UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Defining Cellular Heterogeneity within the Normal Human Breast Epithelial System and BRCA1 Mutation Carriers

Permalink

https://escholarship.org/uc/item/5w73k1wg

Author

Nguyen, Quy Hoa

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Defining Cellular Heterogeneity within the Normal Human Breast Epithelial System and BRCA1 Mutation Carriers

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Quy Hoa Nguyen

Dissertation Committee:
Assistant Professor Kai Kessenbrock, Chair
Professor Xing Dai
Assistant Professor Remi Buisson

2021

© 2018 Frontiers
© 2018 Springer Nature
© 2021 Quy Hoa Nguyen

DEDICATION

To my family.
Thank you for everything.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank everyone who has helped and supported me. Without them, I would not be where I am today.

First, I would like to thank my family, friends, and all those who have been there every day for me.

Second, I would like to specifically thank all of my mentors, current, and past. Dr. Kai Kessenbrock accepted me into his lab and provided me with the tools to succeed. Dr. Devon Lawson has always been there to provide additional support and inputs. Dr. Xing Dai has been more than a mentor, providing valuable advice and input for my projects and career. Dr. Olga Razorenova has gone above and beyond to help me. Dr. Madeline Rasche has taught me so much. Dr. Jo Wu and Professor Guy Dadson provided me with an opportunity of a lifetime to start my career. Dr. Remi Buisson for taking time out to serve on my committee and for valuable inputs.

Especially, thank you to everyone who has helped me. There are too many people to list out specifically. Thank you to those who supported my research. Also, thank you to my collaborators for teaching me so much. Thank you to everyone who has made my life so much easier.

CURRICULUM VITAE

Quy Hoa Nguyen

EDUCATION:

Ph.D. in Biomedical Sciences 2021
University of California, Irvine

B.S. in Biochemistry and Molecular Biology 2014
University of California, Irvine

RESEARCH EXPERIENCE:

Graduate Student Researcher Jan. 2016- Jan. 2021
University of California, Irvine
Kai Kessenbrock Lab
Identifying epithelial subpopulations within the breast epithelium to understand the systems-level changes during breast cancer tumorigenesis using single cell genomics.

CMB Graduate Rotation Student Sept. 2015- Dec. 2015
University of California, Irvine
Angela Fleischman Lab
Investigated the role of mutant Calreticulin in the development and progression of myeloproliferative neoplasm.

Lab Assistant I June 2014- Aug. 2015
University of California, Irvine
Olga Razorenova Lab
Managed lab inventory and expenses. Responsible for lab safety and everyday operations. Managed and maintained laboratory equipment. Investigated mechanism of chemical synthetic lethality of various compounds in renal cell carcinoma.

Undergraduate Student Researcher April 2013- June 2014
University of California, Irvine
Undergraduate Research - Olga Razorenova Lab
Investigated chemical synthetic lethality of various compounds in renal cell carcinoma. Managed lab inventory and expenses. Responsible for lab safety and everyday operations. Managed and maintained laboratory equipment.

Research Volunteer Aug. 2012- Sept. 2012
California State University, Fullerton
Madeline Rasche Lab
Investigated ribofuranosylaminobenzene-5`-phosphate synthase enzyme kinetics in conjunction with various inhibitors and developed a new method for purification for crystallization of the enzyme.

HHMI Summer Scholar June 2012- Aug. 2012
California State University, Fullerton
HHMI Summer Research Program - Madeline Rasche Lab
Investigated enzyme kinetics for ribofuranosylaminobenzene-5`-phosphate synthase and its substrate to verify proposed models for its active site.

TEACHING EXPERIENCE:

Cancer System Biology Course Feb. 2019
University of California, Irvine
University of California Irvine CCBS and CRI
Topic: Single-cell analysis of intratumor heterogeneity and its role in drug resistance.

Cancer System Biology Course May 2018
University of California, Irvine
University of California Irvine CCBS and CRI
Topic: Single-cell analysis of intratumor heterogeneity and its role in drug resistance.

PRESENTATIONS:

Speaker July 2019
JLABS, San Diego
Technological advancements in single-cell sequencing Event
Presented: Profiling the breast epithelial population using single-cell RNA sequencing for a comprehensive understanding of cell types and states.

Speaker May 2019
Farmer & The Seahorse, San Diego
2019 10x User Group Meeting
Presented: Multimodal Profiling of the Breast Epithelium with Single Cell RNA and ATAC Sequencing.

Poster Presenter May 2019
University of California, Irvine
UCI Campus-Wide Symposium on Basic Cancer Research
Presented: Profiling Human Breast Epithelial Cells Using Single Cell RNA Sequencing Identifies Cell Diversity.

CCBS Opportunity Award Presenter March 2019
Sheraton Universal, Universal City
UCI Center for Complex Biological Systems Annual Retreat
Presented: Transitional states regulating macrophage heterogeneity in dystrophic muscle.

Poster Presenter June 2018
CZI Initiative West Coast Retreat
Pilot Project for a Human Cell Atlas
Presented: Profiling Human Breast Epithelial Cells Using Single Cell RNA Sequencing Identifies Cell Diversity

CCBS Opportunity Award Presenter April 2017
Westin Pasadena, Pasadena
UCI Center for Complex Biological Systems Annual Retreat
Presented: Understanding the system-level changes during breast cancer tumorigenesis using single cell RNA-seq.

Poster Presenter Jan. 2017
University of California, Irvine
Southern California System Biology Conference
Presented: Identification and Characterization of the Spectrum of Heterogeneity within Human Breast Epithelium by Single Cell RNA Sequencing.

Poster Presenter Nov. 2016
Hilton Long Beach, Long Beach
Cancer Center Annual Scientific Meeting
Presented: Identification and Characterization of the Spectrum of Heterogeneity within Human Breast Epithelium by Single Cell RNA Sequencing.

Student Presenter May 2014
University of California, Irvine
UCI Undergraduate Research Symposium
Presented: Small Molecule Screen in Renal Cell Carcinoma with von Hippel-Lindau Deficiency Reveals Chemical Synthetic Lethal Interactions.

PUBLICATIONS:

Nicholas Pervolarakis, Quy Nguyen, Guadalupe Gutierrez, Peng Sun, Darisha Jhutty, Grace XY Zheng, Corey M Nemec, Xing Dai, Kazuhide Watanabe, and Kai Kessenbrock. **Integrated single-cell transcriptomics and chromatin accessibility analysis reveals novel regulators of mammary epithelial cell identity**. *Cell Reports*. 2020. 33(3):108273.

Bin Fang, Aarthi Kannan, Stephanie Zhao, Quy H. Nguyen, Samuel Ejadi, Maki Yamamoto, J. Camilo Barreto, Haibo Zhao, Ling Gao. **Inhibition of PI3K by cospanlisib exerts potent antitumor effects on Merkel cell carcinoma cell lines and mouse xenografts**. Scientific Reports. 2020. 10(1):8867.

Daniel Haensel, Suoqin Jin, Peng Sun, Rachel Cinco, Morgan Dragan, Quy Nguyen, Zixuan Cang, Yanwen Gong, Adam L MacLean, Kai Kessenbrock, Enrico Gratton, Qing Nie, and Xing Dai. **Defining epidermal basal cell states during skin homeostasis and wound healing using single-cell transcriptomics**. *Cell Reports*. 2020. 30(11):3932-3947.e6.

Hamad Alshetaiwi, Nicholas Pervolarakis, Laura Lynn McIntyre, Dennis Ma, Quy Nguyen, Jan Akara Rath, Kevin Nee, Grace Hernandez, Katrina Evans, Leona Torosian, Anushka Silva, Craig Walsh, Kai Kessenbrock. **Defining the emergence of myeloid-derived suppressor cells in breast cancer using single-cell transcriptomics**. *Science Immunology*. 2020. 5(44):eaay6017.

Stefan A. Brooks, Daniel M. Kim, Sarah J. Morse, Quy H. Nguyen, Brianna M. Craver, Hew Yeng Lai, Angela G. Fleischman. **Upregulation of endogenous thrombopoietin receptor (MPL) with in vivo passage of calreticulin (CALR) mutant Ba/F3 cells, highlighting MPL as the requisite cytokine receptor for CALR mediated transformation**. *Leukemia Research*. 2019. 82:11-14.

Quy H. Nguyen, Nicholas Pervolarakis, Kevin Nee, and Kai Kessenbrock. **Experimental Considerations for Single Cell RNA Sequencing Approaches**. *Front. Cell Dev. Biol.* 2018. 6:108.

Quy H. Nguyen, Nicholas Pervolarakis, Kerrigan Blake, Dennis Ma, Ryan Tevia Davis, Nathan James, Anh T. Phung, Elizabeth Willey, Raj Kumar, Eric Jabart, Ian Driver, Jason Rock, Andrei Goga, Seema A. Khan, Devon A. Lawson, Zena Werb, and Kai Kessenbrock. **Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity**. *Nature Communications*. 2018. 9(1):2028.

Thompson J.M., Alvarez A., Singha M.K., Pavesic M.W., Nguyen Q.H., Nelson L.J., Fruman D.A., Razorenova O.V. **Targeting the mevalonate pathway suppresses VHL-deficient CC-RCC through a HIF-dependent mechanism**. *Molecular Cancer Therapeutics*. 2018. 17(8):1781-1792.

Thompson JM, Nguyen QH, Singh M, Pavesic MW, Nesterenko I, Nelson LJ, Liao AC, Razorenova OV. **Rho-Associated Kinase 1 (ROCK1) inhibition is Synthetically Lethal with Von Hippel Lindau (VHL) Deficiency in Clear Cell Renal Cell Carcinoma (CC-RCC)**. *Oncogene*. 2016. 36(8):1080-1089.

Thompson J.M., Nguyen Q.H., Singh M., Razorenova O.V. **Approaches to Identifying Synthetic Lethal Interactions in Cancer**. *Yale J Biol Med*. 2015. 88(2): 145–155.

AWARDS:

John Wasmuth Graduate Student Research Seminar Award 2019 June 2019
University of California, Irvine

Presented: Identifying changes in the cellular ecosystem during breast cancer initiation.

CCBS Opportunity Award June 2018
University of California, Irvine
Project: Characterization of macrophage heterogeneity in dystrophic muscle.

Dr. Lorna Carlin Award May 2018
University of California, Irvine

CCBS Opportunity Award June 2016
University of California, Irvine
Project: Understanding the systems-level changes during breast tumorigenesis using single cell RNA-seq.

ABSTRACT OF THE DISSERTATION

Defining Cellular Heterogeneity within the Normal Human Breast Epithelial System and BRCA1 Mutation Carriers

By

Quy Hoa Nguyen

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2021

Assistant Professor Kai Kessenbrock, Chair

The breast epithelial system is a complex network of various cell types that work together to provide milk for the offspring. This complex network requires three distinct cell types to maintain the structure and function of the breast. This system is extremely dynamic and constantly changes in response to hormones secreted during different developmental stages. This constant change requires rapid growth and involution of the network and is where mutations and cellular dysregulation could lead to serious complications such as breast cancer. This dissertation focuses on creating a reference atlas of the normal breast epithelium and explores the pre-malignant changes that occur in human BRCA1 mutation carriers. This work utilizes single-cell RNAseq to profile the breast epithelium, which enable us to identify three main cell types, namely two distinct luminal and one basal cell type. In addition to the known basal and luminal populations, an additional luminal population was identified. This highlights the heterogeneity that exists at the cellular level in the luminal population that has been previously overlooked by bulk profiling approaches. Comparative analysis with BRCA1 mutation carrier samples reveals an expansion of a luminal progenitor

population. This population is over-represented by cells from BRCA1 mutation carriers and have higher overall scEnergy, indicating greater dysregulation. Further characterization of this population reveals markers relating to stem and progenitor identities, suggesting its role as the cancer cell of origin. To perform single-cell profiling and functional assays, we needed to establish a robust method for single-cell isolation and purification. Analysis of the current method for single-cell isolation reveals that biological changes arise and these changes have long-term effects on the growth potential, thereby influencing the results of functional assays. Overall this work takes a high-resolution approach to identify cellular heterogeneity in the breast epithelium and elude the changes that occur during early cancer initiation for BRCA1 mutation carriers.

