human BCs and LPs/LCs by FACS

(A) Human BCs were isolated from freshly dissociated normal human mammoplasty samples according to their $CD10^{+}CD90^{+}CD49f^{+}$ phenotype (blue oval gate), after depleting hematopoietic and endothelial cells using CD45 and CD31, respectively. (B) Regenerated BCs were isolated as in (A, blue rectangular gates), and LCs according to their $EpCAM^{+}CD10^{-}CD90^{-}$ phenotype (red oval gate). (C) In a second experiment, human BCs and LPs were isolated from freshly dissociated mammoplasty samples according to their $EpCAM^{lo}CD49f^{+}$ (blue oval gate), and $EpCAM^{hi}CD49f^{+}$ (red oval gate) phenotypes, after similarly depleting hematopoietic and endothelial cells.

Table 4.2 Estimated numbers of clones in the first two xenografts of normal human cells that would be below the set detection threshold

Cell type analyzed	No. of cells detected above threshold	No. of cells sampled for sequencing	Estimated % of clones below threshold	Estimated no. of clones below threshold (if assuming 100 cells/clone)
Basal	**22**	**1,376**	**98**	**14**
Luminal	**105**	**2,000**	**95**	**19**

The total number of cells in clones detected above the designated threshold and the total number of cells sampled for sequencing were used to calculate the number of cells from clones falling below the designated threshold (and thus not detected). To obtain an estimate of the minimum number of clones found below the threshold, it was assumed that these clones contained ~100 cells - the smallest clone size that was reproducibly detected.

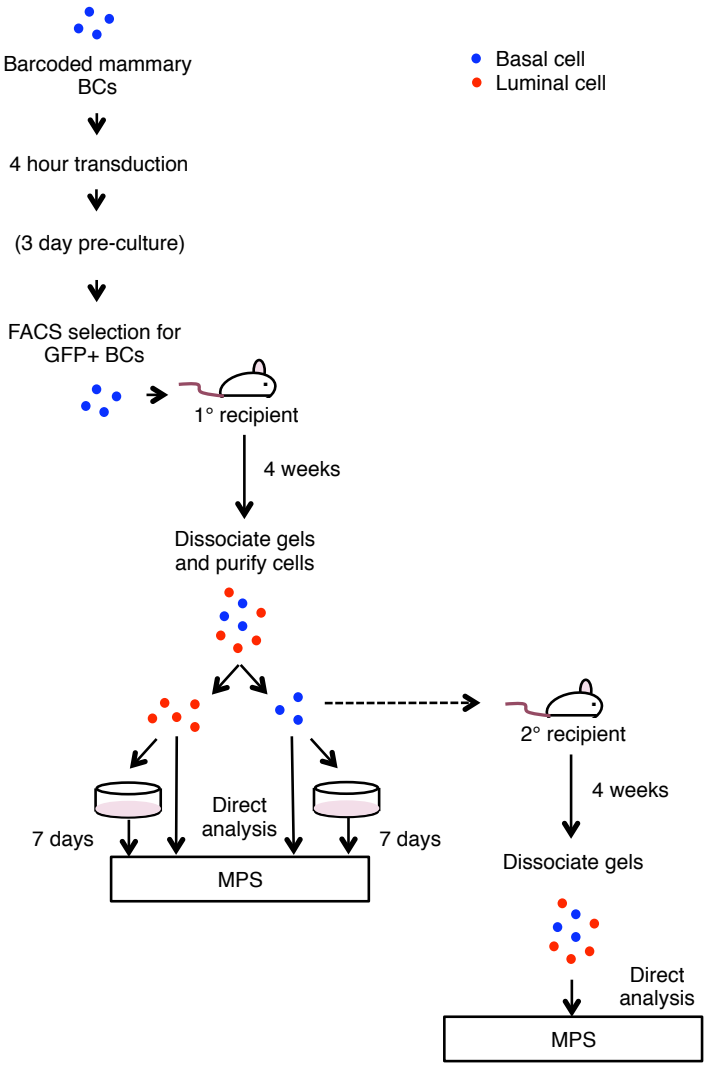

Figure 4.3 Experimental design used to track the regenerative activity of human BCs *in vivo*

Primary human BCs (blue circles) were barcoded and embedded into collagen gels with irradiated fibroblasts, with or without a 4-day period *in vitro* beforehand. The gels were then transplanted under the kidney capsule of NSG mice. After 4 weeks, the gels were retrieved, the regenerated cells dissociated and sorted into GFP^+ BC (blue circles) and LC (red circles) fractions, of which the proportions indicated were taken for barcode analysis immediately, or after being expanded *in vitro* for 7-days. In one experiment, 50% of the

BCs isolated from the primary transplant were transplanted into a secondary mouse from which all of the cells regenerated were then analyzed directly, without prior culture.

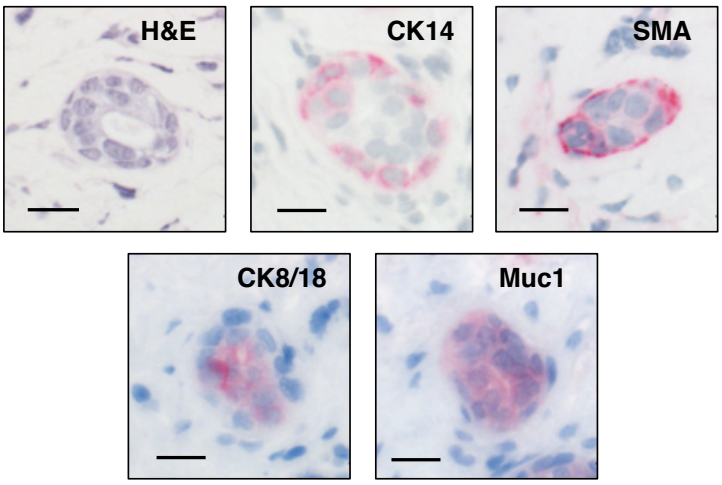

Figure 4.4 Histology of mammary structures regenerated *in vivo* from transplanted human BCs

Histology of bi-layered structures generated from barcoded human basal mammary epithelial cells in collagen gels xenografted under the kidney capsule of NSG mice. The top left panel shows H&E staining. The other panels show immunohistochemical staining for CK14 and SMA - markers of BCs, and CK8/18 and MUC1 - markers of LCs. Positive staining is shown as pink, and all sections were counter-stained with hematoxylin. Black scale bars indicate 20 μm.

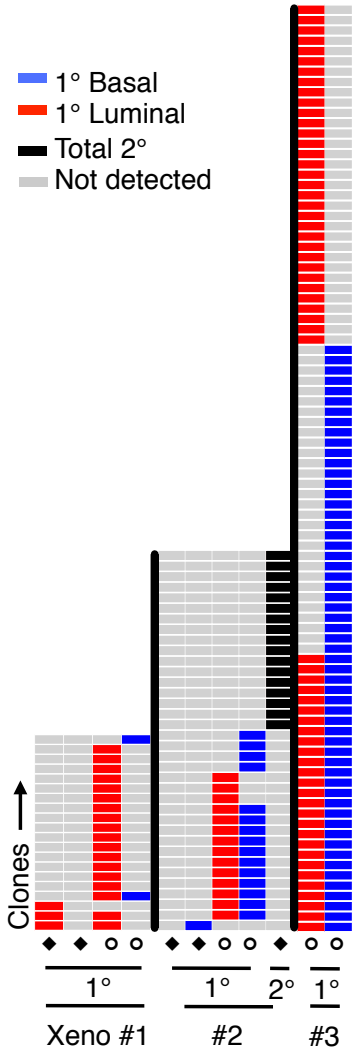

Figure 4.5 Detection of barcoded clones by MPS from transplanted normal human BCs

A depiction of all clones detected in the three primary xenografts (#1 to #3) and in the single secondary (2°) xenograft. The y-axis is the clone number ID within each xenograft analyzed. The columns refer to the clones detected in the cells isolated directly from the *in vivo* assay (◆) or after a further 7 days of cell expansion *in vitro* (O).

Table 4.3 Clones detected in primary and secondary transplants of human BCs

Patient 1:

Clone	BCs	LCs	Basal CFC	Luminal CFC	Secondary clones
1	0	0	508	10,611	0
2	0	0	456	0	0
3	0	0	0	6,912	0
4	0	0	0	5,253	0
5	0	0	0	4,841	0
6	0	0	0	4,624	0
7	0	12	0	4,321	0
8	0	53	0	3,814	0
9	0	0	0	3,323	0
10	0	0	0	2,962	0
11	0	0	0	2,724	0
12	0	0	0	2,624	0
13	0	0	0	2,541	0
14	0	0	0	1,854	0
15	0	0	0	1,748	0
16	0	0	0	1,667	0
17	0	0	0	1,554	0
18	0	0	0	1,274	0
19	0	0	0	1,042	0
20	0	0	0	805	0
21	0	0	0	376	0
22	0	40	0	0	0
23	0	0	0	0	577
24	0	0	0	0	461
25	0	0	0	0	383
26	0	0	0	0	337
27	0	0	0	0	332
28	0	0	0	0	316
29	0	0	0	0	235
30	0	0	0	0	227
31	0	0	0	0	221
32	0	0	0	0	193
33	0	0	0	0	178
34	0	0	0	0	149

(Table continued on subsequent page…)

Clone	BCs	LCs	Basal CFC	Luminal CFC	Secondary clones
35	0	0	0	0	144
36	0	0	0	0	133
37	0	0	0	0	115
38	0	0	0	0	105
39	0	0	0	0	80

Patient 2:

Clone	BCs	LCs	Basal CFC	Luminal CFC
40	0	0	923	6,390
41	0	0	535	0
42	0	0	334	2,229
43	0	0	234	1,579
44	0	0	208	0
45	0	0	204	1,541
46	0	0	155	1,277
47	0	0	137	1,158
48	0	0	108	0
49	0	0	75	0
50	0	0	37	751
51	0	0	55	842
52	0	0	16	644
53	0	0	0	530
54	0	0	0	394
55	0	0	0	174
56	22	0	0	0

Patient 3:

Clone	Basal CFC	Luminal CFC
1	29,738	0
2	25,958	3,367
3	21,761	0
4	20,810	7,995
5	20,589	0
6	18,913	2,772

(Table continued on subsequent page…)

Clone	Basal CFC	Luminal CFC
7	17,119	0
8	14,000	0
9	11,972	0
10	9,989	75
11	8,601	0
12	8,537	3,132
13	7,692	0
14	7,657	47,735
15	7,071	0
16	6,378	2,578
17	6,259	22
18	5,547	5,109
19	5,330	0
20	4,954	2,283
21	4,899	0
22	4,786	3,255
23	4,724	0
24	4,592	2,177
25	4,521	0
26	4,430	0
27	4,247	2,468
28	4,200	2,716
29	3,974	97
30	3,754	1,982
31	3,393	38,107
32	3,384	0
33	3,371	1,710
34	3,113	0
35	2,717	47
36	2,491	1,043
37	2,253	0
38	1,943	0
39	1,572	112
40	1,221	0
41	1,075	538
42	993	0
43	826	0

(Table continued on subsequent page…)

Clone	Basal CFC	Luminal CFC
44	753	147
45	717	0
46	705	0
47	664	369
48	568	0
49	481	0
50	447	582
51	434	0
52	370	0
53	266	0
54	250	3,610
55	240	896
56	198	0
57	140	0
58	0	41,253
59	0	33,640
60	0	28,449
61	0	27,619
62	0	27,581
63	0	23,620
64	0	8,192
65	0	8,186
66	0	4,740
67	0	3,521
68	0	2,888
69	0	2,446
70	0	2,183
71	0	2,091
72	0	1,920
73	0	1,855
74	0	1,701
75	0	1,698
76	0	1,694
77	0	1,274
78	0	1,174
79	0	1,125
80	0	916

(Table continued on subsequent page…)

Clone	Basal CFC	Luminal CFC
81	0	698
82	0	610
83	0	579
84	0	552
85	0	490
86	0	389
87	0	329
88	0	247
89	0	179
90	0	134

All clone sizes have been normalized to total BCs or LCs isolated from the xenograft. Values for "CFCs" are in cell number, as detected from sequencing of the cells expanded *in vitro* from cells isolated from the *in vivo* transplant, normalized to total BCs or LCs isolated from the transplant. Secondary transplants were performed only on those mice or patient samples indicated. Clones identified from the secondary transplant of the regenerated basal cells from Patient 1 are reported as total clone size.

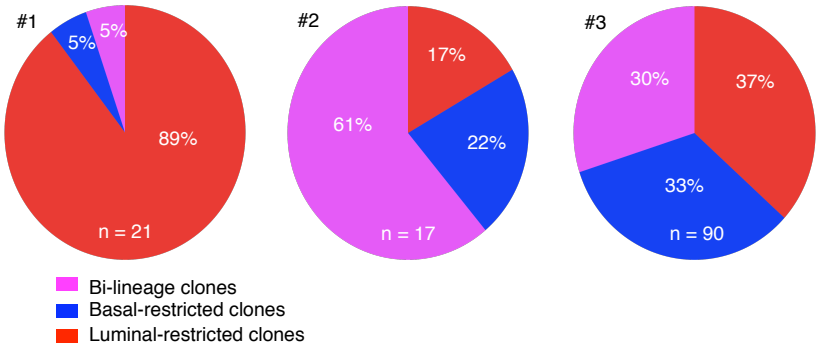

Figure 4.6 Representation of different types of clones in culture-expanded cells from primary xenografts

Each pie chart indicates the proportions of clone types defined according to their content of BCs and/or LCs, and/or their progeny generated *in vitro* ('n' is the total number of clones detected in directly analyzed cells and in *in vitro*-expanded cells). Bi-lineage, luminal-restricted and basal-restricted clones are shown as magenta, red and blue, respectively. The number on the top left of each graph corresponds to the xenografts in Figure 4.4.

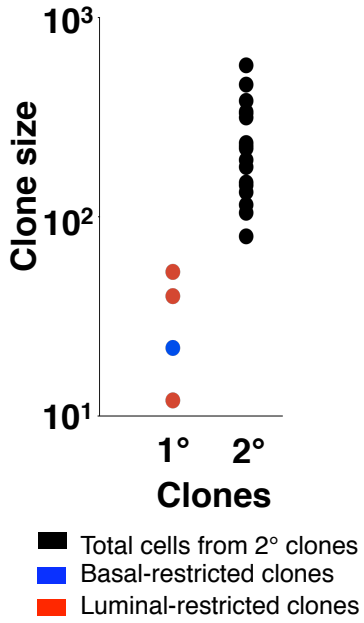

Figure 4.7 Comparison of the size of clones generated in primary and secondary xenografts of normal human BCs

Shown is the total size of each clone measured in primary (1°) and secondary (2°) transplanted mice derived from xenograft #1 (1° and 2°) and 2 (1° only) (basal = blue, luminal = red, total cells = black).

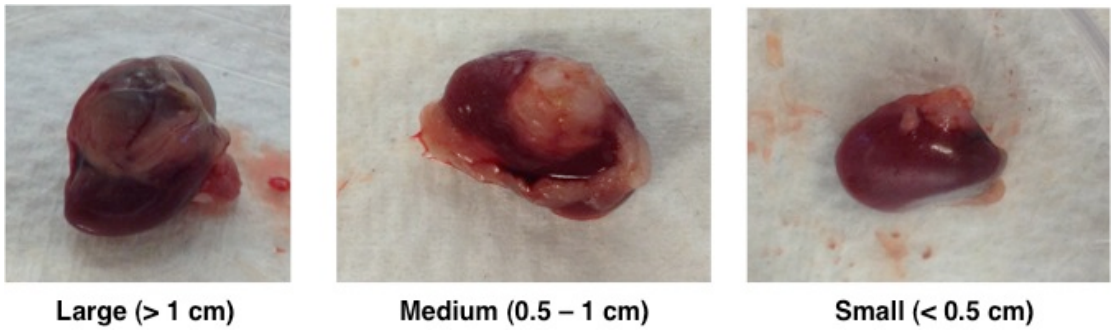

Figure 4.8 Examples of tumours generated from BCs or LPs transduced with *KRAS*^{G12D} + *PIK3CA*^{H1047R} + *TP53*^{R273C}

Tumours were palpable after 6-8 weeks in subrenal xenografts of transduced BCs or LPs. The size of each tumour is classified as small (< 0.5 cm), medium (0.5 to 1 cm) or large (>1 cm).