CHAPTER 1: INTRODUCTION

Incorporates components from published article:

Experimental Considerations for Single-Cell RNA Sequencing Approaches

Quy H. Nguyen, Nicholas Pervolarakis, Kevin Nee and Kai Kessenbrock. Frontier Cell and Dev. Biol. 2018.

INTRODUCTION

Human breast

The anatomical structure of the human breast has been well studied and understood for a long time, and it is well known that the role of the breast is to produce and secrete milk to nurse the offspring during pregnancy. Anatomically, the human breast is described as a fatty tissue that surrounds the mammary epithelium and supports their growth and development throughout a person's life (Tobon and Salazar, 1974). The mammary gland is a glandular tissue made up of approximately 15-20 lobes and terminates into many lobular structures; this is where milk is produced in response to hormones secreted during pregnancy (Macias and Hinck, 2012; Pandya and Moore, 2011). A ductal system connects these lobule units to the other lobes and nipple, channel milk to the nipple, and allows for the nursing of the offspring (Hassiotou and Geddes, 2013). This is a very dynamic structure, and many biological processes take place during these changes, and dysregulation in some of these processes can cause catastrophic changes in breast biology and development. A well-known disease that arises due to dysregulation in the breast normal growth biology is breast cancer. The origin of many cases of breast cancer is known to be caused by abnormal growth in the glandular tissues that provide that function of the breast. However, the molecular changes that cause this dysregulation and aberrant growth are still elusive. To understand how genetic and environmental factors can influence these changes, we must first understand how normal breast develops and changes in response to hormones and other environmental factors.

Breast epithelial system

On the cellular level, the human breast consists of both epithelial and stromal components. These stromal components secrete hormones and other factors that regulate homeostasis and promote growth and involution when appropriate. The epithelial system forms a ductal network composed of two layers, the inner layer made up of luminal cells and an outer layer made up of basal cells (Glukhova et al., 1995; Love and Barsky, 2004). This ductal network of epithelial cells grows into a mainly adipose-rich tissue surrounded by other stromal components such as; immune cells, fibroblast, extracellular matrix, and vascular and lymphatic vessels (Margan et al., 2016). This ductal epithelial network grows and terminates into lobular units, also known as terminal duct lobular units (Rønnov-Jessen et al., 1996), where they serve their purpose in the breast as milk-producing units (Ramsay et al., 2005). These terminal duct lobular units form structures like branches on a tree. They are surrounded by a layer of basal cells and have an inner layer of luminal cells that secret milk during lactation into the lumen of these terminal ducts lobular units. These terminal lobular units are connected to each other, and the nipple by ducts made up of the same epithelial bilayer, an outer basal and inner luminal cells. This ductal epithelial network provides the breast's primary function, which is to produce and provide milk to offspring after pregnancy. A recent study has identified three major cell types within the human breast epithelial system: one basal subpopulation and two luminal subpopulations, luminal 1 and luminal 2 (Nguyen et al., 2018a). The luminal 1 subpopulation is marked by SLPI and has signatures similar to that of progenitor luminal cells. In contrast, the luminal 2 subpopulations are marked by ANKRD30A and have signatures resembling mature luminal

cells. A more in-depth analysis on the single-cell level also reveals a distinction in the luminal 1 population, further separating that cell type into two cell state subpopulations.

Breast epithelial development

This glandular tissue is extremely dynamic and continuously changes throughout a person's life, from early development to menopause. The anatomy of the breast can drastically change depending on a person's age, menstrual cycle, and pregnancy status. Hormones play a big role in the regulations of these changes (Ferguson et al., 1992; Schedin et al., 2004). Early development of the breast is critical for establishing a structure in which the breast can form and developed. The normal glands first establish a rudimentary structure and remain dormant until puberty (Sternlicht, 2006). This is when estrogen and other sex hormones drive significant anatomical changes and development of this rudimentary structure into a fully developed glandular tissue network (Joshi et al., 2010).

At the tissue's base are multipotent stem cells and lineage-specific progenitor cells; they provide the foundation for the initial development and continual morphogenesis in response to hormone secretion (Bach et al., 2017). Estrogen is the main driver behind the initial ductal elongation and branching (Howard and Gusterson, 2000). This occurs as the epithelial cells rapidly expand and form into their bilayer structure. Further growth and expansion typically occur at the terminal duct lobular units (Hens and Wysolmerski, 2005). This is where the progenitor cells are believed to reside and drive the changes that occur during the different cycles of breast development (Van Keymeulen et al., 2011). Recent studies allude to the existence of a multipotent stem cell that resides in the basal layer (Petersen and Polyak, 2010; Zhou et al., 2019). These studies have identified a subpopulation

of basal cells with greater stem cell capacity (Visvader and Stingl, 2014; Yang et al., 2017). Basal cells alone can reconstitute the mammary system in mice after a transplant, giving rise to both basal and luminal cells to complete the epithelial ductal network (Eaves et al., 2006; Shackleton et al., 2006). These mammary stem cells are responsible for the growth and expansion of the epithelial network during dynamic periods and constant cell renewal (Fridriksdottir et al., 2005; Woodward, 2005). At the same time, there is also evidence supporting the idea that the epithelial network's maintenance is maintained by lineage-specific progenitor cells that developed earlier on during the initial development of the system (Casey et al., 2018; Davis et al., 2016). Their direct role during tumor progression has not been well characterized. Disruption of homeostasis within these two cell types can give rise to different subtypes of breast cancer.

Interaction between the epithelial cells and their environment also influences their development (Robinson et al., 1999). Stromal cells in the breast reside near the epithelial network and support their growth and homeostasis. These stromal cells help promote adhesion and migration of epithelial cells by secretion of extracellular factors such as growth factors and extracellular matrix (Hu et al., 2017; Jena and Janjanam, 2018). These secreted factors have influenced ducts and branching morphology (Ferguson et al., 1992). Genetic mutations and other environmental factors could lead to the dysregulation in these interactions and normal cellular response to control growth.

Breast cancer

Breast cancer is one of the most prevalent forms of cancer in women worldwide, with an especially poor prognosis for the most aggressive basal-like subtype and late-stage cancer

diagnosis (Anders and Carey, 2008; Foulkes et al., 2010). Patients' survival drops drastically once the disease has progressed to a later stage. Disease progression and survival are worse for patients with mutations or other factors that put them at higher risk of more aggressive breast cancer. Some of these subtypes are characterized by the fact that this cancer does not have a known targetable receptor. With a recent report of over 250,000 new cases in the US and over 40,000 deaths, breast cancer remains at the top of the list, and a complete understanding and cure still elude us. The 5-year survival for women diagnosed with an early stage has increased in recent years and is almost 99%. This is due to advancements in early diagnostic and surgical procedures. Cellular and molecular understanding of cancer progression has also contributed to a favorable outcome for most patients who are diagnosed early.

Breast cancer arises from the epithelium. Dysregulation of cellular identities and improper cellular signaling leads to an unchecked and rapid expansion of certain epithelial cells. This process is influenced by genetic and environmental factors that can disrupt normal homeostasis and promote abnormal growth. Several different subtypes of breast cancer have been characterized. Each subtype is believed to arise from a different cell type and has distinct behavior in their formation and progression (Børresen-Dale et al., 2003; Eroles et al., 2012; Pergamenschikov et al., 2002). Common to all breast cancer is that it arises from the breast epithelium (Dimri et al., 2005; Hinck and Näthke, 2014). Two major cell types that make up the breast epithelium are basal and luminal epithelial cells, and studies have shown that each one can give rise to a distinct breast cancer subtype.

Recent studies indicate that luminal progenitor cells may be responsible for some cases of triple-negative breast cancer (TNBC) (Lim et al., 2009). Cases of the more aggressive

TNBC are characterized by their lack of estrogen, progesterone, and HER2 receptors (Bianchini et al., 2016; Hedenfalk, 2006). This renders the cancer insensitive and resistant to targeted hormone therapies and allows the tumor to become proliferative and highly aggressive (Cancer et al., 2011). A major driver for this subtype of breast cancer in women is a mutation in the BRCA1 gene (Stevens et al., 2013). Approximately 60% of women who carry this pro-tumorigenic mutation will develop breast cancer, and about 70% of those cases will be characterized by the more aggressive basal-like subtype (Kerr and Ashworth, 2001; Orban and Olah, 2003). These pro-tumorigenic mutations promote destabilization of cellular homeostasis and cancer formation. Due to the lack of targeted therapy for TNBC, BRCA1 mutation carriers commonly undergo a prophylactic mastectomy to reduce their risk of developing breast cancer (Kaas et al., 2010). A majority of those who do develop breast cancer do not respond well to the available chemotherapies. Upon completion of treatment, they relapse with a more aggressive cancer, which results in metastasis and death (Harbeck et al., 2009). Fully understanding how BRCA1 mutations disrupt normal cellular homeostasis remains elusive mainly due to technical difficulties in separating each cell population's expression signatures in previous bulk expression analysis approaches. Therefore it is essential to understand the early stages of breast cancer initiation and progression at the single-cell level. This will form the basis for a systems-level understanding of tumor formation and the identification of the breast cancer cells of origin in BRCA1 mutation carriers.

Approaches to studying cellular heterogeneity

Recent advances in single-cell profiling technology enable us to interrogate cellular heterogeneity at a resolution previously impossible with bulk sequencing. High throughput isolation of single cells enables an unbiased analysis of all cell types within the breast epithelium and will allow for reconstruction of the hierarchy of differentiation. It remains a debated question in the field whether these stem cells are bi-potent in that they can differentiate into both luminal or basal lineages or whether there are lineage-restricted unipotent stem cells within this tissue. Our knowledge of the mammary epithelium is mainly based on bulk analysis approaches that average out individual cell differences and only allow us to generalize the same characteristics for all cells in the system. The unbiased identification of stem cells and other epithelial subpopulations using single-cell RNAseq will allow for a better understanding of how epithelial tissue develops and maintain homeostasis on a cellular and molecular level.

Elucidating cellular heterogeneity represents a major scientific challenge in many areas of biology and biomedical research, including developmental and stem cell biology, immunology, neurobiology, and cancer research (Wagner et al., 2016). The recent convergence of next-generation sequencing and bioengineering approaches to manipulate individual cells has led to unbiased single-cell DNA (Navin et al., 2011), RNA (Pollen et al., 2014; Tanay and Regev, 2017), and ATAC sequencing (Buenrostro et al., 2015). These technological advances redefine our understanding of how biological systems function and have formed the basis for large-scale, international collaborations such as the Human Cell Atlas project (Rozenblatt-Rosen et al., 2017a). In this spirit, a recent endeavor using microwell-based single-cell RNAseq created the first cell atlas to map out most mouse tissues (Han et al., 2018). Moreover, single-cell RNAseq has provided critical new insights into key

developmental processes, such as the early cardiovascular lineage segregation steps in mice (Lescroart et al., 2018). Our recent work utilized single-cell RNAseq to reveal the spectrum of cellular heterogeneity within the human breast epithelium identifying three major cell types, each harboring multiple distinct cell states (Nguyen et al., 2018a). Due to the high sensitivity of these methods, in particular single-cell RNAseq, it can be challenging to choose a good approach to minimize batch effects and unwanted technical variation that may overshadow true biological insights.