Table 4.4 Frequency of tumours generated *de novo* in NSG mice from transplanted transduced primary human mammary BCs and LPs

Co-transduction with multiple genes:

Sample name	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	Patient age
KRAS^{G12D} + PIK3CA^{H1047R} + TP53^{R273C}							
	BCs: 8/18 (44%)			**LPs: 12/18 (67%)**			
3-12	8×10^5	6	L	1×10^6	6	L	21
51-09	1.4×10^6	8.5	M	7×10^5	8.5	M	47
15-13*	1×10^6	7	M	1×10^6	7	M	
14-13*	1×10^6	7	S	1×10^6	7	S	33
172-04	4×10^5	8.5	M	7×10^5	8.5	M	54
18-13	3×10^4	4	S	6×10^4 10^6	4 8	S M	21
22-13*	1×10^5	8	S	7×10^5	8	S	35
199-04	1×10^6	8.5	L	5×10^5	8.5	-	43
17-12*	2.42×10^5	8	-	1×10^6	8	S	48
8-13*	1.25×10^5	8	-	2.8×10^5	8	M	21
17-13*	1.2×10^5	8	-	6.2×10^5	8	S	51
35-11*	1.2×10^5	8	-	5.7×10^5	8	M	33
38-12SQ	1×10^5	6	-	10^5	6	M	25
55-07	1×10^6	6	-	10^6	6	-	25
36-04	1×10^6	6	-	10^6	6	-	43
8-12SQ	1×10^6	6	-	8×10^5	6	-	25
35-10SQ	1×10^5	6	-	10^5	6	-	38
52-08	4×10^5	8.5	-	5.4×10^5	8.5	-	62

(Table continued on subsequent page…)

Sample name	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	Patient age
KRAS^{G12D} + TP53^{R273C}							
	BCs: 3/5 (60%)			**LPs: 6/6 (100%)**			
22-13*	1.6×10^5	8	S	7×10^5	8	S	35
17-13*	1.2×10^5	8	S	6.2×10^5	8	M	51
35-11*	1.2×10^5	8	M	5.7×10^5	8	M	33
17-12*	2.42×10^5	8	-	10^6	8	S	48
18-13	3×10^4	4	-	6×10^4	4	S	21
8-13*				2.8×10^5	8	L	21
KRAS^{G12D} + PIK3CA^{H1047R}							
	BCs: 3/6 (50%)			**LPs: 4/6 (67%)**			
17-12*	2.42×10^5	8	S	10^6	8	M	48
17-13*	1.2×10^5	8	S	6.2×10^5	8	M	51
4-13	3×10^4	4	S	6×10^4	4	-	32
22-13*	1.6×10^5	8	-	7×10^5	8	S	35
35-11*	1.2×10^5	8	-	5.7×10^5	8	M	33
8-13*	1.25×10^5	8	-	2.8×10^5	8	-	21

Transduction with single genes:

Sample name	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	Patient age
TP53^{R273C}							
	BCs: 0/1 (0%)			**LPs: 0/5 (0%)**			
17-12*	2.42×10^5	8	-	10^6	8	-	48
22-13*				7×10^5	8	-	35
17-13*				6.2×10^5	8	-	51
35-11*				5.7×10^5	8	-	33
8-13*				2.8×10^5	8	-	21

(Table continued on subsequent page…)

Sample name	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	No. of cells transplanted	Time *in vivo* (weeks)	Size (S, M or L)	Patient age
PIK3CA^{H1047R}							
	BCs: 0/1 (0%)			**LPs: 1/5 (20%)**			
8-13*				2.8×10^5	8	S	21
17-12*	2.42×10^5	8	-	10^6	8	-	48
22-13*				7×10^5	8	-	35
17-13*				6.2×10^5	8	-	51
35-11*				5.7×10^5	8	-	33
KRAS^{G12D}							
	BCs: 2/6 (33%)			**LPs: 6/6 (100%)**			
17-12*	2.42×10^5	8	S	10^6	8	S	48
22-13*	1.6×10^5	8	S	7×10^5	8	M	35
17-13*	1.2×10^5	8	-	6.2×10^5	8	M	51
35-11*	1.2×10^5	8	-	5.7×10^5	8	S	33
18-13	3×10^4	4	-	6×10^4	4	S	21
8-13*				2.8×10^5	8	M	21
***mCherry* (Control)**							
	BCs: 0/5 (0%)			**LPs: 0/5 (0%)**			
17-12*	2.42×10^5	8	-	10^6	8	-	48
22-13*	1.6×10^5	8	-	7×10^5	8	-	35
8-13*	1.25×10^5	8	-	2.8×10^5	8	-	21
17-13*	1.2×10^5	8	-	6.2×10^5	8	-	51
35-11*	1.2×10^5	8	-	5.7×10^5	8	-	33

In all experiments where the BCs and/or LPs were transduced with the mutant genes shown and simultaneously barcoded with the MPG barcoded lentiviral library (containing a GFP fluorescence reporter indicated by an asterisk (*)), a lentivirus encoding *TP53^{R273C}-YFP* was used instead of the GFP reporter. In this case, cells co-transduced with *PIK3CA^{H1047R}* and/or *TP53^{R273C}* are indistinguishable by fluorescence, but can be distinguished from cells containing a barcode. For three of the samples (indicated with a suffix "SQ" in the sample name), cells were injected subcutaneously with 50% matrigel rather than transplanted under the kidney capsule in collagen gels. Boxes shaded in gray

indicate tumours that did not appear after the indicated duration *in vivo*, and boxes in black indicate samples that were not tested under those conditions.

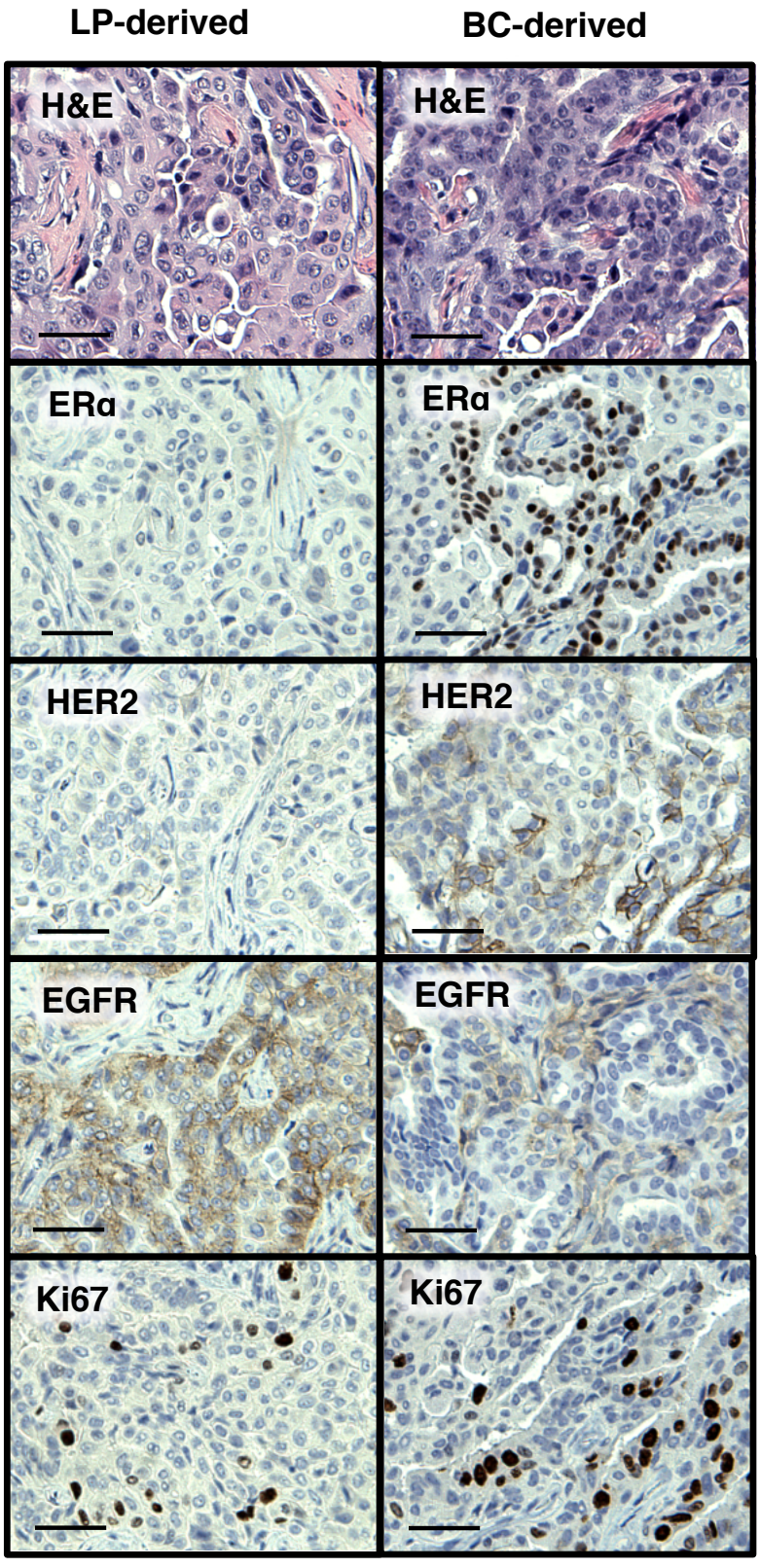
LP-derived
BC-derived
H&E
H&E
ERα
ERα
HER2
HER2
EGFR
EGFR
Ki67
Ki67

Figure 4.9

Figure 4.9 Morphological and immunohistochemical analysis of tumours derived from BCs and LPs

Shown are representative tissue sections of an ER^+ and ER^- tumour derived from $KRAS^{G12D}$ + $PIK3CA^{H1047R}$ + $TP53^{R273C}$ transduced primary human mammary BCs (top row) and LPs (bottom row), respectively. These tissue sections were stained with antibodies for ERα, HER2, EGFR and Ki67.

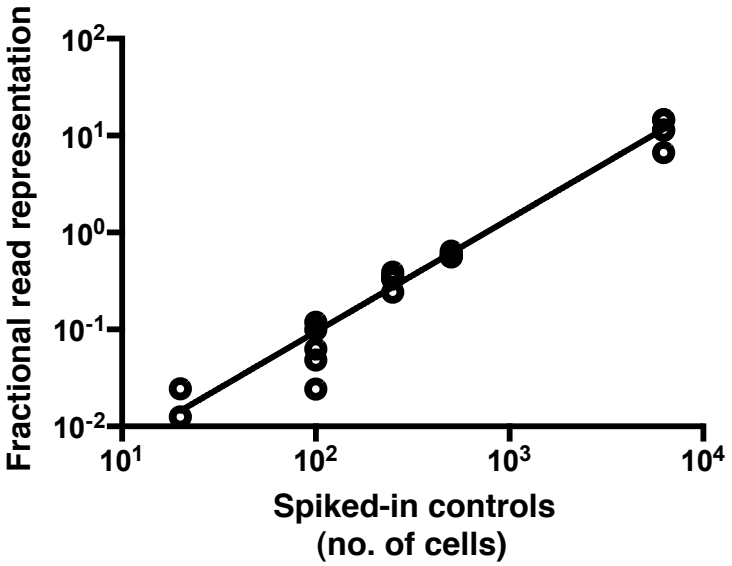

Figure 4.10 Normalization of barcode clones

Shown is a set of 5 spiked-in controls, consisting of 20-6250 cells, each corresponding to a previously determined barcode sequence. The fractional read representation for each control was calculated as the number of sequence reads divided by the sum of the reads for the 20, 100, 250 and 500 cell controls. The linear correlation shown is $y = 0.0019x - 0.1592$, and $R^2 = 0.99$.

Table 4.5 Frequency of tumour clonogenic cells

Transduction conditions	Sample name	BC-derived tumour	LP-derived tumour
$KRAS^{G12D}$ + $PIK3CA^{H1047R}$ + $TP53^{R273C}$	35-11		1/347
	17-13		1/437
	14-13		1/818
	17-12		1/855
	8-13		1/1,244
	15-13	1/470	1/1,453
	18-13		1/2,066
	22-13		1/3,704
Average (95% CI)			**1/816** (1/503 to 1/2,152)
$KRAS^{G12D}$ + $TP53^{R273C}$	8-13		1/143
	35-11		1/338
	22-13		1/626
	17-12		1/781
	17-13	1/181	NA
Average (95% CI)			**1/312** (1/135 to 1/inf)
$KRAS^{G12D}$ + $PIK3CA^{H1047R}$	22-13		1/235
	35-11		1/369
	17-13		1/372
	17-12	1/186	1/1,024
Average (95% CI)			**1/376** (1/209 to 1/1,900)
$KRAS^{G12D}$	8-13		1/138
	35-11		1/254
	22-13		1/354
	17-13		1/374
	17-12	1/169	1/551
Average (95% CI)			**1/270** (1/158 to 1/939)

Table 4.6 Number of clones of each size in barcoded tumours derived from human BCs and LPs

Tumours derived from cells transduced with *KRAS*^{G12D} + *PIK3CA*^{H1047R} + *TP53*^{R273C}:

Clone size (log_2-binned)	BC-derived tumour	LP-derived tumours								
	15-13	**15-13**	**35-11**	**17-13**	**14-13**	**17-12**	**8-13**	**18-13**	**22-13**	**Pooled** (LP-derived tumours)
2^0 to 2^1	0	0	0	0	0	0	0	0	0	0
2^1 to 2^2	0	0	0	0	0	0	0	0	0	0
2^2 to 2^3	0	0	0	0	0	0	0	0	0	0
2^3 to 2^4	0	0	0	0	0	0	0	0	0	0
2^4 to 2^5	0	0	0	0	0	0	0	0	0	0
2^5 to 2^6	0	0	0	0	0	0	0	0	0	0
2^6 to 2^7	0	0	0	0	0	0	0	0	0	0
2^7 to 2^8	0	0	0	0	0	0	0	0	115	115
2^8 to 2^9	1,309	684	360	0	478	0	0	484	27	2,033
2^9 to 2^{10}	432	4	444	0	319	0	0	0	13	780
2^{10} to 2^{11}	231	0	382	0	204	0	193	0	9	788
2^{11} to 2^{12}	105	0	245	264	95	195	19	0	8	826
2^{12} to 2^{13}	27	0	144	204	57	215	6	0	14	640
2^{13} to 2^{14}	21	0	49	216	44	193	1	0	3	506
2^{14} to 2^{15}	4	0	17	207	13	166	0	0	0	403
2^{15} to 2^{16}	0	0	2	173	4	146	1	0	0	326
2^{16} to 2^{17}	0	0	1	168	2	110	0	0	0	281
2^{17} to 2^{18}	0	0	0	97	2	80	0	0	0	179
2^{18} to 2^{19}	0	0	0	79	4	42	0	0	0	125
2^{19} to 2^{20}	0	0	0	10	0	14	0	0	0	24
2^{20} to 2^{21}	0	0	0	0	0	8	0	0	0	8
2^{21} to 2^{22}	0	0	0	0	0	0	1	0	0	1
2^{22} to 2^{23}	0	0	0	0	0	0	3	0	0	3
2^{23} to 2^{24}	0	0	0	0	0	0	1	0	0	1
Total no. of clones	2,129	688	1,644	1,418	1,222	1,169	225	484	189	7,039

Tumours derived from cells transduced with *KRAS*^{G12D}+ *TP53*^{R273C}:

Clone size (log_2-binned)	BC-derived tumour	LP-derived tumours				
	17-13	8-13	35-11	22-13	17-12	Pooled (LP-derived tumours)
2^0 to 2^1	0	0	0	0	0	0
2^1 to 2^2	0	0	0	0	0	0
2^2 to 2^3	0	0	0	0	0	0
2^3 to 2^4	0	0	0	0	0	0
2^4 to 2^5	0	0	0	0	0	0
2^5 to 2^6	0	0	0	0	0	0
2^6 to 2^7	0	0	0	0	0	0
2^7 to 2^8	0	0	177	126	0	303
2^8 to 2^9	92	0	406	97	0	503
2^9 to 2^{10}	44	456	402	262	0	1,120
2^{10} to 2^{11}	168	307	320	217	176	1,020
2^{11} to 2^{12}	138	324	214	159	191	888
2^{12} to 2^{13}	108	303	111	100	215	729
2^{13} to 2^{14}	76	279	52	72	204	607
2^{14} to 2^{15}	21	190	6	66	175	437
2^{15} to 2^{16}	11	89	0	13	165	267
2^{16} to 2^{17}	3	10	0	7	96	113
2^{17} to 2^{18}	1	2	0	0	39	41
2^{18} to 2^{19}	2	1	0	0	12	13
2^{19} to 2^{20}	0	0	0	0	6	6
2^{20} to 2^{21}	0	0	0	0	2	2
2^{21} to 2^{22}	0	0	0	0	0	0
2^{22} to 2^{23}	0	0	0	0	0	0
2^{23} to 2^{24}	0	0	0	0	0	0
Total no. of clones	664	1,961	1,688	1,119	1,281	6,049

Tumours derived from cells transduced with *KRAS^{G12D}+PIK3CA^{H1047R}*:

Clone size (log_2-binned)	BC-derived tumours	LP-derived tumours				
	17-12	17-12	22-13	35-11	17-13	Pooled (LP-derived tumours)
2^0 to 2^1	0	0	0	0	0	0
2^1 to 2^2	0	0	0	0	0	0
2^2 to 2^3	0	0	0	0	0	0
2^3 to 2^4	0	0	0	0	0	0
2^4 to 2^5	0	0	0	0	0	0
2^5 to 2^6	0	0	0	0	0	0
2^6 to 2^7	0	0	0	0	0	0
2^7 to 2^8	182	0	0	118	0	118
2^8 to 2^9	421	0	0	140	0	140
2^9 to 2^{10}	332	0	842	349	421	1,612
2^{10} to 2^{11}	205	172	520	334	270	1,296
2^{11} to 2^{12}	103	161	425	276	292	1,154
2^{12} to 2^{13}	45	149	468	191	231	1,039
2^{13} to 2^{14}	11	145	338	93	210	786
2^{14} to 2^{15}	2	101	228	24	139	492
2^{15} to 2^{16}	1	115	120	17	79	331
2^{16} to 2^{17}	0	65	26	1	12	104
2^{17} to 2^{18}	0	32	6	0	11	49
2^{18} to 2^{19}	0	29	0	1	1	31
2^{19} to 2^{20}	1	8	0	0	0	8
2^{20} to 2^{21}	0	0	0	0	0	0
2^{21} to 2^{22}	0	0	0	0	0	0
2^{22} to 2^{23}	0	0	0	0	0	0
2^{23} to 2^{24}	0	0	0	0	0	0
Total no. of clones	1,303	977	2,973	1,544	1,666	7,160

Tumours derived from cells transduced with *KRAS*^{G12D}:

Clone size (log_2-binned)	BC-derived tumour	LP-derived tumours					
	17-12	17-12	8-13	35-11	22-13	17-13	Pooled (LP-derived tumours)
2^0 to 2^1	0	0	0	0	0	0	0
2^1 to 2^2	0	0	0	0	0	0	0
2^2 to 2^3	0	0	0	0	0	0	0
2^3 to 2^4	0	0	0	0	0	0	0
2^4 to 2^5	0	0	0	0	0	0	0
2^5 to 2^6	0	0	0	0	0	0	0
2^6 to 2^7	0	0	0	0	0	0	0
2^7 to 2^8	0	0	0	0	0	0	0
2^8 to 2^9	333	0	0	0	0	0	0
2^9 to 2^{10}	369	0	0	0	0	0	0
2^{10} to 2^{11}	291	0	0	264	0	0	264
2^{11} to 2^{12}	180	275	285	310	267	308	1,445
2^{12} to 2^{13}	126	267	310	362	304	276	1,519
2^{13} to 2^{14}	94	260	328	369	309	266	1,532
2^{14} to 2^{15}	28	321	346	346	313	233	1,559
2^{15} to 2^{16}	9	252	319	315	274	186	1,346
2^{16} to 2^{17}	3	262	288	197	279	187	1,213
2^{17} to 2^{18}	0	106	126	72	137	155	596
2^{18} to 2^{19}	1	50	28	10	52	42	182
2^{19} to 2^{20}	0	17	2	3	32	5	59
2^{20} to 2^{21}	0	5	1	0	7	1	14
2^{21} to 2^{22}	0	0	0	0	1	0	1
2^{22} to 2^{23}	0	0	0	0	0	0	0
2^{23} to 2^{24}	0	0	0	0	0	0	0
Total no. of clones	1,434	1,815	2,033	2,248	1,975	1,659	9,730

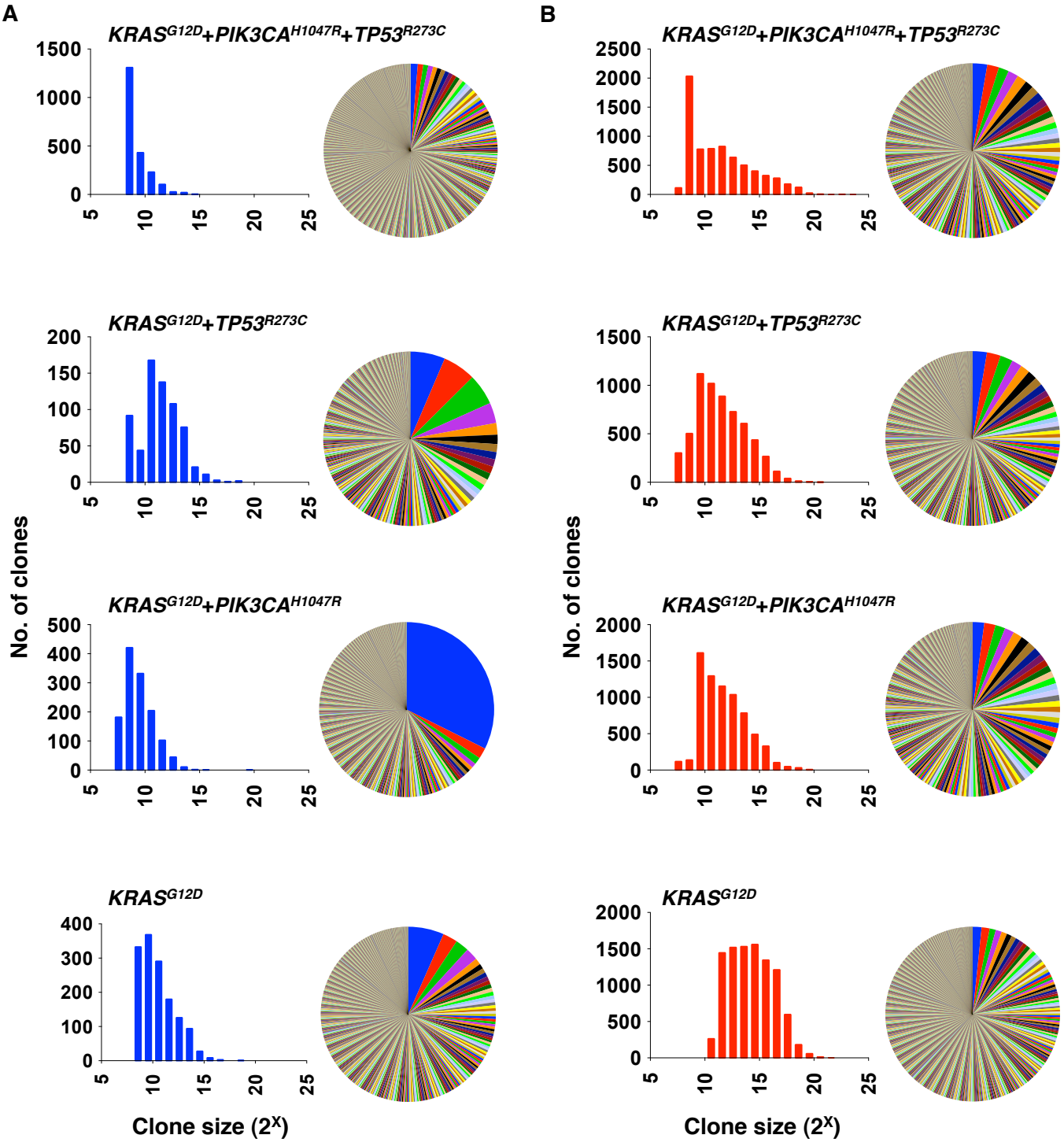
A
B
KRASG12D+PIK3CAH1047R+TP53R273C
KRASG12D+TP53R273C
KRASG12D+PIK3CAH1047R
KRASG12D
No. of clones
Clone size (2X)

Figure 4.11

Figure 4.11 Clonal composition of tumours generated *de novo* from BCs and LPs

Histogram plots showing the clone size distributions (binned in $\log_2$ increments) as estimated from barcode analyses of tumours derived from transduced BCs (blue bars, A) and LPs (red bars, B). The number of tumours pooled to generate each histogram is indicated. The pie charts on the right of each histogram depict the complete distribution of individual clone sizes in a single representative tumour.

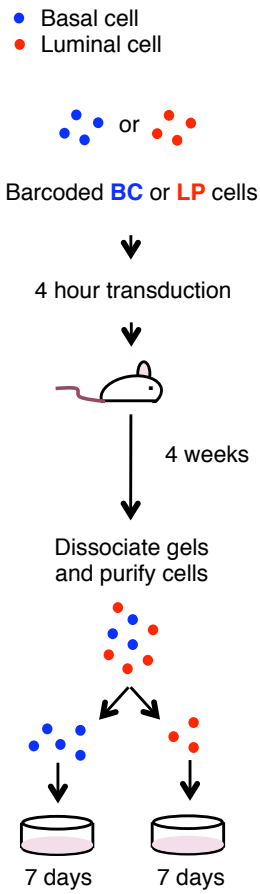

Figure 4.12 Experimental design used to investigate early oncogene-induced changes in human BC and LP growth and differentiation

BCs and LPs, shown as blue and red circles, respectively, were transduced for 4 hours with one or more genes (*KRAS*^{G12D}, *PIK3CA*^{H1047R}, *TP53*^{R273C}), or an empty control vector (encoding *GFP*), and then immediately transplanted into NSG mice. After 4 weeks, the regenerated cells were sorted into basal and luminal fractions (also blue and red, respectively), and assayed for 2D CFC activity.

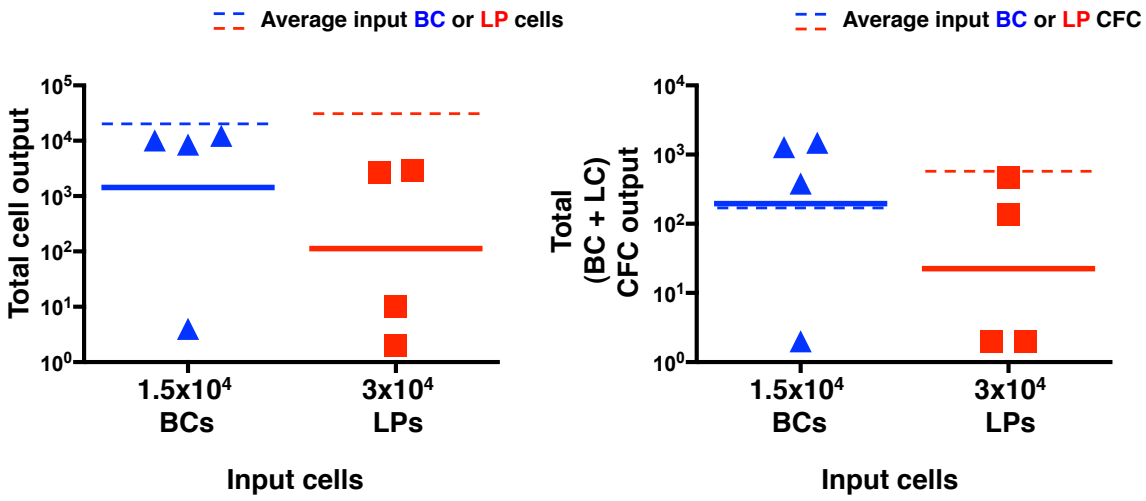

Figure 4.13 Total cell and CFC outputs from control transduced BCs and LPs after 4 weeks *in vivo*

$1.5x10^4$ and $3x10^4$ primary normal human mammary BCs and LPs from 4 different reduction mammoplasty samples were transduced with a lentivirus encoding a fluorescence reporter gene only, and the cells were then transplanted under the kidney capsule of NSG mice. After 4 weeks, the gels were retrieved, the cells dissociated, and the total cell (left) and CFC outputs (right) measured. Blue triangles represent the results for BC transplants, and the red squares for the LP transplants, with the values adjusted to reflect $1.5x10^4$ and $3x10^4$ positively transduced input cells, respectively.

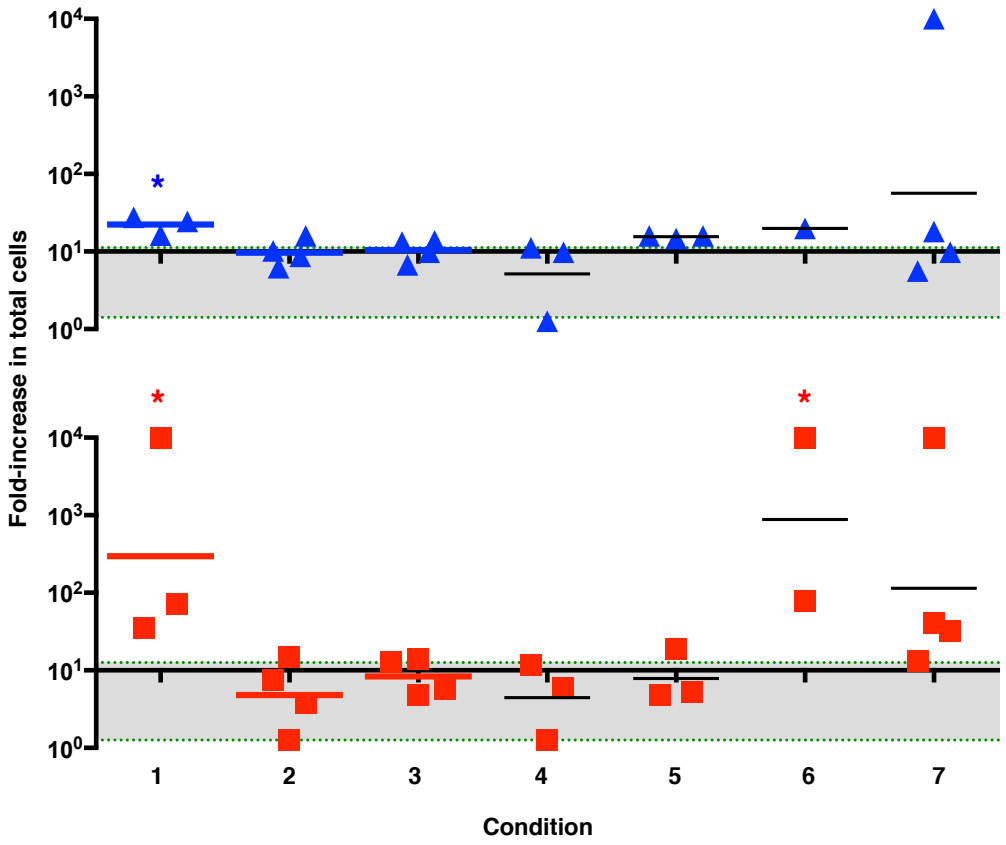

Figure 4.14 Increase after 4 weeks *in vivo* of total cell outputs from BCs and LPs transduced with *KRAS*^{G12D}, *PIK3CA*^{H1047R} and *TP53*^{R273C} alone and in combination

Total cell outputs from transplanted BCs (blue triangles) and LPs (red squares) transduced according to conditions 1 to 7 (described below) are expressed as a fold-increase compared to the average control values (shown in Figure 4.13). The shaded gray area represents the normal range of the control values. Statistically significant differences ($p < 0.05$) are indicated with an asterisk. Condition 1 = *KRAS*^{G12D} only, 2 = *PIK3CA*^{H1047R} only, 3 = *TP53*^{R273C} only, 4 = *PIK3CA*^{H1047R}+*TP53*^{R273C}, 5 = *KRAS*^{G12D}+*PIK3CA*^{H1047R}, 6 = *KRAS*^{G12D}+*TP53*^{R273C}, and 7 = *KRAS*^{G12D}+*PIK3CA*^{H1047R}+*TP53*^{R273C}.

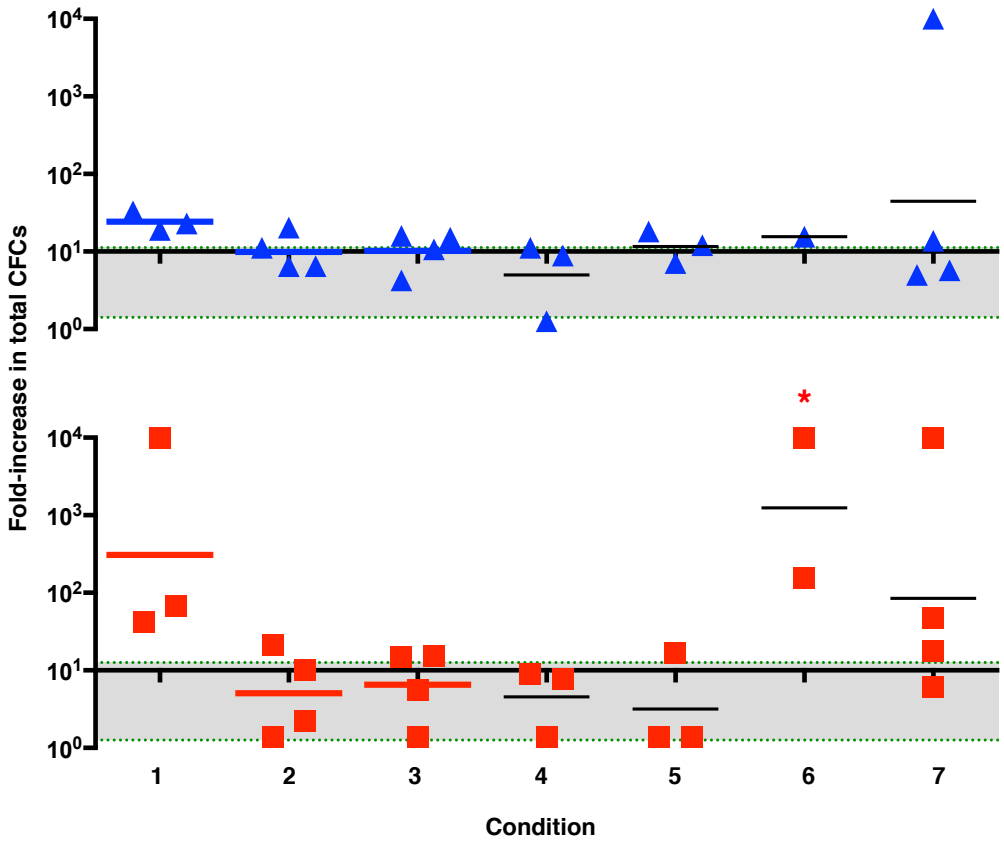

Figure 4.15 Increase after 4 weeks *in vivo* of total CFCs from BCs and LPs transduced with *KRAS*^{G12D}, *PIK3CA*^{H1047R} and *TP53*^{R273C} alone and in combination

Total CFC outputs from transplanted BCs (blue triangles) and LPs (red squares) transduced according to conditions 1 to 7 (described below) are expressed as a fold-increase compared to average control values (shown in Figure 4.13). The shaded gray area represents the normal range of the control values. Statistically significant differences ($p < 0.05$) are indicated with an asterisk. Condition 1 = *KRAS*^{G12D} only, 2 = *PIK3CA*^{H1047R} only, 3 = *TP53*^{R273C} only, 4 = *PIK3CA*^{H1047R}+*TP53*^{R273C}, 5 = *KRAS*^{G12D}+*PIK3CA*^{H1047R}, 6 = *KRAS*^{G12D}+*TP53*^{R273C}, and 7 = *KRAS*^{G12D}+*PIK3CA*^{H1047R}+*TP53*^{R273C}.