Protocols for transcriptome analysis have advanced rapidly, resulting in several robust methods that range in cell and mRNA capture strategy, barcoding, throughput, and automation level (Fan et al., 2015; Macosko et al., 2015). The selection of the optimal approach depends largely on the research question. Recent high-throughput protocols for single-cell RNAseq have dramatically increased scalability through automation, increasing the number of cells that can be processed simultaneously, and decreasing reagent cost through reaction miniaturization. Using microwell-based (Cytoseq, Wayfergen), microfluidics-based (Fluidigm C1 HT), or droplet-based (inDrop, Drop-seq, and 10× Chromium) approaches, hundreds to thousands of cells can be captured in a single experiment (Heath et al., 2016; Islam et al., 2014; Klein et al., 2015; Picelli et al., 2014; Zheng et al., 2017). The newest of these protocols utilize beads functionalized with oligonucleotide primers, each containing a universal PCR priming site, a cell-specific barcode, an mRNA capture sequence, and Unique Molecular Identifiers. Individual cells are captured in wells or droplets with a single bead. Cell-specific barcodes are similar within a droplet, but a unique UMI sequence on the primer allows individual transcripts within a cell to be counted. This provides a quantitative readout of each gene's number of transcripts detected in a cell,

thereby reducing the effects of amplification duplicates that occur with earlier technologies (Patel et al., 2014; Ramsköld et al., 2012). High-throughput 3′-end counting approaches have several important limitations. Since only the 3′-end of each mRNA are sequenced, differential splicing analyses are not feasible (Heath et al., 2016; Macosko et al., 2015). High-throughput approaches typically only achieve ~10% transcriptome coverage, relative to ~40% for full-length sing cell RNAseq protocols that use Switching Mechanism at 5′End of RNA Template (SMART) chemistry (Tirosh et al., 2016; Yuan et al., 2017). This is partly due to lower mRNA capture efficiency but also due to lower sequencing depth. Single-cell qPCR platforms (e.g., Fluidigm C1 and Biomark) remain superior in detecting low-expressed genes (Lawson et al., 2015). Protocols for processing rare cells usually involve an upstream capture step by flow cytometry or micromanipulation, followed by dispensing single cells into microtubes or microwell plates. Studies investigating rare cell populations that require selection via specific markers (e.g., adult tissue stem cell populations) are best performed using these protocols. Single-cell libraries are prepared using SMART-based chemistry, which utilizes a template-switching oligonucleotide (TSO) (Tirosh et al., 2016). This TSO can be used to prime off of the untemplated nucleotides added by the reverse transcriptase, enabling subsequent PCR using a single primer and capture of full-length transcripts (Tirosh et al., 2016; Yuan et al., 2017). PCR then amplifies cDNAs, and libraries are prepared for sequencing using standard protocols. Although several large-scale projects utilize these protocols, because they are manual and utilize larger microliter reaction volumes, they limit the number of cells that can be processed at a reasonable cost.

Sample preparation for single-cell RNAseq

The process of single-cell preparation is arguably the greatest source of unwanted technical variation and batch effects in any single-cell study (Tung et al., 2017). Different tissues can vary significantly in extracellular matrix (ECM) composition, cellularity, and stiffness, and therefore dissociation protocols must be optimized for the specific tissue type of interest. Conventional protocols for single-cell preparation typically involve the following steps: (1) tissue dissection, (2) mechanical mincing, (3) enzymatic/proteolytic ECM breakdown (e.g., dispase, collagenase, trypsin) often accompanied by mechanical agitation, and (4) optional enrichment for cell types of interest by flow cytometry, bead-based immune-selection, differential centrifugation, or sedimentation. Each step can affect the cells' expression signatures and should be carefully optimized to introduce the least artifact. An optimal tissue dissociation protocol will yield as many viable cells as possible in the shortest possible duration without preferentially depleting or significantly altering the frequencies of certain cell types. Recent advances in bioengineering of innovative microfluidic cell dissociation devices (Qiu et al., 2014) can radically change the way tissue samples are dissociated into single cells while avoiding inter-assay variation due to human handling of the tissue. Several microfluidic devices have been optimized for streamlined tissue digestion, cell dissociation, filtering, and polishing. In brief, these devices were designed to work with tissue sequentially through progressively smaller size scales, starting from a tissue specimen, through cellular aggregates and clusters, and finally eluting a solution containing close to 100% single cells, which will be ideal for single-cell RNAseq applications. Also, new semi-automated commercially available systems can help streamline tissue dissociation (e.g., Miltenyi gentleMACS). These devices offer tissue-type specific kits that may allow more reproducible, time-saving, efficient tissue dissociation, and single-cell

preparation (Baldan et al., 2015; Meeson et al., 2013). Ultimately, determining a "best practices" dissociation strategy through heuristic optimization will be critical for downstream single-cell library quality.

There are various methods for isolating specific cell populations or removing unwanted populations that should be optimized for any specific tissue type. Manual isolation utilizing magnetic beads or gradient purification are potential methods for removing unwanted cells such as dead cells. Flow cytometry is a widely used, high-throughput method to enrich for rare cells such as hematopoietic stem cells (Radbruch and Recktenwald, 1995; Will and Steidl, 2010). However, these methods are not without drawbacks since they can introduce artificial stress on cells and change their expression profile(Van Den Brink et al., 2017). Methods that involve antibody binding for purification can also affect the cell expression profile if binding of the antibodies to cell surface molecules induce intracellular signaling (Christaki et al., 2011; Kornbluth and Hoover, 1989). Flow cytometry-isolated cells are exposed to the high pressure during sorting, and these osmotic and pressure changes introduced to cells during cell sorting and handling can induce change to the cell expression profile of multiple cell types (Van Den Brink et al., 2017; Romero-santacreu et al., 2009; Xiong et al., 2002).

Due to the high cost of single-cell sequencing experiments, careful quality control measurements should be executed. The performance of alternative protocols can be assessed using several readouts. A useful first metric can be acquired using imaging of viability, such as using the Countess platform. Flow cytometry is particularly valuable to measure several critical metrics simultaneously, such as cell viability and contamination with doublets and small cell clusters, which can confound single-cell sequencing results.

Flow cytometry can also be used to evaluate whether cell populations of interest, such as immune cells, stromal fibroblasts, or stem cell populations, are maintained in the cell preparation and the appropriate frequency. Finally, an additional metric on RNA quality can be acquired using the RNA integrity number (RIN) method (Schroeder et al., 2006).

Thesis work

Taken all together, breast cancer is still a scientific and public health dilemma. The breast epithelial system is complex and dynamic; changes in cellular processes can lead to disruption in homeostasis and tumorigenesis. With the most new cases and high mortality rate for late stage, we are still a long way from fully understanding how this disease develops and progresses. Previous work has made great strides to elucidate how the breast develops on the cellular level and to characterize many breast cancer subtypes. This allows for better screening and treatment for some cases; it is especially beneficial to those diagnosed with less aggressive subtypes or those who are diagnosed early. Despite advances in screening and treatments for breast cancer, targeted therapies for TNBC are difficult to come by and develop. This is due to the lack of targetable receptors that other subtypes of breast cancer express. Women who have a mutation in the BRCA1 gene have a higher probability of developing this more aggressive subtype of breast cancer, so they typically choose to undergo prophylactic mastectomy. To improve screening methods and help individuals with the BRCA1 mutation, new biomarkers are needed to categorize patients based on their actual risk of developing breast cancer. While some studies focus on treating the disease after tumor formation, this research is innovative because it seeks to characterize the mechanism involved in the early disruption of cellular homeostasis before tumor formation begins. My

application of high throughput single-cell profiling technology will enable unbiased surveying of all cell types and map their roles in maintaining the normal breast epithelium. This will provide the basis for understanding changes in cellular homeostasis during early tumor development and could lead to better diagnosis and prevention therapies. Due to the high cases of disease reoccurrence after treatment and aggressive behavior of metastatic breast cancer, prevention of this disease is preferred over treatment.

CHAPTER 2: Profiling human breast epithelial cells using single-cell RNA sequencing identifies cell diversity

Incorporates components from published article:

Profiling human breast epithelial cells using single-cell RNA sequencing identifies cell diversity

Quy H. Nguyen, Nicholas Pervolarakis, Kerrigan Blake, Dennis Ma, Ryan Tevia Davis, Nathan James, Anh T. Phung, Elizabeth Willey, Raj Kumar, Eric Jabart, Ian Driver, Jason Rock, Andrei Goga, Seema A. Khan, Devon A. Lawson, Zena Werb & Kai Kessenbrock. Nature Communications 2018.

ABSTRACT

Breast cancer arises from breast epithelial cells that acquire genetic alterations leading to the subsequent loss of tissue homeostasis. Several distinct epithelial subpopulations have been proposed, but a complete understanding of the spectrum of heterogeneity and differentiation hierarchy in the human breast remains elusive. Here, we use single-cell mRNA sequencing (scRNAseq) to profile the transcriptomes of 25,790 primary human breast epithelial cells isolated from reduction mammoplasties of seven individuals. Unbiased clustering analysis reveals the existence of three distinct epithelial cell populations, one basal and two luminal cell types, which we identify as secretory L1- and hormone-responsive L2-type cells. Pseudotemporal reconstruction of differentiation trajectories produces one continuous lineage hierarchy that closely connects the basal lineage to the two differentiated luminal branches. Our comprehensive cell atlas provides insights into the human breast epithelium's cellular blueprint and will form the foundation to understand how the system goes awry during breast cancer.

INTRODUCTION

Breast cancer is a highly heterogeneous disease subtyped based on tissue morphology and molecular signatures (Perou et al., 2000). At least six different intrinsic subtypes of breast cancers have been established, namely luminal A, luminal B, HER2-enriched, basal-like, normal breast, claudin-low (Eroles et al., 2012), and more recently, up to ten subtypes have been described (Ali, 2014). Each subtype is speculated to arise from a different cell of origin (Visvader and Stingl, 2014); however, gaps in our understanding of the full spectrum of cellular heterogeneity and the distinct cell types that comprise the

human breast epithelium hinder our ability to investigate their roles in cancer initiation and progression.

Breast cancer arises from the breast epithelium, which forms a ductal network embedded into an adipose tissue that connects the nipple through collecting ducts to an intricate system of 12–20 lobes, which are the milk-producing structures during pregnancy and lactation. The breast epithelium is composed of two known cell types throughout the duct and lobular system, an inner layer of secretory luminal cells and an outer layer of basal/myoepithelial cells. A series of recent reports have indicated that further heterogeneity exists within these two cell layers in mice. Two landmark papers published in 2006 identified a functionally distinct subpopulation of basal epithelial cells that harbors stem cell capacity and can reconstitute a fully developed mammary epithelial network when transplanted into the cleared mammary fat pads of mice (Eaves et al., 2006; Shackleton et al., 2006). Moreover, a subpopulation of luminal progenitor cells identified by high expression of KIT and a subpopulation of mature luminal cells have been identified using flow cytometry (FACS) isolation strategies (Shehata et al., 2012; Stingl et al., 2001). Interestingly, based on comparative bulk expression analyses, these luminal progenitors may have an increased propensity to give rise to triple-negative breast cancers in patients with mutations in the BRCA1 gene (Lim et al., 2009). It remains to be determined if other distinct cell types exist within the breast epithelium and how these relate to the known subtypes of breast cancer.

Advances in next-generation sequencing and microfluidic-based handling of cells and reagents now enable us to explore cellular heterogeneity on a single cell level and reconstruct lineage hierarchies using single-cell mRNA sequencing (scRNAseq) (Pollen et al., 2014; Treutlein et al., 2014). This approach allows an unbiased analysis of the spectrum of

heterogeneity within a population of cells since it utilizes transcriptome reconstruction from individual cells. scRNAseq has been successfully applied to understand the complex subpopulations in normal tissues such as lung (Treutlein et al., 2014) or brain (Pollen et al., 2014) as well as in various cancers including melanoma (Tirosh et al., 2016), glioblastoma (Patel et al., 2014), and within circulating tumor cells from patients with pancreatic cancer (Ting, 2014).