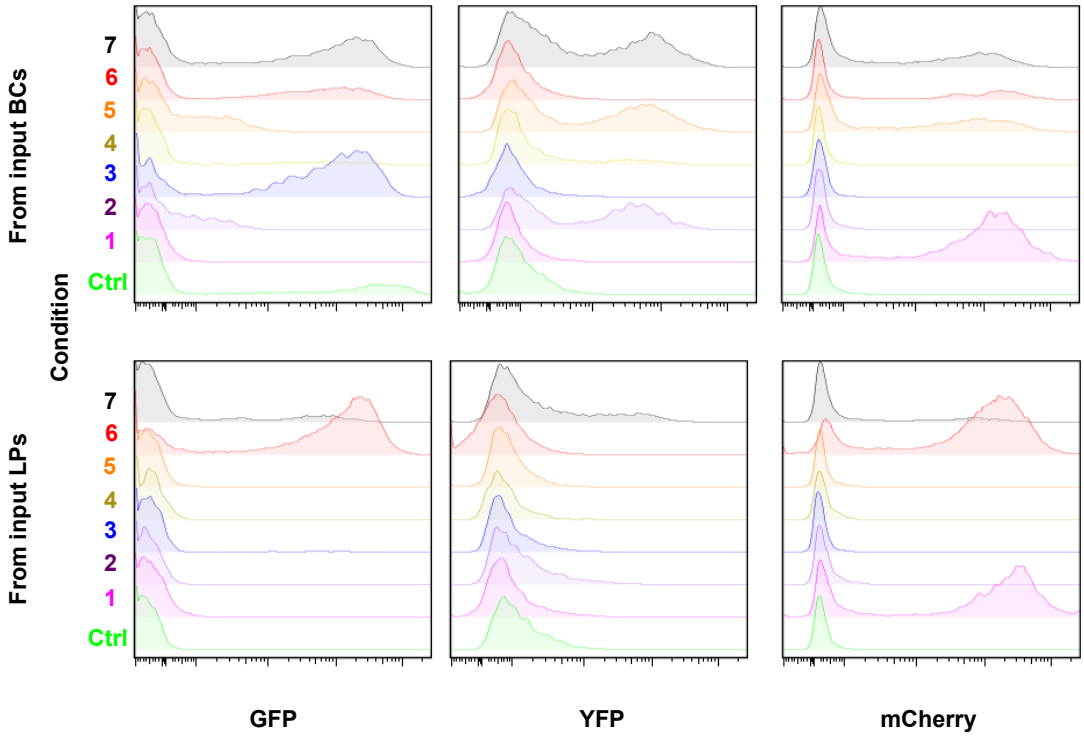

Figure 4.16 Representative flow analysis plots of *in vivo* regenerated cells

GFP, YFP and mCherry expression is shown for the *in vivo* regenerated cells from transplants of BCs (top three histogram plots) and LPs (bottom three histogram plots) transduced with an empty GFP control vector (Ctrl, green) or conditions 1 to 7 (1 = $KRAS^{G12D}$ only, 2 = $PIK3CA^{H1047R}$ only, 3 = $TP53^{R273C}$ only, 4 = $PIK3CA^{H1047R}$+$TP53^{R273C}$, 5 = $KRAS^{G12D}$+$PIK3CA^{H1047R}$, 6 = $KRAS^{G12D}$+$TP53^{R273C}$, and 7 = $KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$).

Table 4.7 Flow cytometric analysis of cells regenerated from transplanted BCs and LPs after 4 weeks *in vivo*

Shown in the tables below are the percentages of GFP^+ (corresponds to either the expression of *GFP* in the control vector or expression of $TP53^{R273C}$), YFP^+ (corresponds to the expression of $PIK3CA^{H1047R}$) or $mCherry^+$ (corresponds to the expression of $KRAS^{G12D}$) cells in the total cells present after 4 weeks in xenografts derived from transplanted BCs and LPs. The four values indicated are from four different mammoplasty samples ("-" indicates a measurement not taken for that particular biological replicate). Condition 1 = $KRAS^{G12D}$ only, 2 = $PIK3CA^{H1047R}$ only, 3 = $TP53^{R273C}$ only, 4 = $PIK3CA^{H1047R}$+$TP53^{R273C}$, 5 = $KRAS^{G12D}$+$PIK3CA^{H1047R}$, 6 = $KRAS^{G12D}$+$TP53^{R273C}$, and 7 = $KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$.

From transplanted BCs:

Condition	Neg	GFP^+	YFP^+	$mCherry^+$	GFP^+/YFP^+	$YFP^+/mCherry^+$	$GFP^+/mCherry^+$	$GFP^+/YFP^+/mCherry^+$
	31	67	0	3	0	0	0	0
	25	75	1	1	1	0	1	0
	85	15	0	1	0	0	0	0
Ctrl	100	0	0	0	0	0	0	0
	2	0	0	98	0	0	0	0
	3	0	0	97	0	0	0	0
	24	0	0	76	0	0	0	0
1	-	-	-	-	-	-	-	-
	20	0	78	2	0	0	0	0
	67	0	33	0	0	0	0	0
	96	0	4	0	0	0	0	0
2	96	0	4	0	0	0	0	0
	18	82	0	0	0	0	0	0
	16	84	0	1	0	0	0	0
	69	29	0	3	0	0	1	0
3	97	3	0	0	0	0	0	0
	41	46	50	0	37	0	0	0
	29	52	67	0	48	0	0	0
	-	-	-	-	-	-	-	-
4	100	0	0	0	0	0	0	0
	6	2	80	89	2	76	2	2
	-	-	-	-	-	-	-	-
	65	0	27	32	0	24	0	0
5	84	2	14	16	2	13	2	2
	8	70	0	92	0	0	70	0
	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-
	7	80	82	88	72	77	77	69
	-	-	-	-	-	-	-	-
	86	5	7	12	4	6	4	3
7	64	10	29	31	6	26	8	5

From transplanted LPs:

Condition	Neg	GFP$^+$	YFP$^+$	mCherry$^+$	GFP$^+$/ YFP$^+$	YFP$^+$/ mCherry$^+$	GFP$^+$/ mCherry$^+$	GFP$^+$/ YFP$^+$/ mCherry$^+$
	97	0	0	3	0	0	0	0
	79	21	0	0	0	0	0	0
	100	0	0	0	0	0	0	0
Ctrl	82	18	0	0	0	0	0	0
	4	0	0	96	0	0	0	0
	-	-	-	-	-	-	-	-
	45	0	0	55	0	0	0	0
1	-	-	-	-	-	-	-	-
	85	0	15	0	0	0	0	0
	100	0	0	0	0	0	0	0
	100	0	0	0	0	0	0	0
2	99	0	1	0	0	0	0	0
	76	24	0	0	0	0	0	0
	54	45	0	0	0	0	0	0
	100	0	0	0	0	0	0	0
3	98	1	0	0	0	0	0	0
	92	0	0	8	0	0	0	0
	67	27	31	1	25	0	0	0
	-	-	-	-	-	-	-	-
4	98	1	2	0	1	0	0	0
	98	0	1	2	0	1	0	0
	-	-	-	-	-	-	-	-
	72	0	18	28	0	18	0	0
5	99	0	0	1	0	0	0	0
	2	85	0	97	0	0	84	0
	1	90	0	99	0	0	90	0
	-	-	0	-	-	-	-	-
6	-	-	-	-	-	-	-	-
	22	43	58	75	33	57	42	33
	-	-	-	-	-	-	-	-
	91	4	6	9	4	6	4	4
7	46	37	39	53	29	39	37	29

Table 4.8 Frequency of premalignant clones evident after 2 weeks *in vivo*

Condition	Cell type transplanted	
	BCs	LPs
$KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$	1 in 174	1 in 284
$KRAS^{G12D}$+$TP53^{R273C}$	1 in 139	1 in 300
$KRAS^{G12D}$+$PIK3CA^{H1047R}$	1 in 90	1 in 220
$PIK3CA^{H1047R}$+$TP53^{R273C}$	1 in 79	1 in 403
$TP53^{R273C}$	1 in 84	1 in 199
$PIK3CA^{H1047R}$	1 in 100	1 in 210
$KRAS^{G12D}$	1 in 90	1 in 202
mCherry	1 in 147	1 in 631

The frequency of clones tabulated here corresponds to the same xenografts shown in Figure 4.17 A-D.

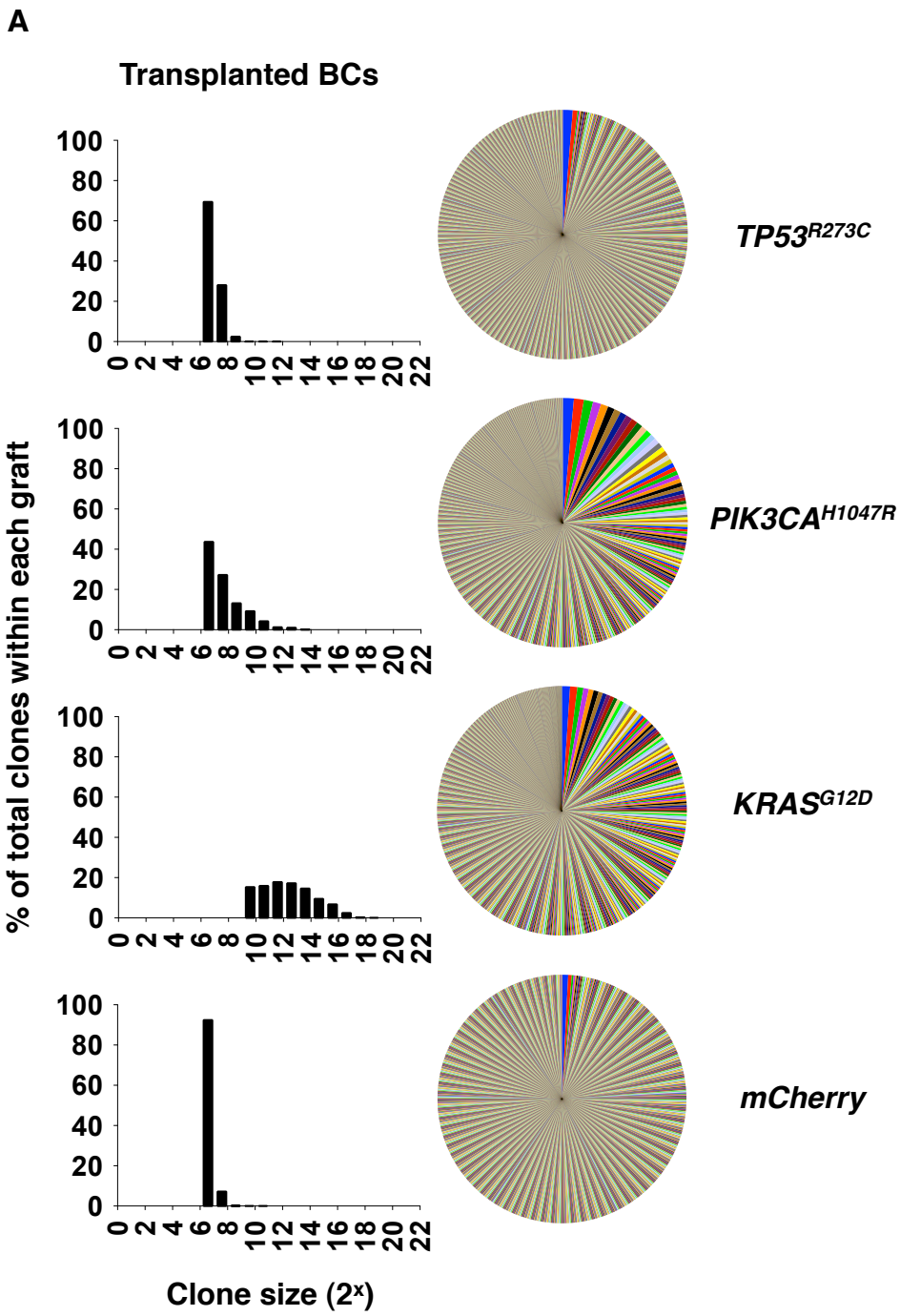
A
Transplanted BCs
% of total clones within each graft
Clone size (2x)
TP53R273C
PIK3CAH1047R
KRASG12D
mCherry

Figure 4.17A

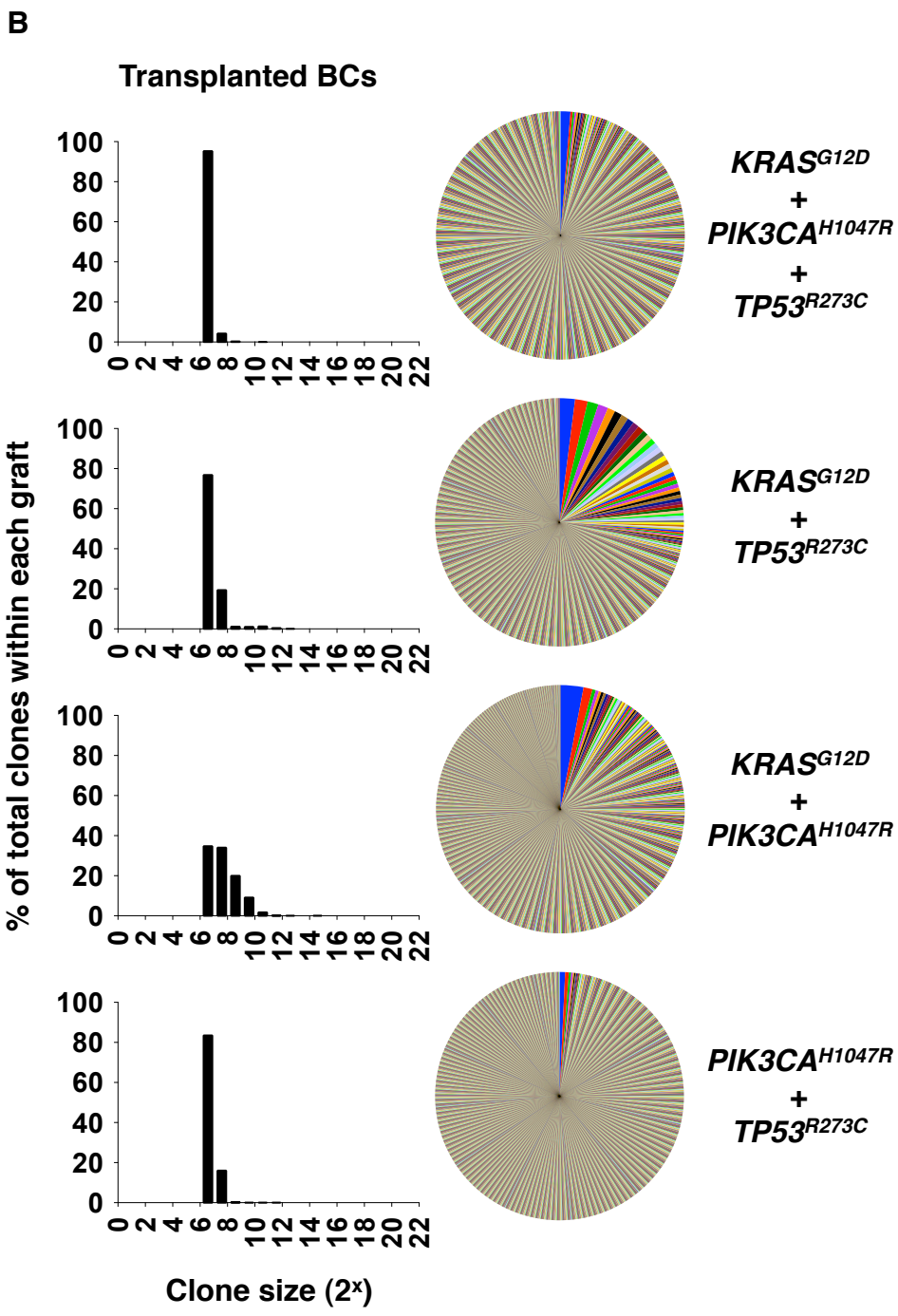
B
Transplanted BCs
% of total clones within each graft
Clone size (2x)
KRASG12D + PIK3CAH1047R + TP53R273C
KRASG12D + TP53R273C
KRASG12D + PIK3CAH1047R
PIK3CAH1047R + TP53R273C

Figure 4.17B

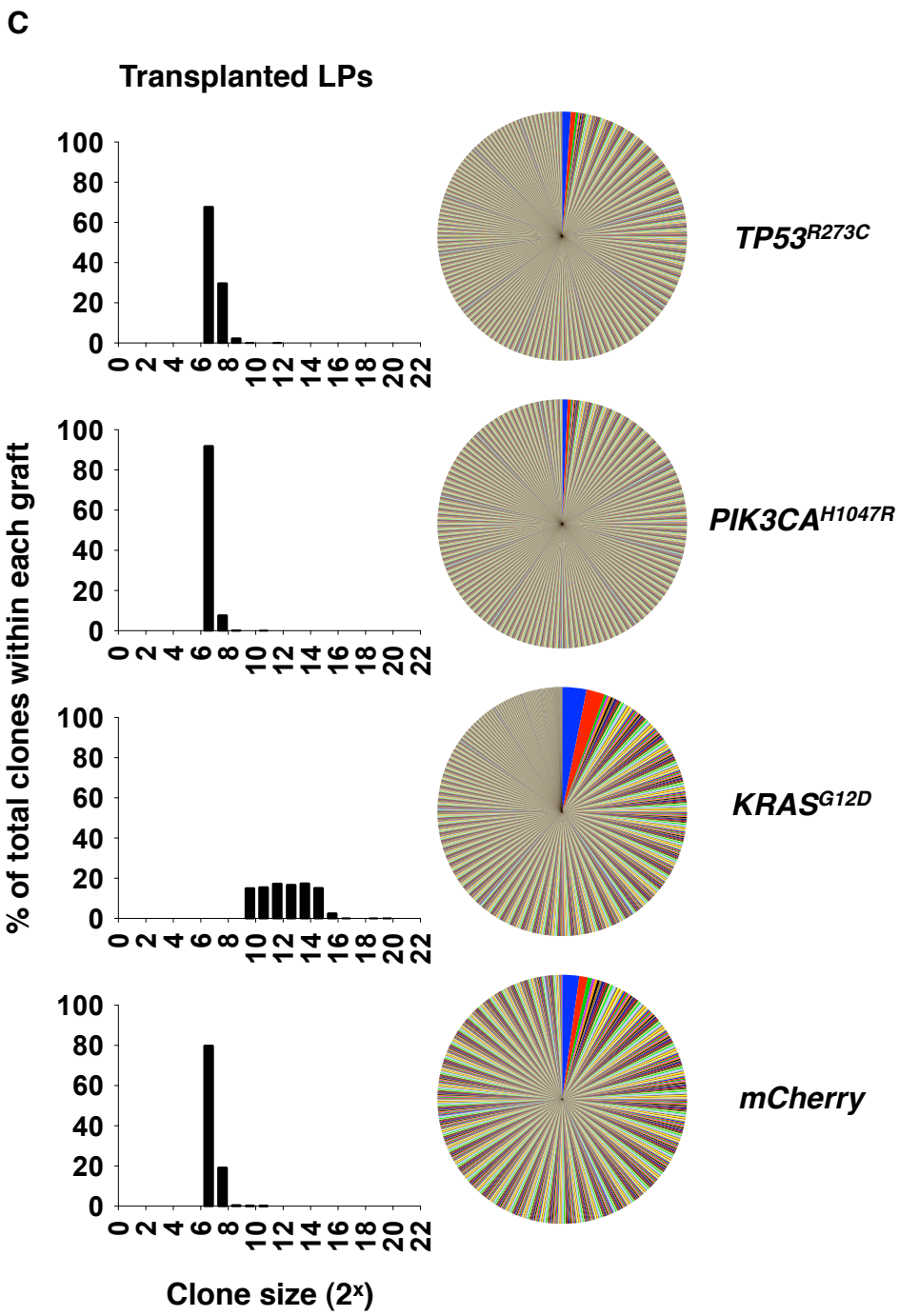
C
Transplanted LPs
% of total clones within each graft
Clone size (2x)
TP53R273C
PIK3CAH1047R
KRASG12D
mCherry