The present study aims to generate a molecular census of cell types and states within the human breast epithelium using unbiased scRNAseq. Focusing on the breast epithelium, our work provides a critical first impetus toward generating large-scale single-cell atlases of the tissues comprising the human body as part of the international human cell atlas initiative (Rozenblatt-Rosen et al., 2017b). This molecular census can shed light on lineage relationships and differentiation trajectories in the human system and its relation to breast cancer. Our single-cell transcriptome analysis provides unprecedented insights into the spectrum of cellular heterogeneity within the human breast epithelium under normal homeostasis. It will serve as a valuable resource to understand how the system changes during early tumorigenesis and tumor progression.

RESULTS

scRNAseq reveals three cell types in the breast epithelium

We collected a cohort of reduction mammoplasties from age- and ethnicity-matched, post-pubertal and pre-menopausal females and performed scRNAseq on purified breast epithelial cells, which were isolated from surrounding stromal cells using flow-cytometry

based on differential expression of CD49f and EpCAM (Lawson et al., 2015). Basal and luminal cells were separately loaded onto the Fluidigm C1 microfluidics-enabled scRNAseq platform (Figure 2.1a). Capture efficiency was monitored by microscopic imaging to exclude doublets and debris from further analysis (Figure 2.2a, b). We used 13 C1 chips in total to capture and sequence transcriptomes of 868 cells from three human individuals. The resulting single-cell cDNA libraries were sequenced in parallel at an average read depth of 1.6 M reads per cell. After removing cells with less than 900 genes detected and additional quality control filtering (see Methods section), we proceeded to analyze 703 single cells at ~4500 genes detected on average per cell, where the gene detection range was comparable between basal and luminal cells (Figure 2.2c).

To identify the main cell types within the breast epithelium that are generalizable across individuals, we performed a combined analysis of all cells from the three individuals using the recently described Seurat pipeline (Macosko et al., 2015). This analysis identified three very distinct clusters of cells (Figure 2.1b), indicating that the breast epithelium is composed of three main cell types. We then explored the significantly up-regulated genes within each cluster (Figure 2.1c), which revealed that these main clusters correspond to one major basal (KRT14+; AUC = 0.83) cell type and two luminal cell types that both express the typical markers KRT8 and KRT18. Importantly, cells representing all three cell types were detected in each of the three individuals (Figure 2.2d). We found several distinct markers for these luminal cell types, such as SLPI (AUC = 0.89) for L1 and ANKRD30A (AUC = 0.81) for L2 (Figure 2.2e). Comparing these signatures to previously published microarray expression analyses of FACS-isolated human breast epithelial cells (Lim, 2010; Lim et al., 2009), we found that L1 corresponds closely to the CD49f+/EpCAM+ population designated as “luminal

progenitors," L2 resembles the CD49f−/EpCAM+ population called "mature luminal," and the basal cluster matched with CD49fhi/EpCAM− "Basal/MaSC." Since basal cells contain a subset of mammary stem cells (MaSCs) (Eaves et al., 2006; Shackleton et al., 2006; Wang, 2015), we examined the basal cell cluster in more detail. Particularly intriguing was observing a subset with increased expression of mesenchymal and stem cell markers ZEB1 (Morel, 2017) and TCF4 (Figure 2.1d). Interestingly, previous work established a direct link between mesenchymal gene expression signatures and MaSC capacity (Ye, 2015), suggesting these ZEB1/TCF4-expressing cells may represent a subset of basal cells with increased MaSC potential.

Droplet-mediated scRNAseq reveals subpopulation diversity

To determine whether additional cellular diversity exists, we next utilized a more scalable droplet-mediated scRNAseq platform (10× Genomics Chromium) (Zheng et al., 2017). Here, we focused on reduction mammoplasty samples from nulliparous women to reduce variability associated with pregnancy-related changes of the breast. We isolated both luminal and basal cells together (EpCAM+/CD49fhi/lo) by flow cytometry and subjected them as one sample to droplet-based scRNAseq targeting on average 5000 cells per sample (Figure 2.3a). We sequenced a total of 24,646 cells from four individuals (Ind4-7) at an average of ~60,000 reads per cell. After quality control filtering to remove cells with low gene detection (<500 genes) and high mitochondrial gene coverage (>10%), detailed clustering analysis of the first individual (Ind4) using Seurat confirmed the existence of three main epithelial cell types, namely Basal (KRT14+), Luminal1 (L1; KRT18+/SLPI+) and Luminal2 (L2; KRT18+/ANKRD30A+) (Figure 2.3b). These analyses also revealed three

additional small clusters; cluster 8 was defined by stromal marker VIM ($P < 9.6 \times 10{-}25$); cluster 9 showed specific expression of endothelial marker gene ESAM ($P < 4.1 \times 10{-}30$); and cluster 10 included a small number of dispersed cells most likely representing outliers. We concluded that these clusters (8–10) were of non-epithelial and denoted them as unclassified (X) in further analyses.

Interestingly, multiple subclusters emerged within each of the main epithelial cell types as indicated by their distinct marker gene signatures (Figure 2.3c). We hypothesized that the main islands of cells (Basal, L1, L2) represent distinct "cell types", whereas subclusters within each island depict "cell states" that are more transient over time (Trapnell, 2015). Within basal cells, we detected three distinct cell states, which showed specific expression of inflammatory mediators (IL24; $P < 1.4 \times 10{-}180$; Cluster 3), markers for myoepithelial cell function (ACTA2; $P < 7.4 \times 10{-}292$; Cluster 4) and specific epithelial keratin expression (KRT17; $P < 1.6 \times 10{-}38$; Cluster 5), respectively. ZEB1 and TCF4, which marked a subset of basal cells in our microfluidics-enabled scRNAseq analysis (Figure 2.1d), were lowly detected and therefore not interpretable droplet-enabled scRNAseq, which is likely due to lower coverage compared to the microfluidics-enabled platform (Svensson et al., 2017).

Within luminal cell type L1, we observed three distinct cell states that were marked by genes associated with milk production (LTF; $P < 8.4 \times 10{-}270$; Cluster 1), high expression of secretory molecules (SAA2; $P < 2.2 \times 10{-}90$; Cluster 0) and distinct epithelial keratin expression (KRT23; $P < 2.5 \times 10{-}157$; Cluster 2). The second luminal cell type L2 harbored two distinct cell states that were marked by expression of hormone-responsive genes (AGR2; $P < 3.1 \times 10{-}144$; Cluster 6) and specific cell surface markers (CD74; $P < 2.9 \times 10{-}121$;

Cluster 7). We next performed detailed individual Seurat clustering analyses for three additional individual datasets from nulliparous women, which confirmed many of the patterns described for Ind4 (Figure 2.3). Like Ind4, the other individuals possessed three main cell clusters corresponding to cell types Basal, L1, and L2, and eight to ten subclusters (Figure 2.4a–c). The number of subclusters per cell type varied across the individuals with Ind5 comprising five Basal, three L1 and one L2 cluster, Ind6 containing seven Basal, three L1 and one L2 cluster, and Ind7 comprising one Basal, three L1 and five L2 clusters (Figure 2.4a–c), which may be due to individual-to-individual variation or anatomical location of the surgical specimens.

To determine cell states that are generalizable across individuals, we developed a comparative approach using a cell scoring method adapted from recently published work (Tirosh et al., 2016). Using the marker gene signatures for each of the 11 clusters (0–10) detected in Ind4 (Figure 2.3b, c), we performed pairwise gene scoring analyses to find matches for every distinct cluster identified in Ind5–7 (Figure 2.4a–c). Comparing Ind4 to Ind5–7 showed that the main cell types (Basal, L1, L2) readily match up across all individuals (Figure 2.5a–c). In addition, it revealed that there are two distinct cell states present within L1 (L1.1 and L1.2) that emerge in all four individuals. The L2 population, which contained two clusters in Ind4, was more homogeneous, and therefore these clusters were combined to a single L2 population. Comparing basal subclusters between individuals suggested that at least two generalizable cell states within basal cells (Figure 2.5a–c). We then performed a separate Seurat analysis using combined basal cells from all four individuals (Figure 2.6a). Several clusters displayed consistently high expression of genes associated with myoepithelial cell function (e.g., ACTA2, TGLN, KRT14). We, therefore, generated a

"myoepithelial cell signature" gene list based on published work (Gudjonsson et al., 2005) to stratify basal cells into either a "Basal" or "Myoepithelial" grouping (Figure 2.6b, c). These results allowed us to include all individual-specific clusters into the final cluster designations, namely Basal (B), Myoepithelial (Myo), Luminal1.1 (L1.1), Luminal1.2 (L1.2), Luminal2 (L2), and the small Unclassified (X) as summarized in Figure 2.6c. These designations were used to perform a combined Seurat analysis of all 24,465 cells from four individuals (Figure 2.5d), which enabled us to determine the common marker genes (e.g., B: APOD; Myo: TAGLN; L1.1: LTF; L1.2: CLDN4; L2: AGR2) for each cell state that is generalizable across all four individuals (Figure 2.5e).

To learn more about the biology underlying these cell states, we used Ingenuity Pathway Analysis (IPA) to identify distinct signaling pathways (Figure 2.6d) and interrogated for transcription factor consensus sites using the Enrichr tool (Kuleshov et al., 2016). These analyses revealed that the Myo state might be controlled by the transcription factors TP63 and PPARγ. It is defined by increased integrin and paxillin signaling, indicating that these cells provide physical integrity within the breast epithelial architecture. The B state was found to be linked to transcription factors STAT3 as well as SOX2, NANOG, and KLF4, which are associated with stem cell capacity and cellular plasticity (Filipczyk, 2015), suggesting that population B may harbor MaSCs. Within the luminal compartment, L1.1 showed distinct signatures of iNOS and IL6 signaling that may indicate a sentinel function of tissue harm and inflammation associated with this cell state. L1.2 displayed increased PI3K/AKT levels and glucocorticoid signaling, which may indicate a link to steroid hormone signaling for this cell population. Within the second luminal cell type L2, we found evidence

for elevated mTOR signaling and aldosterone signaling in epithelial cells, which suggests that this cell type represents a hormone-responsive cell population.

Spatial integration of cell types and states

We next used indirect immunofluorescence analysis to validate our scRNAseq findings on the protein level and to spatially integrate newly discovered cell types and states into the anatomy of the breast. We first focused on the cell states detected within the basal compartment. Immunostaining for ZEB1, which we identified in a subset of basal cells in microfluidics-enabled scRNAseq (Figure 2.1d), showed that this protein is indeed expressed in a small fraction of basal epithelial cells (Figure 2.7a). High ZEB1 and medium KRT14 levels have been recently described in a population of protein C receptor (ProCR) expressing murine MaSCs with in vitro and in vivo stem cell activity (Wang, 2015). Comparison of published gene expression signatures of ProtCR+ MaSCs with the ZEB1+ population identified here showed striking similarity (Figure 2.7b), suggesting that the ZEB1+ basal cells may represent a population of human MaSCs. In addition, staining for TCF4 revealed a similar staining pattern to ZEB1 within the basal (smooth muscle actin-positive) compartment (Figure 2.7c). These findings show that the cell state characterized by ZEB1 and TCF4 expression exists within the intact breast tissue's basal compartment.

KRT14 expression is a hallmark for basal cells, and our differential gene expression analysis confirmed that KRT14 is predominantly expressed within basal cells. However, it exhibited surprising variability across all basal cell populations with particularly high expression in the Myo cell state (Figure 2.7d). Immunofluorescence analysis for KRT14 confirmed this and revealed that KRT14 high cells localized to the basal cell layer within

ductal regions, while lobular basal cells generally displayed lower and more variable staining for KRT14 (Figure 2.7e). Myo cells also expressed high levels of the definitive myoepithelial marker ACTA2 (Figure 2.8a), as well as other genes associated with smooth muscle differentiation and function in other tissues such as MYLK, MYL9, and TAGLN/Transgelin (Robin, 2013).

Surprisingly, basal and luminal markers were not always exclusive, and we noted a distinct fraction of cells that co-express luminal- (e.g., KRT8) and basal- (e.g., KRT14) specific genes, as shown by correlation analysis of our single-cell expression data (Figure 2.8b). To determine whether this population exists in the intact tissue, we performed in situ co-localization analysis by immunofluorescence staining for KRT8 and KRT14. While most areas within the human breast epithelium showed the expected luminal KRT8+/KRT14− or basal KRT8−/KRT14+ pattern, we observed several rare loci within lobular regions of the tissue that indeed showed distinct KRT8+/KRT14+ patterns (Figure 2.8c). Although this cell state has been previously observed in mouse fetal MaSCs (Spike, 2012), our work revealed that this state exists in the human tissue in adult homeostasis.

The scRNAseq analyses revealed that the luminal compartment harbors two discrete epithelial cell types (L1, L2). To determine if L1 and L2 correspond to a ductal and lobular anatomical location within the tissue, we used specific markers for L1 (SLPI) and L2 (ANKRD30A) to identify their spatial distribution within the breast tissue using in situ immunofluorescence. These analyses showed that both L1 and L2 are located next to each other within both ducts and lobules (Figure 2.9a). We next sought to determine their hormone signaling status. Annotation of the single-cell datasets shows that L2 is particularly enriched for ESR1, PGR, and AR (Figure 2.10a), although generally, these genes were found

to be lowly expressed. Consistent with this observation, we also found on the protein-level that L2 marker ANKRD30A commonly overlaps with ER (32.4% of cells), PR (38.0%), and AR (46.8%), whereas SLPI-positive cells showed a markedly lower percentage of hormone receptor expression (Figure 2.9b–d). PGR was also expressed in a sub-fraction of basal cell states, although PR was not detected in basal cells on the protein level (Figure 2.9c).

Proliferation is associated with active progenitor cell capacity within adult epithelial tissues (Fuchs and Nowak, 2008). Interestingly, we observed proliferative cells in all three epithelial cell types (Basal, L1, L2) as evident on the RNA level (Figure 2.10b) and using Ki67 immunostaining (Figure 2.9e–f). Moreover, expression of CDKN1B (p27), which has been previously linked with a quiescent, hormone-responsive progenitor cell population (Choudhury, 2013), was found highest in L2 (Figure 2.10b), while markers for alveolar luminal progenitor cell function such as ELF5 (Harris, 2006) and KIT (Shehata et al., 2012; Stingl et al., 2001) were specifically enriched in luminal subpopulation L1.1 (Figure 2.10c).

L2 was also characterized by higher levels of KRT8 than L1 (Figure 2.9g). To quantify protein expression in individual cells, we utilized a recently developed single-cell western blot application (ProteinSimple, Milo), which performs electrophoretic separation of the protein content of about 2000 cells per chip and subsequently probed with fluorescently labeled antibodies. Applying single-cell western blotting to luminal and basal cells isolated by FACS identified three cell states, namely KRT8-negative, -low, and -high (Figure 2.9h–i), which illustrates the usefulness of single-cell Western blotting as a quantitative validation tool downstream of scRNAseq analyses.

Taken together, these analyses confirmed remarkable concordance between the patterns observed in scRNAseq and on the protein-level in intact tissues. Our spatial analyses

confirmed that the luminal compartment contains two distinct cell types (L1 and L2) intermingle within ducts and lobules. Both contain a subset of proliferative cells, suggesting that they each contain L1- and L2-committed progenitor cells to maintain these cell types. L1 may be committed to secretory function based on their expression signatures, while L2 likely functions as a hormone-sensing unit of the breast epithelium.

Reconstructing lineage hierarchies within the epithelium

To understand how these observed cell types and states are related to each other, we next reconstructed differentiation trajectories by pseudotemporal ordering of single cells using Monocle, which utilizes reverse graph embedding to generate a trajectory plot that can account for both branched and linear differentiation processes (Trapnell, 2014). Applying Monocle to our droplet-based scRNAseq dataset on a subsampled population (4000 cells; 1000 cells per individual) from all four individuals yielded one tightly connected differentiation trajectory that separates into three main branches corresponding to the primary cell types Basal, L1, and L2 (Figure 2.11a). This suggests that the system is maintained through one continuous rather than several disconnected lineages. Considering the substantial evidence supporting the existence of MaSCs within the basal cell compartment (Eaves et al., 2006; Shackleton et al., 2006), we manually set the start of pseudotime within the basal cell type (Figure 2.11b), thus resulting in a trajectory that differentiates into three main branches that are each enriched for Myo, L1, and L2, respectively. Of note, L1.2 is markedly enriched at the branching point between L1 and L2, suggesting that it represents a luminal-restricted bi-potent progenitor. It also precedes L1.1 on the L1 branch, suggesting that L1.2 is a progenitor to L1.1. Interestingly, L1.1 displayed

high ELF5 and KIT expression, previously reported as progenitor cell markers (Shehata et al., 2012; Stingl et al., 2001). Instead, our data suggest that L1.1 represents a second mature, differentiated luminal cell type rather than a luminal progenitor that is upstream of L2. These basic principles were also reflected in our pseudotemporal analysis of the microfluidics-enabled scRNAseq dataset, which projects a bifurcation into luminal and basal lineage emerging from one common population of ZEB1 + progenitor cells (Figure 2.12a b). These results are in line with previous mammary differentiation models mediated by bi-potent stem/progenitor cells (Visvader and Stingl, 2014).

Subpopulations correspond to breast cancer subtypes

To learn more about the relationship of these newly defined subpopulations to existing subtypes of breast cancer, we used our gene scoring approach to directly compare each population's gene signatures to gene signatures associated with each cancer subtype from the Metabric dataset (Trapnell, 2014). This showed that both Luminal A and Luminal B subtypes of breast cancer are closely related to L2-type luminal cells (Figure 2.12c, top), which is in line with previous gene signature analyses of FACS-enriched basal, luminal progenitor, and mature luminal cells (Lim et al., 2009). In addition, a recent report by Lehman et al. used global gene expression analyses to identify molecularly distinct subtypes within triple-negative breast cancer (TNBC) (Lehmann, 2011). We found that Myo showed the highest similarity to the mesenchymal-like subtype of TNBC, while the Basal1 class of TNBC yielded the highest scores in the luminal L1.1 state (Figure 2.12c, bottom). Taken together, these analyses allow us to directly link several defined breast cancer subtypes to

distinct cell populations of epithelial cells suggesting that the subtypes of breast cancer may arise from different tumor cells-of-origin.

DISCUSSION

The current state of knowledge in breast epithelial biology is primarily based on population-level analyses of separated basal and luminal cells following bulk analyses of these distinct epithelial cell types (Shehata et al., 2012). While several distinct subpopulations of murine basal and luminal cells have been reported anecdotally (Visvader and Stingl, 2014), comprehensive knowledge about expression signatures and cellular identities of these subpopulations remains sparse, particularly in the human system. Our scRNAseq analysis of the human breast epithelium from non-diseased, post-puberty, pre-menopausal individuals for the first time allows for unbiased, de novo identification of distinct cell types and states in the adult human breast epithelium before pregnancy-induced changes occur. Strikingly, our approach revealed the existence of three main epithelial cell types (Basal, L1, and L2), in line with a recent scRNAseq analysis of the mouse mammary gland (Pal et al., 2017), although this work referred to these populations as “basal”, “luminal progenitor” and “mature luminal cells”. Our spatial analyses showed that these three cell types intermingle within ducts and lobules and appear to form functionally distinct lineages that contribute to different aspects of breast biology (summarized in Figure 2.13a). three cell types contained a fraction of proliferative cells suggests that cycling, lineage-restricted progenitor cell subpopulations may maintain each cell type during normal homeostasis.

Our unbiased clustering analysis and pseudotemporal reconstruction of differentiation trajectories strongly suggest that these cell types represent three main

branches of specified, differentiated cells: basal/myoepithelial, secretory L1, hormone-responsive L2 cells (Figure 2.13b). The lineage hierarchy likely starts with basal MaSCs (Eaves et al., 2006; Shackleton et al., 2006) that differentiate either into specified myoepithelial cells or into a common luminal progenitor, which gives rise to the two distinct luminal cell types L1 and L2. Interestingly, the ELF5/KIT-expressing subpopulation L1.1 represents a mature differentiated luminal cell state as it was predominantly located at the end of the L1 branch, suggesting that ELF5/KIT may be crucial for differentiation into the secretory L1 cell type rather than promoting progenitor cell function as previously described (Shehata et al., 2012; Stingl et al., 2001). It appears to be the L1.2 cell state within the L1 cell type that harbors a luminal-restricted bi-potent progenitor capacity for differentiation into the more specified secretory L1.1 or hormone-responsive L2 cells.

A currently unresolved question of active debate is whether MaSCs act as bi-potent stem cells that give rise to both lineages of basal and luminal cells37, or whether homeostasis is mediated through distinct uni-potent, lineage-restricted basal and luminal stem cells (Van Keymeulen et al., 2011). Considering these two models, Monocle could have yielded a sparsely connected differentiation trajectory separating basal and luminal lineages, which would have supported a trajectory driven by lineage-restricted basal and luminal unipotent progenitor cells on both ends of the spectrum. Instead, the outcome of our Monocle analysis is in favor of the existence of the bi-potent stem/progenitor model as it identified one continuous trajectory indicative of a common source for both basal and luminal cell differentiation.

Understanding the origins of breast cancer in its earliest phases has the potential to advance methods of cancer early detection and may ultimately form the basis to prevent

cancer progression before it turns into a life-threatening disease. Here, we asked whether the newly identified cell states correspond to specific subtypes of breast cancer and may represent potential cancer cells-of-origin for the specific breast cancer subtypes. The luminal epithelial cell type L2 showed the clearest correlation with both Luminal A and B subtypes from the Metabric dataset (Curtis, 2012), which is in line with previously reported similarities between a FACS-enriched population of mature luminal cells and the luminal-like breast cancer subtypes (Lim et al., 2009). The fact that several L2 markers are independently known as breast cancer-associated antigens such as SYTL2 and ANKRD30A (Seil, 2007), and that it shows the highest expression of CDKN1B/p27 as a marker for potential breast cancer cells of origin(Choudhury, 2013) further corroborates the link between the hormone-responsive L2 cell type to breast cancer in general. Interestingly, the cell state closest related to the TNBC Basal subtype was the luminal progenitor-like population L1.1. The concept that a luminal cell may be the cell-of-origin for basal-type breast cancer is not new and has been previously proposed in the context of BRCA1-driven disease (Lim et al., 2009). Interestingly, those cell states containing subsets of proliferative cells, namely B, L1.1, and L2, are predominantly linked to breast cancer subtypes, which is in line with previous reports showing an association of mammary epithelial cell proliferation in normal tissues with increased breast cancer risk (Huh, 2016).

In summary, our results provide crucial insights into the spectrum of cellular heterogeneity within the human breast epithelium in unprecedented resolution. Our unbiased analysis of the single-cell gene signatures from seven human individuals provides evidence for defined differentiation trajectories to maintain homeostasis in the adult human breast and distinct subpopulations of both basal and luminal lineage that may serve as cells

of origin for the different subtypes of breast cancer. Our single-cell atlas comprising the human breast epithelium will serve as a resource to map out the defined changes occurring during breast cancer and therefore form the basis for improved methods of cancer early detection and possible strategies for cancer prevention.