Figure 4.17C

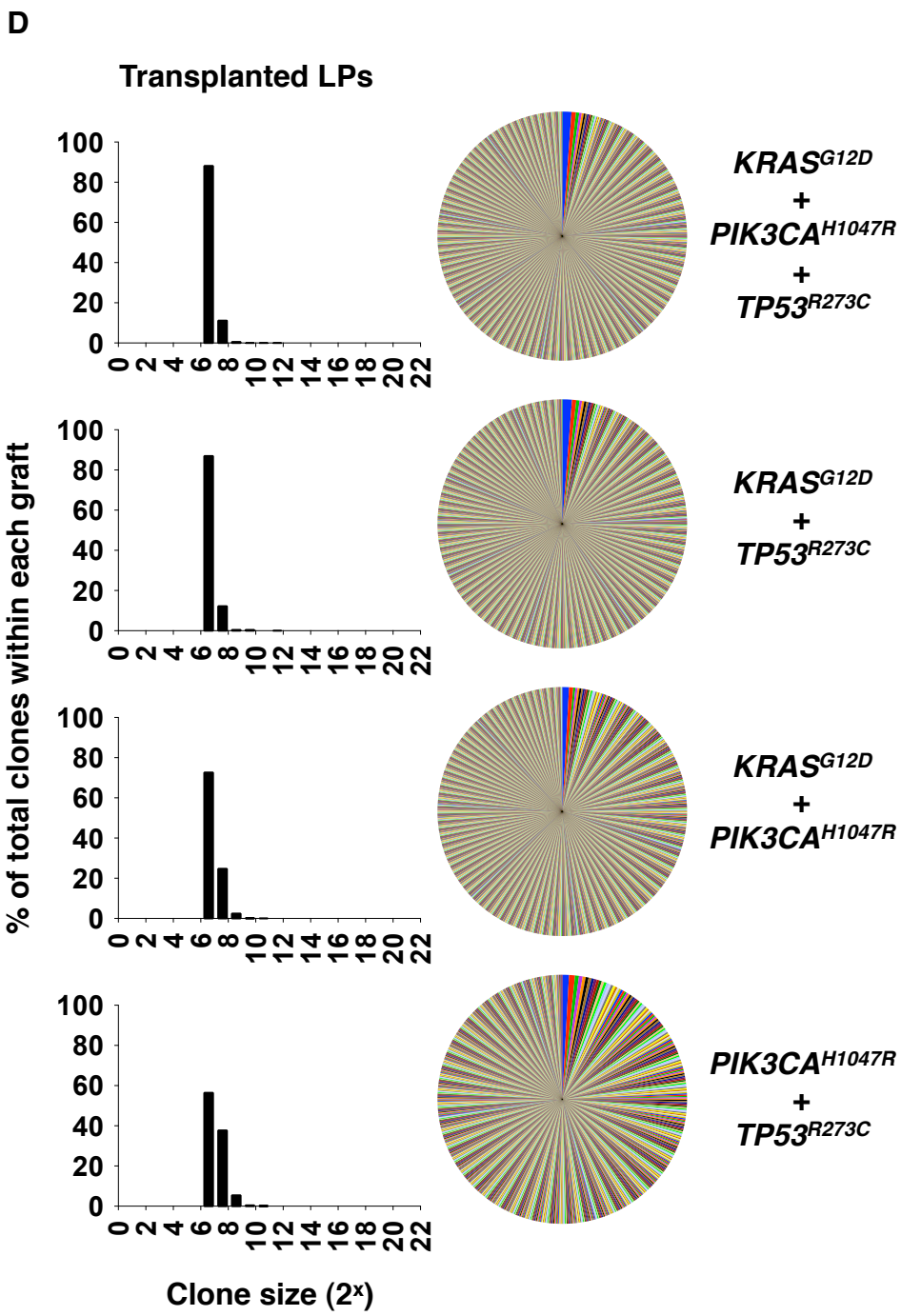
D
Transplanted LPs
% of total clones within each graft
Clone size (2x)
KRASG12D + PIK3CAH1047R + TP53R273C
KRASG12D + TP53R273C
KRASG12D + PIK3CAH1047R
PIK3CAH1047R + TP53R273C

Figure 4.17D

Figure 4.17 Clonal composition of xenografts of transduced BCs and LPs after 2 weeks *in vivo*

Histogram plots of the clone size distributions (binned in $\log_2$ increments) as estimated from barcode analyses of transplanted BCs or transplanted LPs transduced with one or more oncogenes (A/B and C/D, respectively). Each histogram represents a single xenograft. The pie charts to the right depict the proportion of cells contributed by each individual clone to the total cell output.

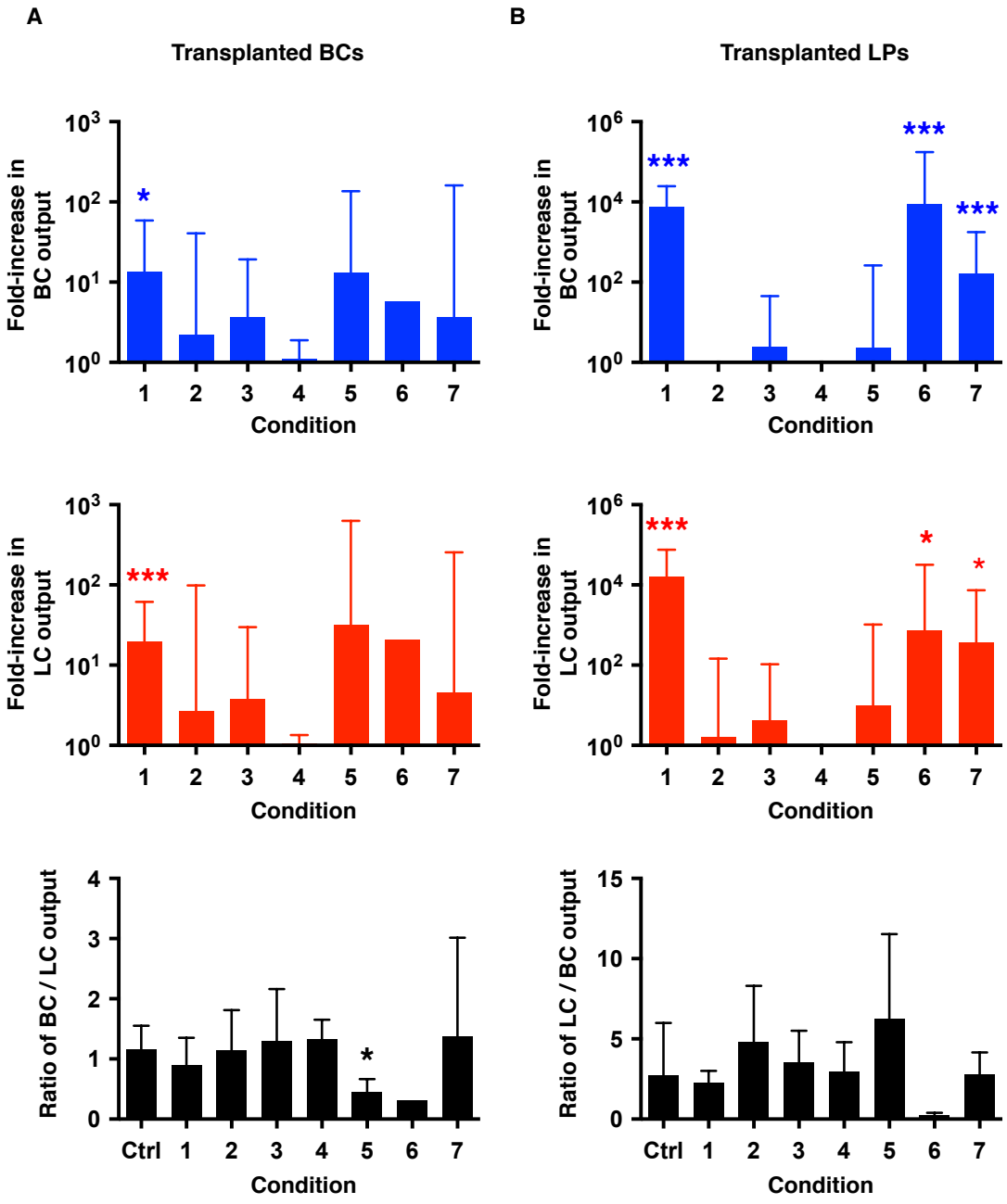
A
Transplanted BCs
Fold-increase in BC output
Condition
Fold-increase in LC output
Condition
Ratio of BC / LC output
Ctrl 1 2 3 4 5 6 7
Condition
B
Transplanted LPs
Fold-increase in BC output
Condition
Fold-increase in LC output
Condition
Ratio of LC / BC output
Ctrl 1 2 3 4 5 6 7
Condition

Figure 4.18

Figure 4.18 Phenotype analysis of cells present in xenografts generated from BCs and LPs transduced with various oncogenes after 4 weeks

The fold-increase in BC (blue) and LC (red) cell outputs from test vector-transduced as compared to control transduced cells (shown in Figure 4.13) are shown for $1.5x10^4$ and $3x10^4$ transplanted BCs (A) and LPs (B), respectively. Condition 1 = $KRAS^{G12D}$ only, 2 = $PIK3CA^{H1047R}$ only, 3 = $TP53^{R273C}$ only, 4 = $PIK3CA^{H1047R}$+$TP53^{R273C}$, 5 = $KRAS^{G12D}$+$PIK3CA^{H1047R}$, 6 = $KRAS^{G12D}$+$TP53^{R273C}$, and 7 = $KRAS^{G12D}$+$PIK3CA^{H1047R}$+$TP53^{R273C}$. Also shown is the ratio of total BC to LC or total LC to BC numbers derived from transplanted BCs and LPs (bottom bar plots, black, respectively). For all bar plots in A and B, statistically significant differences (compared to control) are indicated with an asterisk (* and ***, for p-values <0.05 and <0.01, respectively).

A

No. of cells per clone (log_2-increments)	No. of clones	
	Replicate #1	Replicate #2
$< 2^6$	0	0
2^6 to 2^7	0	82
2^7 to 2^8	166	123
2^8 to 2^9	58	80
2^9 to 2^{10}	32	36
2^{10} to 2^{11}	26	24
2^{11} to 2^{12}	28	28
2^{12} to 2^{13}	36	28
2^{13} to 2^{14}	16	26
2^{14} to 2^{15}	15	17
2^{15} to 2^{16}	10	3
2^{16} to 2^{17}	17	1
2^{17} to 2^{18}	13	6
2^{18} to 2^{19}	1	0
$> 2^{19}$	0	0

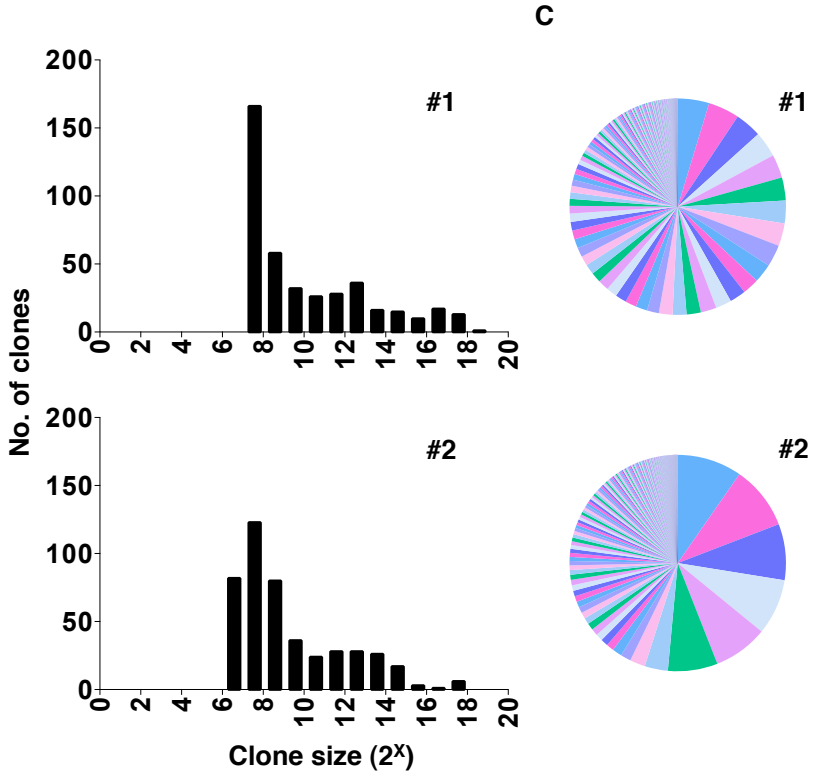

Figure 4.19 Clonal composition of breast tumour xenografts from the serial *in vivo* propagation of a single pleural effusion sample

A pleural effusion sample from a patient with advanced breast cancer (originally an ER^+ tumour) was generated in a primary NSG mouse, and the cells from this first passage were then barcoded and injected into two secondary NSG mice. The clonal composition of the tumours obtained in the two secondary recipients (Replicate #1 and #2) as determined by barcoding are shown as a clone size distributions (number of cells per clone binned in log_2-increments, A and B). The proportion of cells in each individual clone that contributed to the entire tumour is depicted as a pie chart (C), where each wedge indicates a unique clone.

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS

5.1 Possibilities for further improving cellular barcoding technology

DNA barcoding of cells using diverse vector libraries offers an extraordinarily powerful approach to analyzing the clonal outputs of single cells without the necessity to visualize or isolate their progeny (Cheung et al., 2013; Gerrits et al., 2010; Grosselin et al., 2013; Lu et al., 2011; Naik et al., 2013; Nguyen et al., 2014; Schepers et al., 2008). The strategy for analyzing MPS data from cellular barcoding experiments presented in this thesis reinforces this principle and illustrates its applicability to analyze the *in vivo* growth behavior of large numbers of transplanted normal and genetically transformed mammary epithelial cells of defined phenotypes. As detailed in Chapter 2, this was enabled by our development and validation of an improved approach for inferring clone size from barcode sequence data using "spiked-in controls" that consisted of known numbers of cells carrying a known barcode. The importance of this approach is particularly relevant to the resolving power of the method, due to the level of background noise inherent in the use of the MPS platform for deep sequencing.

The use of spiked in control cells allowed a threshold of clone sizes to be detected at defined confidence levels, but also revealed that an appreciable number of smaller clones were missed (typically <100 cells per clone in our experiments). This emerged as a potentially significant issue when it became apparent that many clones only appeared in secondary transplants of either normal mouse or human mammary epithelial cells, suggesting an initial delay in their growth in the primary hosts. In addition, we found that the smaller numbers of spiked-in controls used were often missed when combined with very large clones that saturated the capacity of the Illumina MiSeq as proved to be the

situation for the tumours generated *de novo* from normal human mammary epithelial cells. In this case, we found many clones that contained millions of cells. The use of a broader range of spiked-in control cell numbers could address these short-comings. However, this solution would also decrease the proportion of sequence reads available for detection of experimental clones, thus reducing its efficiency.

To circumvent this latter limitation, an alternative to the use of spiked-in controls altogether would be desirable. One approach could be to use an additional barcoding step to serve as an internal calibration of the sequencing run (Figure 5.1). When individual cells are initially transduced, each is assumed to have incorporated a single unique DNA barcode that is faithfully replicated in each daughter cell. Thus, the number of different template barcodes for PCR amplification obtained from a population in which multiple clones are represented in unknown numbers, should depend directly on the number of cells in each clone. Thus, we would propose that another DNA barcode (a short stretch of degenerate nucleotides) could be added to the primers used to amplify the primary barcodes that have been extracted from the cells. A single cycle of linear-amplification PCR would then be used to mark each starting template with a unique "secondary" barcode. A reaction cleanup would then remove all of the secondary barcodes, and allow for subsequent exponential amplification to be performed using common flanking primers. Thus, when the number of unique secondary barcodes is tabulated corresponding to a single primary barcode, this number of secondary barcodes would correspond to the number of cells per clone, providing a direct quantitation of clone size, replacing the use of spiked-in controls to infer this number. The used of staggered primers would also eliminate the need for a PhiX spike-in (exogenous DNA, which reduces the sequence

coverage allocated to the experimental clones), used to improve cluster recognition on the Illumina HiSeq and MiSeq MPS platforms. Undoubtedly, this proposed approach will be found to have its own set of limitations, such as the efficiency of the first single-cycle linear-amplification PCR that will directly impact the accuracy of clone size estimation. Nevertheless, it appears to offer significant advantages, particularly for the kind of *in vivo* tracking studies described in this thesis.

5.2 Barcoding reveals unexpected patterns of *in vivo* regenerative activity exhibited by both mouse and human mammary epithelial cells

A key issue investigated in this thesis was to determine the kinetics of clone development from normal mammary epithelial (basal) cells with long-term (sustained) *in vivo* regenerative activity as revealed in secondary hosts by comparison to those able to regenerate detectable populations directly in primary hosts. Serial transplantability has long been used as a surrogate indicator to infer the presence of cells with "self-renewal" capacity. Although rigorous evidence that secondary hosts have been repopulated with cells that contributed substantially to initially regenerated populations exists for mouse mammary cells (Shackleton et al., 2006; Stingl et al., 2006a), such data are extremely limited and do not exist for human mammary cells although bulk serial transplants have been performed (Eirew et al., 2008).