METHODS

Origin of tissue samples

Anonymous reduction mammoplasty samples were acquired from NCI Cooperative Human Tissue Network (CHTN) and from the Department of Surgery, Feinberg School of Medicine, Northwestern University. Other investigators may have received specimens from the same tissue specimens obtained through NCI CHTN. Specimens were anonymized then collected and distributed by CHTN, specimens are covered under collection/distribution of tissues under consent or waiver of consent. Samples were washed in PBS (Corning 21-031-CV) and mechanically dissociated using a razor blade. Dissociated samples were digested overnight in DMEM (Corning 10-013-CV) with Collagenase Type I, 2 mg/mL (Life Technologies 17100-017). Viable organoids were separated using differential centrifugation and viably frozen in 50% FBS (Omega Scientific FB-12), 40% DMEM, and 10% DMSO (Sigma-Aldrich D8418) by volume.

Single-cell RNA sequencing

Viable organoids were thawed and washed using DMEM, and digested with 0.05% trypsin (Corning 25-052-CI) containing DNase (Sigma Aldrich D4263-5VL) to generate a

single-cell suspension. Cells were stained for FACS using fluorescently labeled antibodies for CD31 (eBiosciences 48-0319-42), CD45 (eBiosciences 48-9459-42), EpCAM (eBiosciences 50-9326-42), CD49f (eBiosciences 12-0495-82), and SytoxBlue (Life Technologies S34857). We only proceeded with samples showing at least 80% viability as measured using SytoxBlue in FACS.

Sorted cells were washed and resuspended at a concentration of ~500 cells/μl. For microfluidics-enabled scRNAseq, cell suspensions were mixed with Fluidigm C1 Suspension Reagents (Fluidigm 100-5315) at a ratio of 8:2 before loading mix onto C1 chip (Fluidigm 100-5760). Bright-field images of captured cells were collected using a Keyence BZ-X710 microscope (Keyence Corporation, Itasca, Illinois, USA). Single-cell RNA isolation and amplification were performed using the Fluidigm C1 Single-Cell Auto Prep IFC following the Fluidigm Protocol: 100-7168 I1. RNA spike-in controls were omitted. cDNA library preparation were performed following the Fluidigm C1 Protocol: 100-7168 I1.

For droplet-enabled scRNAseq, flow cytometry sorted cells were washed in PBS with 0.04% BSA and resuspended at a concentration of ~1000 cells/μl. Library generation for 10× Genomics v1 chemistry was performed following the Chromium Single Cell 3' Reagents Kits User Guide: CG00026 Rev B. Library generation for 10× Genomics v2 chemistry were performed following the Chromium Single Cell 3' Reagents Kits v2 User Guide: CG00052 Rev B.

Quantification of cDNA libraries was performed using Qubit dsDNA HS Assay Kit (Life Technologies Q32851) and high-sensitivity DNA chips (Agilent. 5067-4626). Quantification of library construction was performed using KAPA qPCR (Kapa Biosystems KK4824). For microfluidics-enabled scRNAseq libraries, we generally multiplexed 96 cells per lane on an

Illumina HiSeq2500 resulting in a calculated depth of ~1.6 million reads per cell (Illumina Rapid PE kit v2 402-4002 and Rapid SBS kit v2 FC 401-4022). For droplet-enabled scRNAseq, we used the Illumina HiSeq4000 platform to achieve an average of 50,000 reads per cell.

Processing of scRNAseq data

After demultiplexing sequencing libraries to individual cell FASTQ files (observed average read depth per cell was found to be ~1.6 Million reads), each library was aligned to an indexed GRCh38 RefSeq genome using RSEM version 1.2.1242, and bowtie2 version 2.2.3 with the following options enabled: rsem-calculate- expression -p $CORES—bowtie2—paired-end -output- genome-bam. Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values were quantified and concatenated into a resulting gene expression matrix for each library, which was then loaded into R for subsequent computational analysis. For quality control filtering, we generally excluded libraries with less than 900 genes detected. In addition, genes that were not detected in at least 3 of the cells after this trimming were also removed from further analysis. Alignment of 3' end counting libraries from droplet-enabled scRNAseq analyses was completed utilizing 10× Genomics Cell Ranger 1.3.1. Each library was aligned to an indexed GRCh38 genome using Cell Ranger Count. "Cell Ranger Aggr" function was used to normalize the number of confidently mapped reads per cell across the libraries from different individuals utilizing 10× v2 chemistry.

Cluster identification using Seurat

For cluster identification in both microfluidics- and droplet-enabled scRNAseq datasets, we utilized the Seurat pipeline17. The data matrices were imported into R and were

processed with the Seurat R package version 1.2.1, where the FPKM values were transformed into log-space after the aforementioned trimming steps (each gene was expressed in at least three cells, each cell has at least 900 genes). PCA was performed using highly variable genes in the trimmed dataset. Using the first two PC's as input, we then performed density clustering to identify groupings in the data and t-distributed statistical neighbor embedding (tSNE) to visualize. Using further Seurat functionality, marker genes for each respective cluster were identified and used for subsequent analysis.

For droplet-enabled scRNAseq data, we used the Seurat R package version 2.0.0. Data was read into R as a counts matrix and transformed into log-space. Due to the difference in gene detection across the two platforms, differences in chemistry for the library prep, as well as sequencing depth per cell, a minimum cutoff of 500 and a maximum cut-off of 6000 genes per cell for this dataset was used. In addition, cells with a percentage of total reads that aligned to the mitochondrial genome (referred to as percent mito) greater than 10% were removed, since increased detection of mitochondrial genes can be associated with cells undergoing stress and cell death (Ilicic, 2016).

To account for the possibility of individual cell complexity driving cluster separation, we employed Seurat's "RegressOut" function to reduce the contribution of both the number of UMI's and the percent mito. Variable genes were then determined for subsequent PCA for each separate individual. For tSNE projection and clustering analysis, we used the first ten principal components. We used the feature plot function to highlight the expression of known marker genes for basal (e.g., KRT5, KRT14) and luminal cells (e.g., KRT8, KRT18) to identify which clusters belonged to which epithelial cell type. The specific markers for each cluster identified by Seurat were determined using the "FindAllMarkers" function.

Cluster comparisons and assignment

Cluster specific marker genes from the individual library analyses were used as input lists to the previously described gene scoring method (described in more detail below) to compare cluster signatures in a pairwise manner between individuals. To visualize pairwise gene scoring results, we generated heatmaps displaying averaged gene scoring results for each cluster. We overlaid individual-specific cluster designations onto these heatmaps to find which individual clusters best match to each other. Clusters were merged together in the case that multiple clusters scored highly. We performed a separate Seurat analysis using combined basal cells from all four individuals, and then matched clusters using the gene scoring method on a set of genes curated to represent a myoepithelial cell fate25 to score and classify the clusters as either Basal (B) or Myoepithelial (Myo) cell state.

Gene scoring

To compare gene signatures and pathways in epithelial subpopulations, we utilized individual gene scores as described previously12. Briefly, each score was generated by calculating total gene expression for each of the analyzed genes and separating them into 25 bins of similar expression. For every gene in each target pathway or signature, 100 "control" genes were selected from its corresponding bin and added to a "control" pathway. The resulting "control" pathway contained an equivalent expression distribution as the target pathway and its average represents an equivalent sampling of 100 pathways of equal size to the target pathway. The expression of genes in the target pathway and the "control" pathways was averaged across each cell to generate a target score (STarget) and control

score (SCtrl). The cell's score for the target pathway (SPath) is the difference between the target score and control score: SPath = STarget – SCtrl. To determine statistical significance, we used the unpaired Wilcox test with a 95% confidence interval.

Gene set and pathway analysis

Cells belonging to subpopulations were averaged to serve as a representation of each subgroup, and trimmed to their respective marker genes as determined by Seurat following log2 transformation. Each subpopulation sample was then uploaded to Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com) core analysis feature and compared. A p-value of 0.05 was used as a cut-off to determine significant enrichment of a pathway or annotated gene grouping present in the Ingenuity Knowledge base. In addition, comprehensive gene set enrichment was done using Enrichr26 based on the cell type and state-specific marker genes identified by Seurat.

Immunofluorescence analysis

Tissues were fixed in 4% formaldehyde for 24 h, dehydrated in solutions of increasing concentrations of ethanol, cleared with xylene, and embedded in paraffin. Slides of 10-µm sections were prepared using a Leica SM2010 R Sliding Microtome (Leica Biosystems, Wetzlar, Germany). Slides were heated at 65 °C for 1 h, followed by two 5-min incubations in Histo-Clear (National Diagnostics, Cat. No. HS-200, Atlanta, Georgia, USA) for paraffin removal. Tissues were rehydrated with solutions of decreasing concentrations of ethanol, washed in double-distilled H2O and PBS, and subjected to antigen retrieval using a microwave pressure cooker with 10 mM citric acid buffer (0.05% Tween 20, pH 6.0). Tissues

were blocked in blocking solution (0.1% Tween 20 and 10% Goat Serum in PBS) for 20 min at room temperature, incubated with primary antibodies prepared in blocking solution at 4 °C overnight, washed in PBS, incubated with secondary antibodies diluted in PBS for 1 h at room temperature, and washed in PBS. Slides were mounted with VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories, Cat. No. H-1200, Burlingame, California, USA) and micrographs were taken with the BZ-X700 Keyence fluorescent microscope. For quantification of staining (e.g., ZEB1 and KRT14 staining), we manually counted positive cells as signal around nuclei (DAPI) and utilized the BZH Hybrid Cell Count software (Keyence) in at least three different fields of view using a 40× objective in at least two different samples.

Primary Antibodies: Estrogen Receptor (ER) rat mAb diluted 1:50 (Cat. No. 916201); KRT14 rabbit pAb diluted 1:500 (Cat. No. PRB-155P) (Biolegend, San Diego, CA, USA); Androgen Receptor (AR) rabbit mAb diluted 1:400 (Cat. No. 5153); Progesterone Receptor (PR) rabbit mAb diluted 1:1000 (Cat. No. 8757) (Cell Signaling, Danvers, MA, USA); KRT8 (TROMA-1) mouse mAb diluted 1:500 (DSHB, Iowa City, Iowa, USA); SLPI goat pAb diluted 1:200 (R&D Systems, Cat No. AF1274-SP, Minneapolis, MN, USA); α-Smooth Muscle Actin mouse mAb diluted 1:500 (Cat. No GTX60466), Ki67 mAb diluted 1:200 (Cat. No. GTX16667); TP63 rabbit pAb diluted 1:500 (Cat. No. GTX102425), MUC1 rabbit pAb diluted 1:500 (Cat. No. GTX15481), ACTA2 mouse mAb diluted 1:500 (Cat. No. GTX60466); TCF4 rabbit pAb diluted 1:500 (Cat. No. GTX54531); E-cadherin (DCH1) rabbit pAb diluted 1:500 (Cat. No. GTX100443); KRT18 rabbit pAb diluted 1:500 (Cat. No. GTX112978) (GeneTex, Inc., Irvine, California, USA); ACTA2 mouse mAb diluted 1:500 (Cat. No. MA511547); NY-BR-1 mouse mAb diluted 1:500 (Cat. No. MS-1932-P0); KRT14 mouse mAb diluted 1:100 (Cat. No.

MA511599); and KRT18 mouse mAb diluted 1:100 (Cat. No. MA512104) (Thermo Fisher Scientific Inc., Carlsbad, California, USA).

Secondary Antibodies: Donkey anti-mouse Cy5.5-conjugated IgG (Novus Biologicals, Cat. No. NBP1-73774, Littleton, CO, USA); Goat anti-rabbit IgG conjugated with Alexa Fluor 568 and 488 (Cat. No. A21069 & A11034); Goat anti-mouse IgG conjugated with Alexa Fluor 568 and 488 (Cat. No. A11004 & A11001); Goat anti-rat IgG conjugated with Alexa Fluor 488 (Cat. No. A11006); Donkey anti-rabbit FITC-conjugated IgG (Cat. No. A16030); and Donkey anti-goat IgG conjugated to FITC and Alexa Fluor 568 (Cat. No. A16006 & A11057) (Thermo Fisher Scientific Inc., Carlsbad, California, USA).

Single-cell western blot

Single-cell western blots were completed using the Single-Cell Western instrument Milo, scWest chips, and reagents from ProteinSimple (San Jose, CA). A standard 6%T scWest chip was re-hydrated in 1× Suspension Buffer for 15 min at room temperature. A volume of 1 mL of flow cytometry-sorted human mammary epithelial cells (combined basal and luminal) at 100,000 cells/mL were settled in medium onto the scWest chip for 15 min at room temperature. Un-captured cells were washed away with 1 mL of media. Captured cells were lysed for 10 s, then individual cell protein lysates were electrophoretically separated for 1 min at 240 V, and proteins were UV-captured for 4 min. After running on Milo, the scWest chip was washed 2 × 10 min in 1× Wash Buffer, then probed for mouse anti-cytokeratin 8 (Abcam ab9023) at 200 µg/mL and rabbit anti-β-tubulin (Abcam ab6046) at 100 µg/mL for 2 h at room temperature. Primary antibodies were diluted in 1 × Wash Buffer (final) containing 5% (w/v) BSA. After 3 × 10-min washes in 1× Wash Buffer, the scWest chip

was incubated with donkey anti-rabbit IgG Alexa 647 (A-31573 ThermoFisher Waltham, MA) and donkey anti-mouse IgG Alexa 488 (A-21202 ThermoFisher) at 100 μg/mL in 1× Wash Buffer containing 5% BSA for 1 h in the dark at room temperature. The chip was then washed 3 × 15 min in 1× Wash Buffer, dried, and imaged using a Molecular Devices Genepix 4400A (Sunnyvale, CA) (Standard Blue Filter 500 gain, Standard Red Filter 600 gain). Images were saved as single-color tiffs and analyzed using Scout software (ProteinSimple).

Reconstructing differentiation trajectories using Monocle

Cell fate decisions and differentiation trajectories were reconstructed with the Monocle 2 package, which utilizes reverse graph embedding based on a user-defined gene list to generate a pseudotime plot that can account for both branched and linear differentiation processes. For pseudotemporal analysis of breast epithelial cells in C1 data, we used Monocle version 2.2.0, ordered a combined set of cells from all three individuals on a list of marker genes as determined by Seurat analysis using up to 20 genes per cluster with least 0.5 power (Supplementary Data 2). Labels of basal and luminal cells respectively were assigned according to the identity of the cells from the initial cell sorting and ZEB1 positive cells were labeled based on expression level >0. For pseudotemporal analysis of droplet-based scRNAseq data, we first ordered the four individuals in Monocle 2.2.0 separately using cell type markers identified in the C1 analysis along with the top 20 marker genes for each subpopulation in Seurat. Next, for each of these four datasets, we identified genes differentially expressed between trajectory clusters (States), averaged the gene expression values for all cells within each State, and generated a Pearson correlation matrix for these average gene expression values across States. We averaged the four correlation matrices into

one matrix and kept only genes that had an average Pearson correlation of 0.8 with at least one other gene. Finally, we ordered a random subsample of 4000 cells (1000 cells from each individual) by the genes from our correlation analysis that overlapped with Seurat identified subpopulation marker genes (Supplementary Data 2).

Comparison of subpopulations to breast cancer subtypes

To learn more about the relationship of the newly defined normal breast epithelial subpopulations to the known breast cancer subtypes, we used the gene scoring method to compare each subpopulation to previously described triple-negative breast cancer subtypes. To this end, we utilized the genes that are specifically up-regulated in each subtype as previously reported (Curtis, 2012; Lehmann, 2011). To compare each subpopulation to METABRIC derived molecular subtype signatures, the METABRIC microarray expression dataset was downloaded and processed using the R Bioconductor package Limma version 3.30.13. Samples were grouped by their annotated molecular subtype, and differentially expressed genes was calculated for each group. The top 20% of the upregulated genes was sorted by log-fold change were then used for downstream scoring.

Figure 2.1. Identification of three major epithelial cell types and their markers using scRNAseq. **a** Overview of scRNAseq approach using primary human breast tissue samples that were processed into single-cell suspension, followed by FACS isolation of basal (CD49f-hi, EPCAM+) and luminal (CD49f+, EPCAM-hi), and scRNAseq analysis using the microfluidics-enabled scRNAseq. **b** Combined tSNE projection of cells from all three microfluidics-enabled scRNAseq datasets. The major basal cluster is highlighted in red; Luminal1 (L1) in green; Luminal2 (L2) in blue. **c** Heatmap displaying the scaled expression patterns of top marker genes within each cell type with selected marker genes highlighted; yellow indicating high expression of a particular gene, and purple indicating low expression. **d** Feature plots showing the scaled expression of TCF4 and ZEB1 marking a subpopulation of basal cells and gene plot showing co-expression of TCF4 and ZEB1 in the same cells. See Supplemental Figure 2.1 capture site imaging, gene detection, individual principal component analysis, tSNE plot colored by individual-derived cells, and feature plots of cell type-specific markers. (Panel b-d, generated by Kai Kessenbrock)

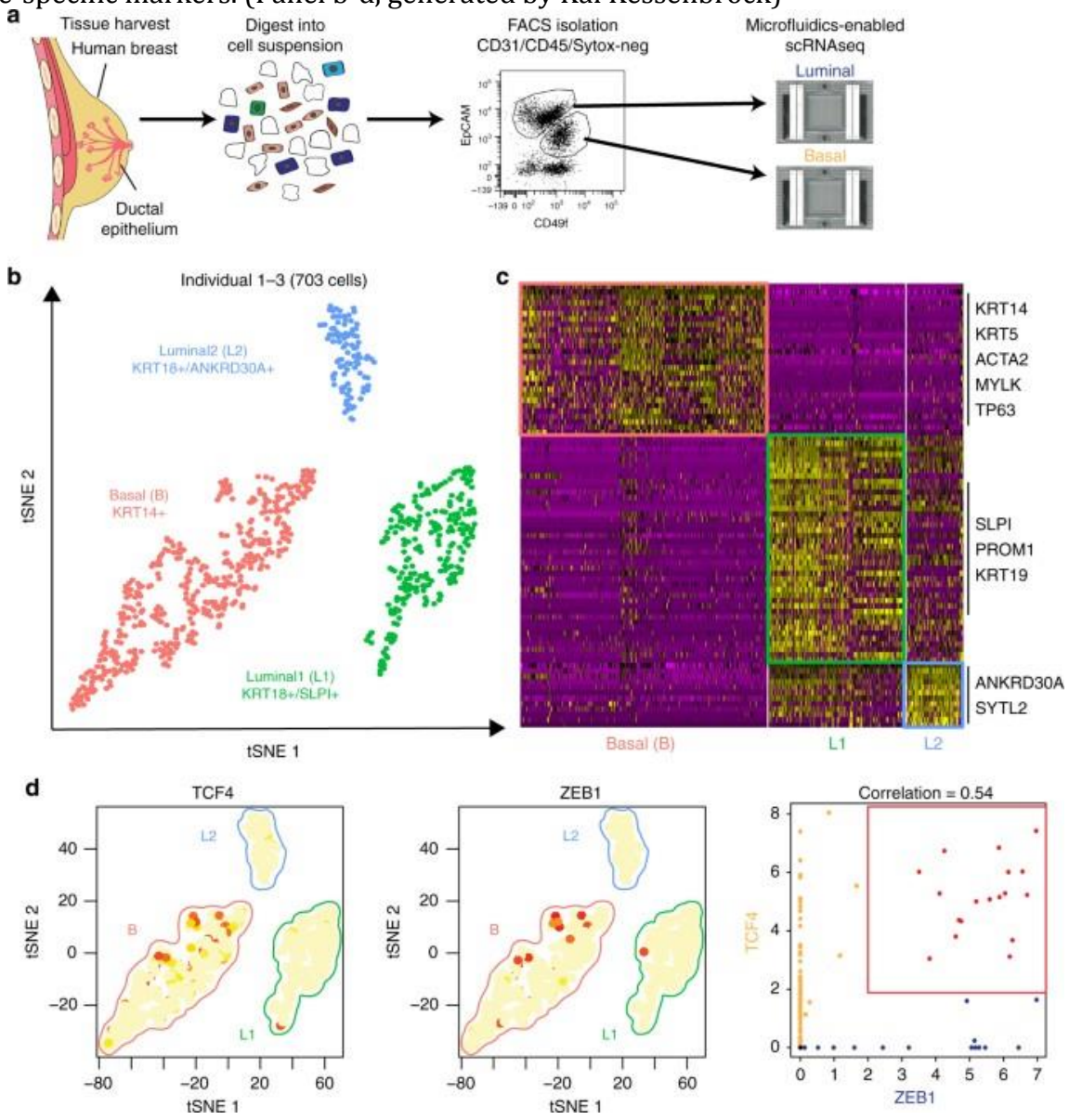


Figure 2.2. Technical information and supportive data on microfluidics-en-abled scRNAseq. (a-b) All 96 capture sites were imaged using the Keyence BZ-X700 microscope to confirm single-cell capture (a), and to exclude capture sites that contained doublets or multiplets (b). (c) Number of genes detected per cell were distributed in a comparable manner between basal (red) and luminal cells (green). (d) tSNE projection of data generated on microfluidics-enabled scRNAseq data, with cells colored by the individual sample source. (e) Feature plots showing the scaled expression of ANKRD30A marking cell type L2, KRT18 marking both luminal cell types, SLPI marking cell type L1 with greater specificity, and KRT14 showing the highest expression in basal cells. (Panel a-e, generated by Kai Kessenbrock)