The results presented in this thesis confirm my initial hypothesis that when non-limiting numbers of basal cells are transplanted, a spectrum of differentiation activity can be observed from clones that regenerate in vivo – some clones display bi-lineage differentiation whereas others are lineage-restricted. However, I did not anticipate the

significant proportion of clones that would demonstrate a latent potential for growth and differentiation, which was a very significant observation here shared by both mouse and human basal mammary epithelial cells. In the syngeneic mouse transplant model, transplantation of non-limiting numbers of normal basal mammary epithelial cells demonstrated a diverse range of clonal growth patterns, that included initially luminal-restricted clones that displayed bi-lineage differentiation activity in secondary hosts, indicating a latent potential for bi-lineage differentiation. Furthermore, many clones that were not detectable after 8 weeks in a primary host then expanded in a secondary, indicating a latent growth potential, as well as a capacity for long-term survival (up to 8 weeks). Interestingly, the few initially basal-restricted clones did not demonstrate robust regeneration in secondary mice. In the human xenograft transplant model, however, all of the clones detected in a secondary xenograft were not previously detected in the primary, which alters our understanding of the use of measuring secondary MRU activity as a readout for "self-renewal", since the clones that demonstrate primary MRU activity may not be the same clones that demonstrate MRU activity in a secondary transplant.

These results contrast with previous reports suggesting that mouse basal mammary cells with regenerative activity *in vivo* appear bipotent at least when transplanted at limiting dilution (Shackleton et al., 2006; Stingl et al., 2006a). Our barcode results also contrast with those obtained from *in situ* lineage-tracing studies that suggest at times, the differentiation of BCs is restricted to the myoepithelial lineage (van Amerongen et al., 2012; Van Keymeulen et al., 2011). In addition, these latter lineage-tracing studies reported that luminal mammary cells sustain cells of the luminal lineage through several rounds of pregnancy, lactation and involution (Booth et al., 2007; Rios et

al., 2014; van Amerongen et al., 2012; Van Keymeulen et al., 2011; Wagner et al., 2002). However, in transplantation assays, the finding of rare *luminal* cells that can regenerate bi-lineage mammary structures post-transplant *in vivo* has been independently confirmed by two groups (Makarem et al., 2013; Shehata et al., 2012) and now also confirmed by our transplants of barcoded mouse luminal mammary cells. In addition, we confirm that human luminal mammary cells like their murine counterparts usually do not generate significant numbers of cells after 4 weeks *in vivo*. This variability in growth and differentiation displayed by individually assessed mouse and human mammary epithelial cells that nevertheless contribute to normally appearing tissue suggests that their ultimate regenerative activity may be subject to control mechanisms that act at multiple stages in their expansion. Thus single-cell or limiting dilution transplants may represent one extreme, where each cell is highly stimulated to display its capacity for growth, perhaps ultimately constrained only by mechanisms that operate at the whole tissue level (eg. the size of the cleared mammary fat pad). On the other hand, *in situ* lineage tracing studies may represent an opposite extreme where the environment is already nearly saturated with sufficient numbers of both mammary BCs and LCs and the stimulus for growth is limited to replacement requirements. These possibilities help reconcile the discrepancy in growth and differentiation activities observed in the transplantation studies (Nguyen et al., 2014; Shackleton et al., 2006; Stingl et al., 2006a) compared to *in situ* lineage tracing (van Amerongen et al., 2012; Van Keymeulen et al., 2011). It is important to note, however, that FACS isolation of regenerated basal and luminal cells from these transplants, and subsequent barcode analysis both have limits of detection that may result in the few basal cells of a luminal-restricted clones, or vice versa, to not be detected.

Thus, a luminal-restricted clone may be biased in its output of luminal cells and not truly "restricted" to a single lineage. In fact, we know this to be the case, since many of these luminal-restricted clones in the syngeneic mouse transplant model gave rise to robust bi-lineage clones even when only regenerated basal cells from the primary transplant were transplanted into secondary mice. The same applies to the newly detected clones in the secondary, which must have also been present in the primary mice (below the limit of detection). Nevertheless, it is still unclear how a clone that can demonstrate MRU activity upon serial *in vivo* transplantation relates to a clone that replenishes cells *in situ* under normal physiological demands.

The concept of latent regenerative cell potential being displayed was not anticipated in mammary cells, although it is notable that such a phenomenon has been shown in other tissues. For example, in both the skin and the gut, "reserve" populations of stem cells able to regenerate all components of the skin (Blanpain et al., 2004) or gut epithelium (Barker et al., 2007; Sangiorgi and Capecchi, 2008; Tian et al., 2011) have been revealed under conditions when the normally active stem cells are eliminated. A similar observation has been made in xenotransplants of human cord blood cells using both Southern blotting (Kreso et al., 2013) or barcoding to detect the emergence of new clones in secondary mice (Cheung et al., 2013).

5.3 Acquisition and analysis of tumor-initiating ability

We also describe the use of cellular barcoding to analyze the clonal growth of premalignant and established human mammary tumours; the latter including a single patient's primary xenograft as well as a cohort of tumours that we created *de novo* from

either isolated BCs or LPs isolated from normal human mammary tissue by transducing the cells with specific oncogenes ($KRAS^{G12D} \pm PI3KCA^{H1047R} \pm TP53^{R273C}$).

Interestingly, the frequency of clones that contributed to the BC-derived tumours was remarkably similar to the frequency of MRUs (1/500 BCs) determined from previously reported limiting dilution subrenal transplants, and to the frequency of regenerated clones produced from similarly transplanted barcoded BCs but at higher numbers. It is thus inviting to speculate that the normal cells most susceptible to transformation (in the system we employed) are those that already possess the potential for extensive cell output. However, this would not explain the even higher frequency of tumours obtained from normal LPs using the same oncogenes nor the fact that the frequency of clones in both LP and BC-derived tumours was also similar. It is possible that some normal human LPs possess robust bipotent growth capacity like their mouse counterparts, but the conditions to elicit such activity in human cells have not yet been discovered. Nevertheless, our findings do not support the concept that the nascent regenerative activity of the cell of origin is an essential driving factor in its ability to initiate tumour formation.

It is interesting that LPs not only produced tumours efficiently, they also lacked expected features including expression of HER2 and ERα which were, however, common in the tumours similarly derived from normal BCs. In addition, after a similar period of growth, the LP-derived tumours contained clones that were 4 to 8-fold larger than the largest clones in the tumours derived from BCs.

The concept of latent growth potential and its implications for tumorigenesis has been highlighted with the recent advances in the induction of pluripotency from normal

adult cells using more recent protocols that require only the transient expression of the "Yamanaka factors" (OCT3/4, SOX2, KLF4, and c-MYC)(Takahashi et al., 2007b; Takahashi and Yamanaka, 2006; Yu et al., 2007). Interestingly, it has been found using a mouse model that if induction of pluripotency is halted during the process, tumours develop (Ohnishi et al., 2014). Similarly, forced expression of the Yamanaka factors, specifically NANOG (thought to be critical for maintaining pluripotency)(Miyanari and Torres-Padilla, 2012) and MYC in mouse mammary cells, and OCT4 in human mammary cells induces tumorigenesis (Beltran et al., 2011; Horiuchi et al., 2012; Lu et al., 2013; Moumen et al., 2013). It would thus be informative to determine whether and when the process of tumorigenesis initiated here using $KRAS^{G12D}$, $PIK3CA^{H1047R}$ and $TP53^{R273C}$ might be reversible. Such a question might be addressed by first analyzing the transcriptomes of the *de novo* generated tumours, and comparing their profiles to either human embryonic stem cells or induced pluripotent stem cells. Then, it would be useful to understand at which point after expression of these mutant genes tumorigenesis becomes irreversible, and how this may be associated with certain epigenetic modifications and changes in gene expression. One such approach would be to design inducible vectors with which expression of these genes can be controlled. The changes that occur on a molecular level to induce tumorigenesis in LPs and BCs are also likely to be different, given their known transcriptome differences including transcription factors like NOTCH and GATA3 that are implicated in the determination of cell fate of normal mammary epithelial cells (Asselin-Labat et al., 2007; Bouras et al., 2008; Dontu et al., 2004; Raouf et al., 2008). An analysis of the transcriptome of the *de novo* tumours as compared to normal datasets of purified human BC and LP cells, would thus also be

informative to understand how changes in phenotype are induced during their transformation.

It has already been discussed in the discussion section of Chapter 4 that a limitation of this model, where we generate human breast tumours *de novo* in a forward fashion, is that these tumours form much more rapidly than those in patients and in some mouse models. However, it is possible that the frequency of tumour formation may actually be higher than I report here, and that because of the time restraints imposed by health problems incurred by the transplanted mice (induced by the use of slow-release estrogen and progesterone pellets), tumours that would appear after a longer latency period are not detected. From the experiments that examine the cell outputs at 2 and 4 weeks, it appears that some conditions ($KRAS^{G12D}$ alone) induce changes in clone size (detected by barcode analysis) as early as 2 weeks, whereas other changes (eg. $KRAS^{G12D}$+$TP53^{R273C}$, particularly in LPs) occur at 4 weeks (but not at 2 weeks). Thus, it would be interesting to investigate whether hormone pellet implants are necessary for tumour formation, and whether tumours of different characteristics may appear at a later timepoint, particularly for combinations of mutant genes (eg. $PIK3CA^{H1047R}$+$TP53^{R273C}$) that are common in spontaneous human breast tumours, but do not appear to induce any early changes in cell output or produce any tumours by 8 weeks from BCs or LPs. Another avenue of investigation not explored in my thesis, but is pertinent to understanding how the cell of origin influences the characteristics of resultant *de novo* breast tumours, is the effect of other oncogenes in inducing early changes in cell output and/or tumorigenesis.

5.4 Clinical implications

The findings presented in this thesis support the hypothesis that the cell of origin influences the phenotypic and functional characteristics of de novo generated human breast tumours. Perhaps the most interesting findings relevant to breast cancer are the observation that a majority of normal mouse and human *basal* mammary epithelial cells produced apparently *luminal-restricted* clones in transplanted primary mice, and that human BC-derived tumours proved to have an ER^+ phenotype (typically associated with luminal-like breast tumours)(Perou et al., 2000; Sorlie et al., 2001). Conversely, some transplanted mouse luminal mammary cells displayed bi-lineage differentiation ability, and human LP-derived tumours proved to be largely ER^- (typically associated with basal-like breast tumours)(Curtis et al., 2012; Perou et al., 2000; Shah et al., 2012; Sorlie et al., 2001). In addition, tumorigenesis appeared to be more frequently induced in LP cells than BCs, and LP-derived tumors were frequently larger than their BC-derived sample-matched counterparts. These findings are consistent with the natural history of luminal versus basal-like human breast tumours in patients, where tumours with a basal phenotype are typically more aggressive, and have a shorter latency period (Sorlie et al., 2001). Further characterization of these *de novo* generated tumours should help to determine how similar they are to breast cancer that arise in patients, for example, by comparison of their phenotypic, genomic and transcriptomic profiles. It will also be important to determine how these *de novo* generated tumours respond to hormone and targeted therapies and whether they develop subclones with metastatic properties. The work described in this thesis establishes the feasibility for such future studies and sets the stage for their pursuit.

5.5 Figures & tables

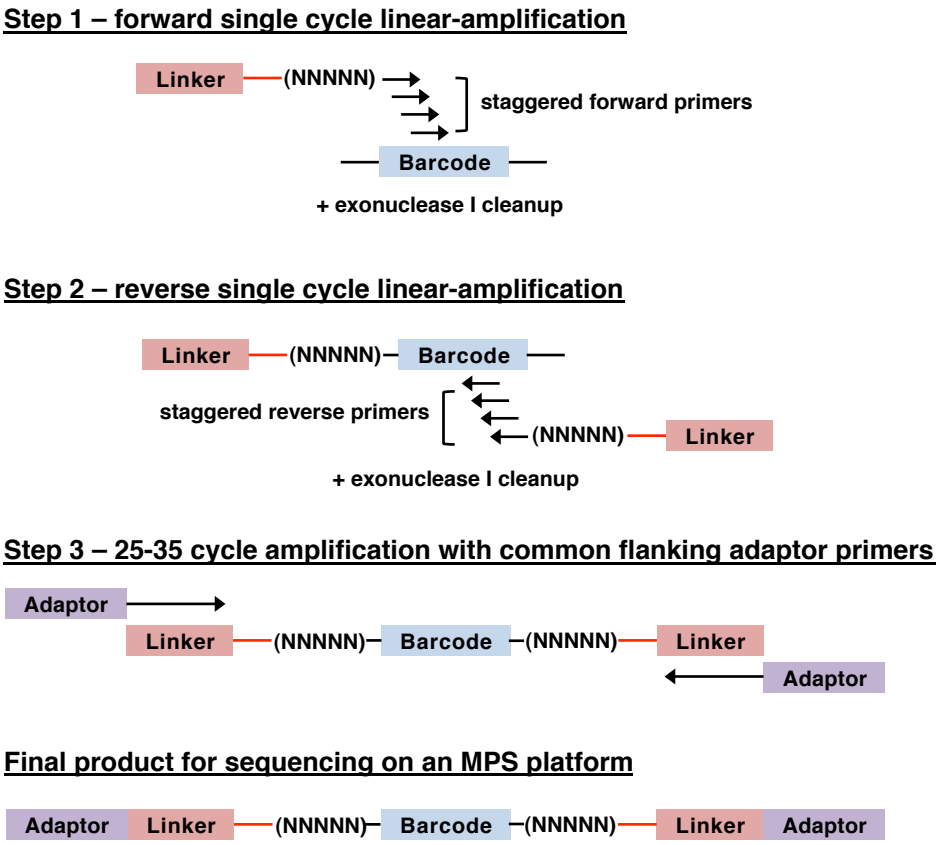

Figure 5.1 Proposed approach for the direct quantitation of barcoded clones

Depicted is a 3-step protocol to directly quantify the number of starting templates (and thus the number of cells) for barcode clones in experimental samples of unknown barcode composition. A single forward and reverse cycle of linear-amplification PCR serves to label the starting template strands with a unique secondary barcode (depicted as "(NNNNN)", that when amplified with common flanking primers (containing no secondary barcode), can then be analyzed and serve as a direct count of cell number per unique primary barcode clone (depicted in blue). The staggered primers serve to increase the complexity of the amplicon library for sequencing on the Illumina MPS platforms, so that cluster recognition is improved without a sample of spiked-in exogenous DNA.

REFERENCES

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proceedings of the National Academy of Sciences of the United States of America *100*, 3983-3988.

Aloia, L., Di Stefano, B., and Di Croce, L. (2013). Polycomb complexes in stem cells and embryonic development. Development *140*, 2525-2534.

Alvi, A.J., Clayton, H., Joshi, C., Enver, T., Ashworth, A., Vivanco, M., Dale, T.C., and Smalley, M.J. (2003). Functional and molecular characterisation of mammary side population cells. Breast cancer research : BCR *5*, R1-8.

Aparicio, S., and Caldas, C. (2013). The implications of clonal genome evolution for cancer medicine. The New England journal of medicine *368*, 842-851.

Asselin-Labat, M.L., Shackleton, M., Stingl, J., Vaillant, F., Forrest, N.C., Eaves, C.J., Visvader, J.E., and Lindeman, G.J. (2006). Steroid hormone receptor status of mouse mammary stem cells. Journal of the National Cancer Institute *98*, 1011-1014.

Asselin-Labat, M.L., Sutherland, K.D., Barker, H., Thomas, R., Shackleton, M., Forrest, N.C., Hartley, L., Robb, L., Grosveld, F.G., van der Wees, J.*, et al.* (2007). Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. Nature cell biology *9*, 201-209.

Asselin-Labat, M.L., Vaillant, F., Shackleton, M., Bouras, T., Lindeman, G.J., and Visvader, J.E. (2008). Delineating the epithelial hierarchy in the mouse mammary gland. Cold Spring Harbor symposia on quantitative biology *73*, 469-478.

Balinsky, B.I. (1950). On the prenatal growth of the mammary gland rudiment in the mouse. Journal of anatomy *84*, 227-235.

Barker, N., van Es, J.H., Kuipers, J., Kujala, P., van den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P.J.*, et al.* (2007). Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature *449*, 1003-1007.

Barnes, D.W.H., Ford, C.E., Gray, S.M., and Loutit, J.F. (1959). Spontaneous and induced changes in cell populations in heavily irradiated mice. Progress in Nuclear Energy - Biological sciences *2*, 1-10.

Bastien, R.R., Rodriguez-Lescure, A., Ebbert, M.T., Prat, A., Munarriz, B., Rowe, L., Miller, P., Ruiz-Borrego, M., Anderson, D., Lyons, B.*, et al.* (2012). PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC medical genomics *5*, 44.

Beltran, A.S., Rivenbark, A.G., Richardson, B.T., Yuan, X., Quian, H., Hunt, J.P., Zimmerman, E., Graves, L.M., and Blancafort, P. (2011). Generation of tumor-initiating

cells by exogenous delivery of OCT4 transcription factor. Breast cancer research : BCR *13*, R94.

Benz, C., Copley, M.R., Kent, D.G., Wohrer, S., Cortes, A., Aghaeepour, N., Ma, E., Mader, H., Rowe, K., Day, C.*, et al.* (2012). Hematopoietic stem cell subtypes expand differentially during development and display distinct lymphopoietic programs. Cell stem cell *10*, 273-283.

Beristain, A.G., Narala, S.R., Di Grappa, M.A., and Khokha, R. (2012). Homotypic RANK signaling differentially regulates proliferation, motility and cell survival in osteosarcoma and mammary epithelial cells. Journal of cell science *125*, 943-955.

Blanpain, C. (2013). Tracing the cellular origin of cancer. Nature cell biology *15*, 126-134.

Blanpain, C., Lowry, W.E., Geoghegan, A., Polak, L., and Fuchs, E. (2004). Self-renewal, multipotency, and the existence of two cell populations within an epithelial stem cell niche. Cell *118*, 635-648.

Blanpain, C., and Simons, B.D. (2013). Unravelling stem cell dynamics by lineage tracing. Nature reviews Molecular cell biology *14*, 489-502.

Bocchinfuso, W.P., and Korach, K.S. (1997). Estrogen receptor residues required for stereospecific ligand recognition and activation. Molecular endocrinology *11*, 587-594.

Bogden, A.E., Haskell, P.M., LePage, D.J., Kelton, D.E., Cobb, W.R., and Esber, H.J. (1979). Growth of human tumor xenografts implanted under the renal capsule of normal immunocompetent mice. Experimental cell biology *47*, 281-293.

Bonnefoix, T., and Callanan, M. (2009). Reassessing the human mammary stem cell concept by modeling limiting dilution transplantation assays. Nature medicine *15*, 602-4.

Booth, B.W., Boulanger, C.A., and Smith, G.H. (2007). Alveolar progenitor cells develop in mouse mammary glands independent of pregnancy and lactation. Journal of cellular physiology *212*, 729-736.

Bouras, T., Pal, B., Vaillant, F., Harburg, G., Asselin-Labat, M.L., Oakes, S.R., Lindeman, G.J., and Visvader, J.E. (2008). Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. Cell stem cell *3*, 429-441.

Brisken, C., and O'Malley, B. (2010). Hormone action in the mammary gland. Cold Spring Harbor perspectives in biology *2*, a003178.

Brisken, C., Park, S., Vass, T., Lydon, J.P., O'Malley, B.W., and Weinberg, R.A. (1998). A paracrine role for the epithelial progesterone receptor in mammary gland development. Proceedings of the National Academy of Sciences of the United States of America *95*, 5076-5081.

Buck, A.C. (1963). Differentiation of First-and Second-Set Grafts of Neonatal Testis, Ovary, Intestine and Spleen Implanated beneath the Kidney Capsule of Adult Albino Rat Hosts. The American journal of anatomy *113*, 189-213.

Bystrykh, L.V., Verovskaya, E., Zwart, E., Broekhuis, M., and de Haan, G. (2012). Counting stem cells: methodological constraints. Nature methods *9*, 567-574.

Caliskan, M., Gatti, G., Sosnovskikh, I., Rotmensz, N., Botteri, E., Musmeci, S., Rosali dos Santos, G., Viale, G., and Luini, A. (2008). Paget's disease of the breast: the experience of the European Institute of Oncology and review of the literature. Breast cancer research and treatment *112*, 513-521.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cariati, M., Marlow, R., and Dontu, G. (2011). Xenotransplantation of breast cancers. Methods in molecular biology *731*, 471-482.

Chaffer, C.L., Brueckmann, I., Scheel, C., Kaestli, A.J., Wiggins, P.A., Rodrigues, L.O., Brooks, M., Reinhardt, F., Su, Y., Polyak, K.*, et al.* (2011). Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. Proceedings of the National Academy of Sciences of the United States of America *108*, 7950-7955.

Cheung, A.M., Nguyen, L.V., Carles, A., Beer, P., Miller, P.H., Knapp, D.J., Dhillon, K., Hirst, M., and Eaves, C.J. (2013). Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. Blood *122*, 3129-3137.

Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265*, 346-355.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.*, et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352.

Daniel, C.W., De Ome, K.B., Young, J.T., Blair, P.B., and Faulkin, L.J., Jr. (1968). The in vivo life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. Proceedings of the National Academy of Sciences of the United States of America *61*, 53-60.

Daniel, C.W., Silberstein, G.B., and Strickland, P. (1987). Direct action of 17 beta-estradiol on mouse mammary ducts analyzed by sustained release implants and steroid autoradiography. Cancer research *47*, 6052-6057.

Daniel, C.W., and Young, L.J. (1971). Influence of cell division on an aging process. Life span of mouse mammary epithelium during serial propagation in vivo. Experimental cell research *65*, 27-32.

Dawson, S.J., Rueda, O.M., Aparicio, S., and Caldas, C. (2013). A new genome-driven integrated classification of breast cancer and its implications. The EMBO journal *32*, 617-628.

Deome, K.B., Faulkin, L.J., Jr., Bern, H.A., and Blair, P.B. (1959). Development of mammary tumors from hyperplastic alveolar nodules transplanted into gland-free mammary fat pads of female C3H mice. Cancer research *19*, 515-520.

DeRose, Y.S., Wang, G., Lin, Y.C., Bernard, P.S., Buys, S.S., Ebbert, M.T., Factor, R., Matsen, C., Milash, B.A., Nelson, E.*, et al.* (2011). Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. Nature medicine *17*, 1514-1520.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L.*, et al.* (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999-1005.

Dontu, G., Abdallah, W.M., Foley, J.M., Jackson, K.W., Clarke, M.F., Kawamura, M.J., and Wicha, M.S. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. Genes & development *17*, 1253-1270.

Dontu, G., Jackson, K.W., McNicholas, E., Kawamura, M.J., Abdallah, W.M., and Wicha, M.S. (2004). Role of Notch signaling in cell-fate determination of human mammary stem/progenitor cells. Breast cancer research : BCR *6*, R605-615.

Duss, S., Andre, S., Nicoulaz, A.L., Fiche, M., Bonnefoi, H., Brisken, C., and Iggo, R.D. (2007). An oestrogen-dependent model of breast cancer created by transformation of normal human mammary epithelial cells. Breast cancer research : BCR *9*, R38.

Edge, S., Byrd, D.R., Compton, C.C., Fritz, A.G., Green, F.L., and Trotti, A., eds. (2010). AJCC Cancer Staging Manual, 7th edn.

Eirew, P., Stingl, J., and Eaves, C.J. (2010). Quantitation of human mammary epithelial stem cells with in vivo regenerative properties using a subrenal capsule xenotransplantation assay. Nature protocols *5*, 1945-1956.

Eirew, P., Stingl, J., Raouf, A., Turashvili, G., Aparicio, S., Emerman, J.T., and Eaves, C.J. (2008). A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. Nature medicine *14*, 1384-1389.

Ellis, M.J., Suman, V.J., Hoog, J., Lin, L., Snider, J., Prat, A., Parker, J.S., Luo, J., DeSchryver, K., Allred, D.C.*, et al.* (2011). Randomized phase II neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: clinical and biomarker outcomes and predictive value of the baseline PAM50-based intrinsic subtype--ACOSOG Z1031. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *29*, 2342-2349.

Emerman, J.T., Stingl, J., Petersen, A., Shpall, E.J., and Eaves, C.J. (1996). Selective growth of freshly isolated human breast epithelial cells cultured at low concentrations in the presence or absence of bone marrow cells. Breast cancer research and treatment *41*, 147-159.

Fata, J.E., Kong, Y.Y., Li, J., Sasaki, T., Irie-Sasaki, J., Moorehead, R.A., Elliott, R., Scully, S., Voura, E.B., Lacey, D.L.*, et al.* (2000). The osteoclast differentiation factor osteoprotegerin-ligand is essential for mammary gland development. Cell *103*, 41-50.

Feng, G., Mellor, R.H., Bernstein, M., Keller-Peck, C., Nguyen, Q.T., Wallace, M., Nerbonne, J.M., Lichtman, J.W., and Sanes, J.R. (2000). Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. Neuron *28*, 41-51.

Fialkow, P.J., Gartler, S.M., and Yoshida, A. (1967). Clonal origin of chronic myelocytic leukemia in man. Proceedings of the National Academy of Sciences of the United States of America *58*, 1468-1471.

Fisher, C.R., Graves, K.H., Parlow, A.F., and Simpson, E.R. (1998). Characterization of mice deficient in aromatase (ArKO) because of targeted disruption of the cyp19 gene. Proceedings of the National Academy of Sciences of the United States of America *95*, 6965-6970.

Fisher, E.R., Land, S.R., Fisher, B., Mamounas, E., Gilarski, L., and Wolmark, N. (2004). Pathologic findings from the National Surgical Adjuvant Breast and Bowel Project: twelve-year observations concerning lobular carcinoma in situ. Cancer *100*, 238-244.

Fonseca, R., Hartmann, L.C., Petersen, I.A., Donohue, J.H., Crotty, T.B., and Gisvold, J.J. (1997). Ductal carcinoma in situ of the breast. Annals of internal medicine *127*, 1013-1022.

Fu, N., Lindeman, G.J., and Visvader, J.E. (2014). The mammary stem cell hierarchy. Current topics in developmental biology *107*, 133-160.

Gabriel, R., Eckenberg, R., Paruzynski, A., Bartholomae, C.C., Nowrouzi, A., Arens, A., Howe, S.J., Recchia, A., Cattoglio, C., Wang, W.*, et al.* (2009). Comprehensive genomic access to vector integration in clinical gene therapy. Nature medicine *15*, 1431-1436.

Gentner, B., Laufs, S., Nagy, K.Z., Zeller, W.J., and Fruehauf, S. (2003). Rapid detection of retroviral vector integration sites in colony-forming human peripheral blood progenitor cells using PCR with arbitrary primers. Gene therapy *10*, 789-794.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P.*, et al.* (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine *366*, 883-892.

Gerrits, A., Dykstra, B., Kalmykowa, O.J., Klauke, K., Verovskaya, E., Broekhuis, M.J., de Haan, G., and Bystrykh, L.V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood *115*, 2610-2618.

Gonzalez-Angulo, A.M., Iwamoto, T., Liu, S., Chen, H., Do, K.A., Hortobagyi, G.N., Mills, G.B., Meric-Bernstam, F., Symmans, W.F., and Pusztai, L. (2012). Gene expression, molecular class changes, and pathway analysis after neoadjuvant systemic therapy for breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research *18*, 1109-1119.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306-313.

Grosselin, J., Sii-Felice, K., Payen, E., Chretien, S., Roux, D.T., and Leboulch, P. (2013). Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. Stem cells *31*, 2162-2171.

Gu, B., Sun, P., Yuan, Y., Moraes, R.C., Li, A., Teng, A., Agrawal, A., Rheaume, C., Bilanchone, V., Veltmaat, J.M.*, et al.* (2009). Pygo2 expands mammary progenitor cells by facilitating histone H3 K4 methylation. The Journal of cell biology *185*, 811-826.

Gu, B., Watanabe, K., and Dai, X. (2012). Pygo2 regulates histone gene expression and H3 K56 acetylation in human mammary epithelial cells. Cell cycle *11*, 79-87.

Hadjantonakis, A.K., Macmaster, S., and Nagy, A. (2002). Embryonic stem cells and mice expressing different GFP variants for multiple non-invasive reporter usage within a single animal. BMC biotechnology *2*, 11.

Hahn, W.C., Counter, C.M., Lundberg, A.S., Beijersbergen, R.L., Brooks, M.W., and Weinberg, R.A. (1999). Creation of human tumour cells with defined genetic elements. Nature *400*, 464-468.

Harkey, M.A., Kaul, R., Jacobs, M.A., Kurre, P., Bovee, D., Levy, R., and Blau, C.A. (2007). Multiarm high-throughput integration site detection: limitations of LAM-PCR technology and optimization for clonal analysis. Stem cells and development *16*, 381-392.

Herschkowitz, J.I., Simin, K., Weigman, V.J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K.E., Jones, L.P., Assefnia, S., Chandrasekharan, S.*, et al.* (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome biology *8*, R76.

Hope, K., and Bhatia, M. (2011). Clonal interrogation of stem cells. Nature methods *8*, S36-40.

Horiuchi, D., Kusdra, L., Huskey, N.E., Chandriani, S., Lenburg, M.E., Gonzalez-Angulo, A.M., Creasman, K.J., Bazarov, A.V., Smyth, J.W., Davis, S.E.*, et al.* (2012). MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. The Journal of experimental medicine *209*, 679-696.

Howard, B.A. (2012). In the beginning: the establishment of the mammary lineage during embryogenesis. Seminars in cell & developmental biology *23*, 574-582.

Hu, Y., and Smyth, G.K. (2009). ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. Journal of immunological methods *347*, 70-8.

Imagawa, W., Bandyopadhyay, G.K., and Nandi, S. (1990). Regulation of mammary epithelial cell growth in mice and rats. Endocrine reviews *11*, 494-523.

Imren, S., Fabry, M.E., Westerman, K.A., Pawliuk, R., Tang, P., Rosten, P.M., Nagel, R.L., Leboulch, P., Eaves, C.J., and Humphries, R.K. (2004). High-level beta-globin expression and preferred intragenic integration after lentiviral transduction of human cord blood stem cells. The Journal of clinical investigation *114*, 953-962.

Jessen, S., Gu, B., and Dai, X. (2008). Pygopus and the Wnt signaling pathway: a diverse set of connections. BioEssays : news and reviews in molecular, cellular and developmental biology *30*, 448-456.

Joshi, P.A., Jackson, H.W., Beristain, A.G., Di Grappa, M.A., Mote, P.A., Clarke, C.L., Stingl, J., Waterhouse, P.D., and Khokha, R. (2010). Progesterone induces adult mammary stem cell expansion. Nature *465*, 803-807.

Joshi, P.A., and Khokha, R. (2012). The mammary stem cell conundrum: is it unipotent or multipotent? Breast cancer research : BCR *14*, 305.

Kannan, N., Huda, N., Tu, L., Droumeva, R., Aubert, G., Chavez, E., Brinkman, R.R., Lansdorp, P., Emerman, J., Abe, S.*, et al.* (2013). The luminal progenitor compartment of the normal human mammary gland constitutes a unique site of telomere dysfunction. Stem cell reports *1*, 28-37.

Kannan, N., Makarem, M., Nguyen, L.V., Dong, J., Eirew, P., Raouf, A., Emerman, J., and Eaves, C.J. (submitted). Glutathione-dependent and independent oxidative-stress control mechanisms distinguish normal human mammary epithelial cell subsets.

Keller, P.J., Arendt, L.M., Skibinski, A., Logvinenko, T., Klebba, I., Dong, S., Smith, A.E., Prat, A., Perou, C.M., Gilmore, H.*, et al.* (2012). Defining the cellular precursors to human breast cancer. Proceedings of the National Academy of Sciences of the United States of America *109*, 2772-2777.

Kendall, S.D., Linardic, C.M., Adam, S.J., and Counter, C.M. (2005). A network of genetic events sufficient to convert normal human cells to a tumorigenic state. Cancer research *65*, 9824-9828.

Kim, H.S., Park, I., Cho, H.J., Gwak, G., Yang, K., Bae, B.N., Kim, K.W., Han, S., Kim, H.J., and Kim, Y.D. (2012). Analysis of the potent prognostic factors in luminal-type breast cancer. Journal of breast cancer *15*, 401-406.

Kordon, E.C., and Smith, G.H. (1998). An entire functional mammary gland may comprise the progeny from a single cell. Development *125*, 1921-1930.

Krege, J.H., Hodgin, J.B., Couse, J.F., Enmark, E., Warner, M., Mahler, J.F., Sar, M., Korach, K.S., Gustafsson, J.A., and Smithies, O. (1998). Generation and reproductive phenotypes of mice lacking estrogen receptor beta. Proceedings of the National Academy of Sciences of the United States of America *95*, 15677-15682.

Kreso, A., O'Brien, C.A., van Galen, P., Gan, O.I., Notta, F., Brown, A.M., Ng, K., Ma, J., Wienholds, E., Dunant, C.*, et al.* (2013). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science *339*, 543-548.

Kuperwasser, C., Chavarria, T., Wu, M., Magrane, G., Gray, J.W., Carey, L., Richardson, A., and Weinberg, R.A. (2004). Reconstruction of functionally normal and malignant human breast tissues in mice. Proceedings of the National Academy of Sciences of the United States of America *101*, 4966-4971.

Kustikova, O., Fehse, B., Modlich, U., Yang, M., Dullmann, J., Kamino, K., von Neuhoff, N., Schlegelberger, B., Li, Z., and Baum, C. (2005). Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science *308*, 1171-1174.

Lee, C.H., Xue, H., Sutcliffe, M., Gout, P.W., Huntsman, D.G., Miller, D.M., Gilks, C.B., and Wang, Y.Z. (2005). Establishment of subrenal capsule xenografts of primary human ovarian tumors in SCID mice: potential models. Gynecologic oncology *96*, 48-55.

Lee, H.J., Gallego-Ortega, D., Ledger, A., Schramek, D., Joshi, P., Szwarc, M.M., Cho, C., Lydon, J.P., Khokha, R., Penninger, J.M.*, et al.* (2013). Progesterone drives mammary secretory differentiation via RankL-mediated induction of Elf5 in luminal progenitor cells. Development *140*, 1397-1401.

Lee, H.J., Hinshelwood, R.A., Bouras, T., Gallego-Ortega, D., Valdes-Mora, F., Blazek, K., Visvader, J.E., Clark, S.J., and Ormandy, C.J. (2011). Lineage specific methylation of the Elf5 promoter in mammary epithelial cells. Stem cells *29*, 1611-1619.

Lim, E., Vaillant, F., Wu, D., Forrest, N.C., Pal, B., Hart, A.H., Asselin-Labat, M.L., Gyorki, D.E., Ward, T., Partanen, A.*, et al.* (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature medicine *15*, 907-913.

Ling, J., Kang, Y., Zhao, R., Xia, Q., Lee, D.F., Chang, Z., Li, J., Peng, B., Fleming, J.B., Wang, H.*, et al.* (2012). KrasG12D-induced IKK2/beta/NF-kappaB activation by IL-1alpha and p62 feedforward loops is required for development of pancreatic ductal adenocarcinoma. Cancer cell *21*, 105-120.

Liu, S., Dontu, G., Mantle, I.D., Patel, S., Ahn, N.S., Jackson, K.W., Suri, P., and Wicha, M.S. (2006). Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells. Cancer research *66*, 6063-6071.

Livet, J., Weissman, T.A., Kang, H., Draft, R.W., Lu, J., Bennis, R.A., Sanes, J.R., and Lichtman, J.W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature *450*, 56-62.

Logan, A.C., Nightingale, S.J., Haas, D.L., Cho, G.J., Pepper, K.A., and Kohn, D.B. (2004). Factors influencing the titer and infectivity of lentiviral vectors. Human gene therapy *15*, 976-988.

Lu, R., Neff, N.F., Quake, S.R., and Weissman, I.L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature biotechnology *29*, 928-933.

Lu, X., Mazur, S.J., Lin, T., Appella, E., and Xu, Y. (2013). The pluripotency factor nanog promotes breast cancer tumorigenesis and metastasis. Oncogene.

Luis, N.M., Morey, L., Di Croce, L., and Benitah, S.A. (2012). Polycomb in stem cells: PRC1 branches out. Cell stem cell *11*, 16-21.

Makarem, M., Kannan, N., Nguyen, L.V., Knapp, D.J., Balani, S., Prater, M.D., Stingl, J., Raouf, A., Nemirovsky, O., Eirew, P.*, et al.* (2013). Developmental changes in the in vitro activated regenerative activity of primitive mammary epithelial cells. PLoS biology *11*, e1001630.

Mallepell, S., Krust, A., Chambon, P., and Brisken, C. (2006). Paracrine signaling through the epithelial estrogen receptor alpha is required for proliferation and morphogenesis in the mammary gland. Proceedings of the National Academy of Sciences of the United States of America *103*, 2196-2201.

Mikkers, H., Allen, J., Knipscheer, P., Romeijn, L., Hart, A., Vink, E., and Berns, A. (2002). High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nature genetics *32*, 153-159.

Miyanari, Y., and Torres-Padilla, M.E. (2012). Control of ground-state pluripotency by allelic regulation of Nanog. Nature *483*, 470-473.

Molyneux, G., Geyer, F.C., Magnay, F.A., McCarthy, A., Kendrick, H., Natrajan, R., Mackay, A., Grigoriadis, A., Tutt, A., Ashworth, A.*, et al.* (2010). BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. Cell stem cell *7*, 403-417.

Moumen, M., Chiche, A., Decraene, C., Petit, V., Gandarillas, A., Deugnier, M.A., Glukhova, M.A., and Faraldo, M.M. (2013). Myc is required for beta-catenin-mediated mammary stem cell amplification and tumorigenesis. Molecular cancer *12*, 132.

Mueller, S.O., Clark, J.A., Myers, P.H., and Korach, K.S. (2002). Mammary gland development in adult mice requires epithelial and stromal estrogen receptor alpha. Endocrinology *143*, 2357-2365.

Naik, S.H., Perie, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., and Schumacher, T.N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature *496*, 229-232.

NCI (2014). Breast Cancer Treatment (National Institutes of Health).

Nguyen, L.V., Makarem, M., Carles, A., Moksa, M., Kannan, N., Pandoh, P., Eirew, P., Osako, T., Kardel, M., Cheung, A.M.*, et al.* (2014). Clonal Analysis via Barcoding Reveals Diverse Growth and Differentiation of Transplanted Mouse and Human Mammary Stem Cells. Cell stem cell *14*, 253-263.

Nguyen, L.V., Vanner, R., Dirks, P., and Eaves, C.J. (2012). Cancer stem cells: an evolving concept. Nature reviews Cancer *12*, 133-143.

Niho, S., Yokose, T., Suzuki, K., Kodama, T., Nishiwaki, Y., and Mukai, K. (1999). Monoclonality of atypical adenomatous hyperplasia of the lung. The American journal of pathology *154*, 249-254.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* (2012). The life history of 21 breast cancers. Cell *149*, 994-1007.

Novaro, V., Roskelley, C.D., and Bissell, M.J. (2003). Collagen-IV and laminin-1 regulate estrogen receptor alpha expression and function in mouse mammary epithelial cells. Journal of cell science *116*, 2975-2986.

Nowell, P.C., and Hungerford, D.A. (1960). Chromosome studies on normal and leukemic human leukocytes. Journal of the National Cancer Institute *25*, 85-109.

Oakes, S.R., Naylor, M.J., Asselin-Labat, M.L., Blazek, K.D., Gardiner-Garden, M., Hilton, H.N., Kazlauskas, M., Pritchard, M.A., Chodosh, L.A., Pfeffer, P.L.*, et al.* (2008). The Ets transcription factor Elf5 specifies mammary alveolar cell fate. Genes & development *22*, 581-586.

Ohnishi, K., Semi, K., Yamamoto, T., Shimizu, M., Tanaka, A., Mitsunaga, K., Okita, K., Osafune, K., Arioka, Y., Maeda, T.*, et al.* (2014). Premature Termination of Reprogramming In Vivo Leads to Cancer Development through Altered Epigenetic Regulation. Cell *156*, 663-677.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T.*, et al.* (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine *351*, 2817-2826.

Pal, B., Bouras, T., Shi, W., Vaillant, F., Sheridan, J.M., Fu, N., Breslin, K., Jiang, K., Ritchie, M.E., Young, M.*, et al.* (2013). Global changes in the mammary epigenome are induced by hormonal cues and coordinated by Ezh2. Cell reports *3*, 411-426.

Perie, L., Hodgkin, P.D., Naik, S.H., Schumacher, T.N., de Boer, R.J., and Duffy, K.R. (2014). Determining Lineage Pathways from Cellular Barcoding Experiments. Cell reports.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A.*, et al.* (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.

Pham, T.T., Angus, S.P., and Johnson, G.L. (2013). MAP3K1: Genomic Alterations in Cancer and Function in Promoting Cell Survival or Apoptosis. Genes & cancer *4*, 419-426.

Pietersen, A.M., Evers, B., Prasad, A.A., Tanger, E., Cornelissen-Steijger, P., Jonkers, J., and van Lohuizen, M. (2008). Bmi1 regulates stem cells and proliferation and differentiation of committed cells in mammary epithelium. Current biology : CB *18*, 1094-1099.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast cancer research : BCR *12*, R68.

Prat, A., and Perou, C.M. (2011). Deconstructing the molecular portraits of breast cancer. Molecular oncology *5*, 5-23.

Press, M.F., Sauter, G., Bernstein, L., Villalobos, I.E., Mirlacher, M., Zhou, J.Y., Wardeh, R., Li, Y.T., Guzman, R., Ma, Y.*, et al.* (2005). Diagnostic evaluation of HER-2 as a molecular target: an assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. Clinical cancer research : an official journal of the American Association for Cancer Research *11*, 6598-6607.

Proia, D.A., and Kuperwasser, C. (2006). Reconstruction of human mammary tissues in a mouse model. Nature protocols *1*, 206-214.

Proia, T.A., Keller, P.J., Gupta, P.B., Klebba, I., Jones, A.D., Sedic, M., Gilmore, H., Tung, N., Naber, S.P., Schnitt, S.*, et al.* (2011). Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate. Cell stem cell *8*, 149-163.

Raouf, A., Brown, L., Vrcelj, N., To, K., Kwok, W., Huntsman, D., and Eaves, C.J. (2005). Genomic instability of human mammary epithelial cells overexpressing a truncated form of EMSY. Journal of the National Cancer Institute *97*, 1302-1306.

Raouf, A., Zhao, Y., To, K., Stingl, J., Delaney, A., Barbara, M., Iscove, N., Jones, S., McKinney, S., Emerman, J.*, et al.* (2008). Transcriptome analysis of the normal human mammary cell commitment and differentiation process. Cell stem cell *3*, 109-118.

Regan, J.L., Kendrick, H., Magnay, F.A., Vafaizadeh, V., Groner, B., and Smalley, M.J. (2012). c-Kit is required for growth and survival of the cells of origin of Brca1-mutation-associated breast cancer. Oncogene *31*, 869-883.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenetics and cell genetics *14*, 9-25.

Rios, A.C., Fu, N.Y., Lindeman, G.J., and Visvader, J.E. (2014). In situ identification of bipotent stem cells in the mammary gland. Nature.

Russo, J., and Russo, I.H. (1987). The Mammary Gland (NY: Plenum Publishing Corporation).

Russo, J., and Russo, I.H. (2004). Development of the human breast. Maturitas *49*, 2-15.

Sangiorgi, E., and Capecchi, M.R. (2008). Bmi1 is expressed in vivo in intestinal stem cells. Nature genetics *40*, 915-920.

Schepers, K., Swart, E., van Heijst, J.W., Gerlach, C., Castrucci, M., Sie, D., Heimerikx, M., Velds, A., Kerkhoven, R.M., Arens, R.*, et al.* (2008). Dissecting T cell lineage relationships by cellular barcoding. The Journal of experimental medicine *205*, 2309-2318.

Schmidt, M., Hoffmann, G., Wissler, M., Lemke, N., Mussig, A., Glimm, H., Williams, D.A., Ragg, S., Hesemann, C.U., and von Kalle, C. (2001). Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Human gene therapy *12*, 743-749.

Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell *110*, 521-529.

Shackleton, M., Vaillant, F., Simpson, K.J., Stingl, J., Smyth, G.K., Asselin-Labat, M.L., Wu, L., Lindeman, G.J., and Visvader, J.E. (2006). Generation of a functional mammary gland from a single stem cell. Nature *439*, 84-88.

Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J.*, et al.* (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809-813.

Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G.*, et al.* (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395-399.

Sheffield, L.G., and Welsch, C.W. (1988). Transplantation of human breast epithelia to mammary-gland-free fat-pads of athymic nude mice: influence of mammotrophic hormones on growth of breast epithelia. International journal of cancer Journal international du cancer *41*, 713-719.

Shehata, M., Teschendorff, A., Sharp, G., Novcic, N., Russell, A., Avril, S., Prater, M., Eirew, P., Caldas, C., Watson, C.J.*, et al.* (2012). Phenotypic and functional

characterization of the luminal cell hierarchy of the mammary gland. Breast cancer research : BCR *14*, R134.

Shenkier, T., Weir, L., Levine, M., Olivotto, I., Whelan, T., Reyno, L., Steering Committee on Clinical Practice Guidelines for the, C., and Treatment of Breast, C. (2004). Clinical practice guidelines for the care and treatment of breast cancer: 15. Treatment for women with stage III or locally advanced breast cancer. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne *170*, 983-994.

Smalley, M.J., Titley, J., and O'Hare, M.J. (1998). Clonal characterization of mouse mammary luminal epithelial and myoepithelial cells separated by fluorescence-activated cell sorting. In vitro cellular & developmental biology Animal *34*, 711-721.

Snippert, H.J., and Clevers, H. (2011). Tracking adult stem cells. EMBO reports *12*, 113-122.

Snippert, H.J., van der Flier, L.G., Sato, T., van Es, J.H., van den Born, M., Kroon-Veenboer, C., Barker, N., Klein, A.M., van Rheenen, J., Simons, B.D.*, et al.* (2010). Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. Cell *143*, 134-144.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S.*, et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America *98*, 10869-10874.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B.*, et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute *98*, 262-272.

Spike, B.T., Engle, D.D., Lin, J.C., Cheung, S.K., La, J., and Wahl, G.M. (2012). A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. Cell stem cell *10*, 183-197.

Stingl, J., Eaves, C.J., Kuusk, U., and Emerman, J.T. (1998). Phenotypic and functional characterization in vitro of a multipotent epithelial cell present in the normal adult human breast. Differentiation; research in biological diversity *63*, 201-213.

Stingl, J., Eaves, C.J., Zandieh, I., and Emerman, J.T. (2001). Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. Breast cancer research and treatment *67*, 93-109.

Stingl, J., Eirew, P., Ricketson, I., Shackleton, M., Vaillant, F., Choi, D., Li, H.I., and Eaves, C.J. (2006a). Purification and unique properties of mammary epithelial stem cells. Nature *439*, 993-997.

Stingl, J., Emerman, J.T., and Eaves, C.J. (2005). Enzymatic dissociation and culture of normal human mammary tissue to detect progenitor activity. Methods in molecular biology *290*, 249-263.

Stingl, J., Raouf, A., Eirew, P., and Eaves, C.J. (2006b). Deciphering the mammary epithelial cell hierarchy. Cell cycle *5*, 1519-1522.

Takahashi, K., Kohno, T., Matsumoto, S., Nakanishi, Y., Arai, Y., Yamamoto, S., Fujiwara, T., Tanaka, N., and Yokota, J. (2007a). Clonal and parallel evolution of primary lung cancers and their metastases revealed by molecular dissection of cancer cells. Clinical cancer research : an official journal of the American Association for Cancer Research *13*, 111-120.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007b). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Tanos, T., Rojo, L.J., Echeverria, P., and Brisken, C. (2012). ER and PR signaling nodes during mammary gland development. Breast cancer research : BCR *14*, 210.

Teng, D.H., Perry, W.L., 3rd, Hogan, J.K., Baumgard, M., Bell, R., Berry, S., Davis, T., Frank, D., Frye, C., Hattier, T.*, et al.* (1997). Human mitogen-activated protein kinase kinase 4 as a candidate tumor suppressor. Cancer research *57*, 4177-4182.

Tian, H., Biehs, B., Warming, S., Leong, K.G., Rangell, L., Klein, O.D., and de Sauvage, F.J. (2011). A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. Nature *478*, 255-259.

Tikoo, A., Roh, V., Montgomery, K.G., Ivetac, I., Waring, P., Pelzer, R., Hare, L., Shackleton, M., Humbert, P., and Phillips, W.A. (2012). Physiological levels of Pik3ca(H1047R) mutation in the mouse mammary gland results in ductal hyperplasia and formation of ERalpha-positive tumors. PloS one *7*, e36924.

Tsai, Y.C., Lu, Y., Nichols, P.W., Zlotnikov, G., Jones, P.A., and Smith, H.S. (1996). Contiguous patches of normal human mammary epithelium derived from a single stem cell: implications for breast carcinogenesis. Cancer research *56*, 402-404.

Tutt, A., Wang, A., Rowland, C., Gillett, C., Lau, K., Chew, K., Dai, H., Kwok, S., Ryder, K., Shu, H.*, et al.* (2008). Risk estimation of distant metastasis in node-negative, estrogen receptor-positive breast cancer patients using an RT-PCR based prognostic expression signature. BMC cancer *8*, 339.

van Amerongen, R., Bowman, A.N., and Nusse, R. (2012). Developmental stage and time dictate the fate of Wnt/beta-catenin-responsive stem cells in the mammary gland. Cell stem cell *11*, 387-400.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., *et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine *347*, 1999-2009.

van Dijk, J.P., Heuver, L.H., van der Reijden, B.A., Raymakers, R.A., de Witte, T., and Jansen, J.H. (2002). A novel, essential control for clonality analysis with human androgen receptor gene polymerase chain reaction. The American journal of pathology *161*, 807-812.

Van Keymeulen, A., Rocha, A.S., Ousset, M., Beck, B., Bouvencourt, G., Rock, J., Sharma, N., Dekoninck, S., and Blanpain, C. (2011). Distinct stem cells contribute to mammary gland development and maintenance. Nature *479*, 189-193.

Veltmaat, J.M., Mailleux, A.A., Thiery, J.P., Bellusci, S. (2003). Mouse embryonic mammogenesis as a model for the molecular regulation of pattern formation. Differentiation *71*, 1-17.

Villadsen, R., Fridriksdottir, A.J., Ronnov-Jessen, L., Gudjonsson, T., Rank, F., LaBarge, M.A., Bissell, M.J., and Petersen, O.W. (2007). Evidence for a stem cell hierarchy in the adult human breast. The Journal of cell biology *177*, 87-101.

Visvader, J.E. (2009). Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. Genes & development *23*, 2563-2577.

Visvader, J.E. (2011). Cells of origin in cancer. Nature *469*, 314-322.

Visvader, J.E., and Lindeman, G.J. (2011). The unmasking of novel unipotent stem cells in the mammary gland. The EMBO journal *30*, 4858-4859.

Vogelstein, B., Fearon, E.R., Hamilton, S.R., and Feinberg, A.P. (1985). Use of restriction fragment length polymorphisms to determine the clonal origin of human tumors. Science *227*, 642-645.

Vogelstein, B., Fearon, E.R., Hamilton, S.R., Preisinger, A.C., Willard, H.F., Michelson, A.M., Riggs, A.D., and Orkin, S.H. (1987). Clonal analysis using recombinant DNA probes from the X-chromosome. Cancer research *47*, 4806-4813.

Wagner, K.U., Boulanger, C.A., Henry, M.D., Sgagias, M., Hennighausen, L., and Smith, G.H. (2002). An adjunct mammary epithelial cell population in parous females: its role in functional adaptation and tissue renewal. Development *129*, 1377-1386.

Welm, B.E., Tepera, S.B., Venezia, T., Graubert, T.A., Rosen, J.M., and Goodell, M.A. (2002). Sca-1(pos) cells in the mouse mammary gland represent an enriched progenitor cell population. Developmental biology *245*, 42-56.

Wiegand, K.C., Shah, S.P., Al-Agha, O.M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M.K., Anglesio, M.S., Kalloger, S.E., *et al.* (2010). ARID1A mutations in

endometriosis-associated ovarian carcinomas. The New England journal of medicine *363*, 1532-1543.

Willman, C.L., Busque, L., Griffith, B.B., Favara, B.E., McClain, K.L., Duncan, M.H., and Gilliland, D.G. (1994). Langerhans'-cell histiocytosis (histiocytosis X)--a clonal proliferative disease. The New England journal of medicine *331*, 154-160.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science *300*, 1749-1751.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A.*, et al.* (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114-1117.

Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell *154*, 1112-1126.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R.*, et al.* (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917-1920.

Zeng, Y.A., and Nusse, R. (2010). Wnt proteins are self-renewal factors for mammary stem cells and promote their long-term expansion in culture. Cell stem cell *6*, 568-577.

Zhang, X., Claerhout, S., Prat, A., Dobrolecki, L.E., Petrovic, I., Lai, Q., Landis, M.D., Wiechmann, L., Schiff, R., Giuliano, M.*, et al.* (2013). A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. Cancer research *73*, 4885-4897.

Zhou, J., Chehab, R., Tkalcevic, J., Naylor, M.J., Harris, J., Wilson, T.J., Tsao, S., Tellis, I., Zavarsek, S., Xu, D.*, et al.* (2005). Elf5 is essential for early embryogenesis and mammary gland development during pregnancy and lactation. The EMBO journal *24*, 635-644.