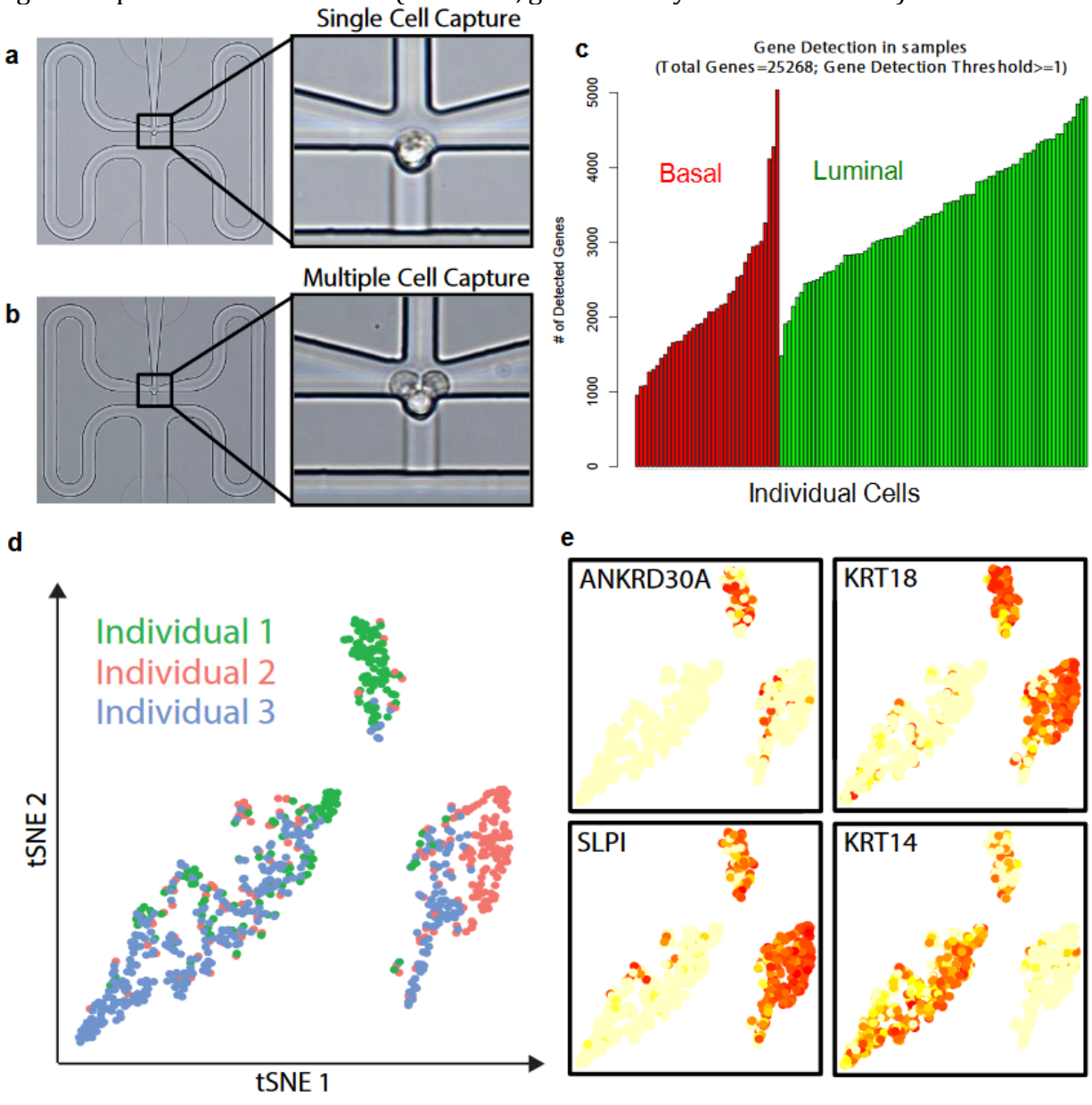


Figure 2.3. High throughput droplet-mediated scRNAseq reveals additional epithelial cell states. **a** Overview for droplet-enabled scRNAseq approach as described above; basal and luminal epithelial cells were sorted together and subjected to combined scRNAseq analysis using the droplet-based scRNAseq. **b** Data from individual four was analyzed using Seurat and the distinct clusters (0–10) are displayed in tSNE projection with selected marker gene for each cluster, and main epithelial cell types (Basal, L1, L2) are outlined. Feature plots of characteristic markers for the three main cell types are shown on the right showing expression levels as a gradient of purple. **c** Heatmap showing the top ten marker genes for each cluster as determined by Seurat analysis with three selected genes per cluster highlighted on the right. See Supplemental Figure 2.2 for individual clustering and marker gene analyses for Individuals 5–7.

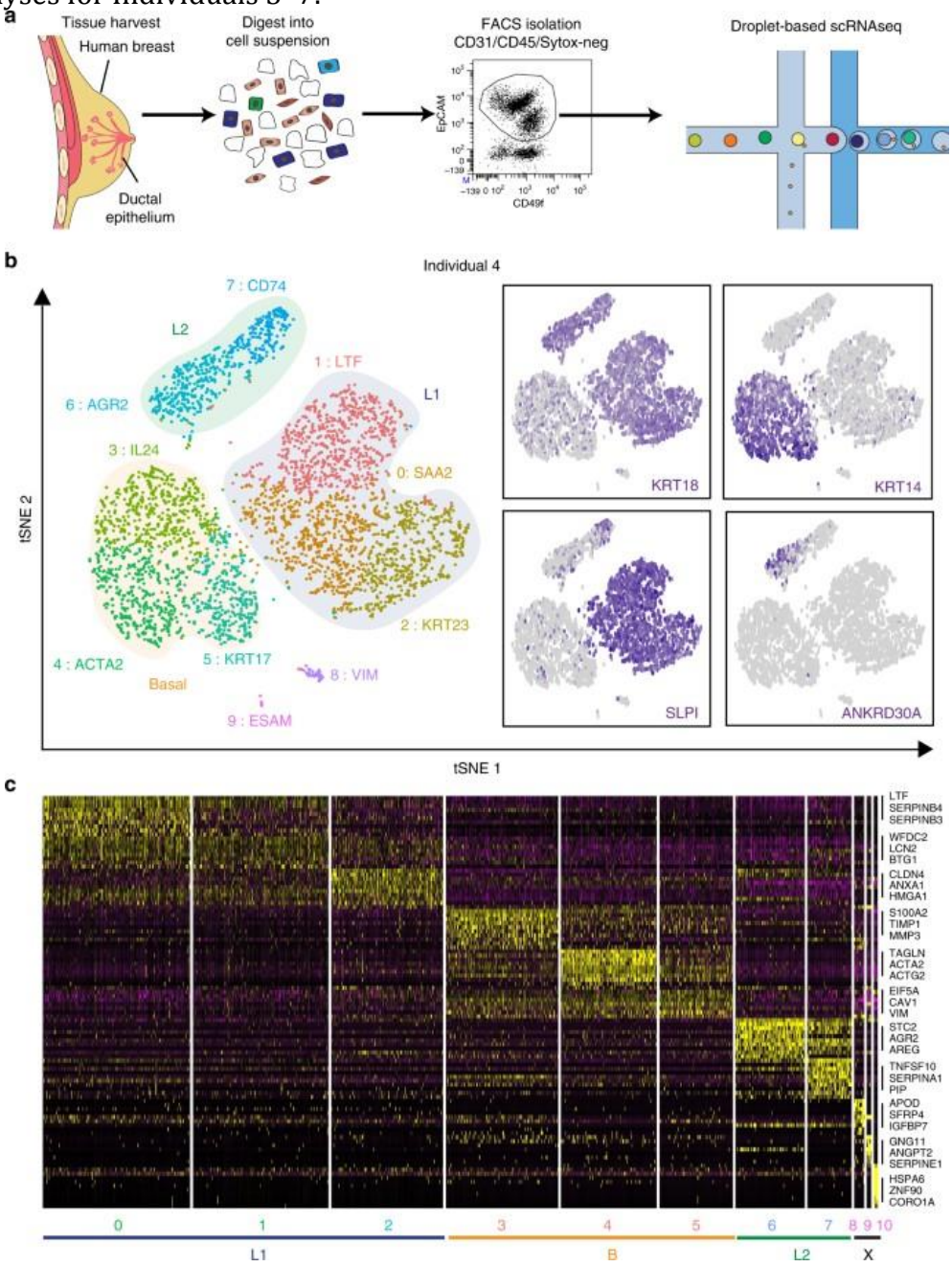


Figure 2.4. Clustering analysis and marker gene determination for individuals 5-7. (a-c) The individual data matrices for Individual 5 (a), 6 (b), and 7 (c) were analyzed using Seurat and their initial cluster determinations are displayed using tSNE projection. Feature plots of characteristic markers of highlighting the three main cell types Basal, L1, and L2 are shown. Additional less frequent non-epithelial populations were detected in some individuals and were designated unclassified (X). Heatmaps showing the top 10 marker genes of each cluster are displayed highlighting selected marker genes for each cluster.

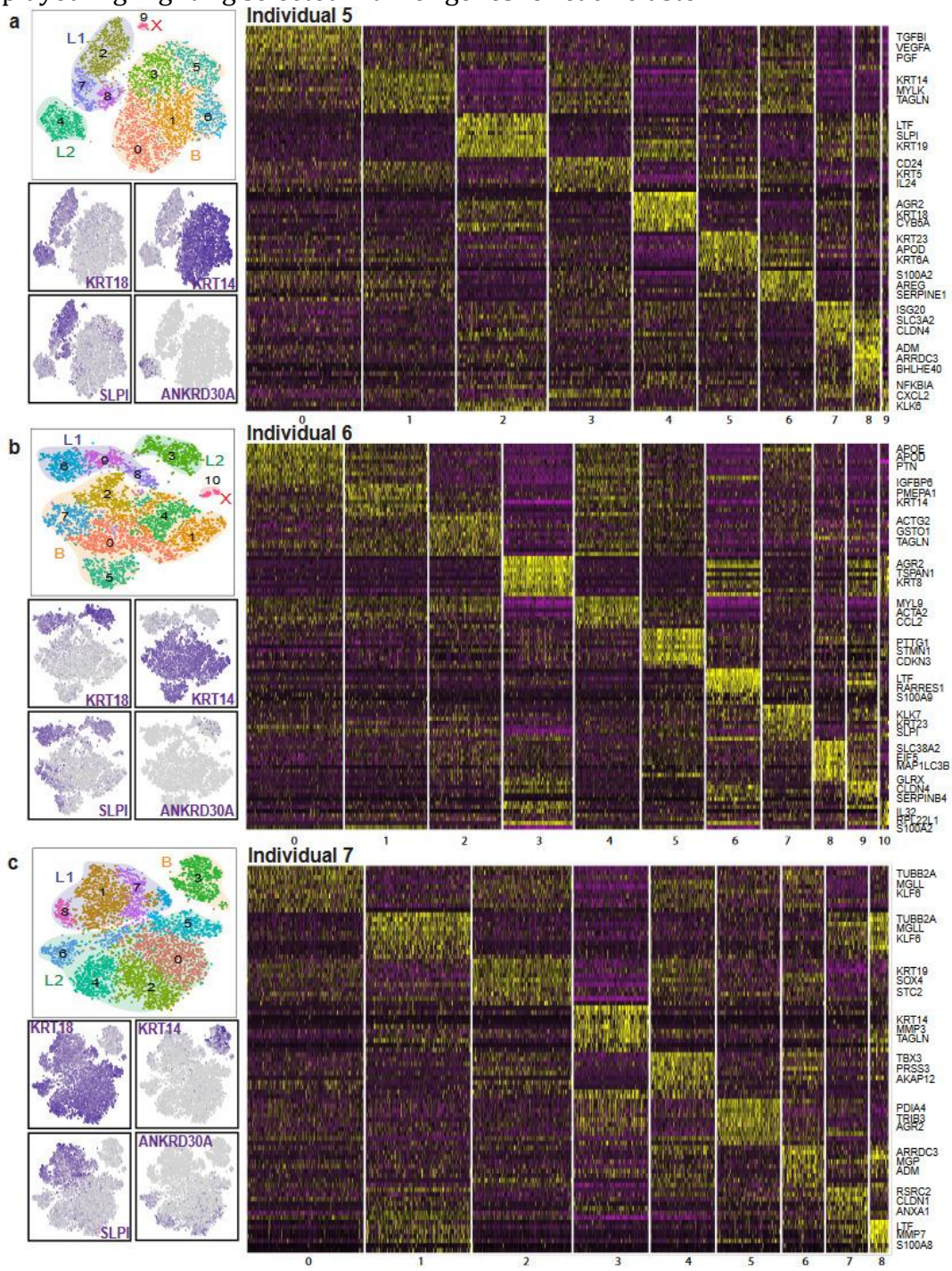


Figure 2.5. Combined droplet-based RNAseq data to identify generalizable cell types and states. **a**–**c** Heatmaps showing gene scoring results using marker genes for Ind4 clusters (0–10; on bottom of heatmap) in all clusters from Ind5 (**a**), Ind6 (**b**), and Ind7 (**c**). Individual-specific cluster IDs are shown in different colors on the right and bottom, and cell type IDs for Basal (**b**), L1, L2, X are indicated for every cluster. Data are shown as *Z* scores from purple (low) to yellow (high). Two distinct cell states L1.1 and L1.2 were found within L1 in all pairwise comparisons as highlighted by colored boxes on the heatmap. **d** Combined tSNE projection of all individual datasets (outlined) is shown including the cell state identity marked by different colors. **e** Heatmap showing the expression pattern of the top ten markers per cell state with selected markers indicated (yellow = high expression; purple = low expression). See Supplemental Figure 2.4 for separate basal cell Seurat analysis, summary of cell state designations, and Ingenuity Pathway Analysis. (Panel a-e, generated by Nicholas Pervolarakis)

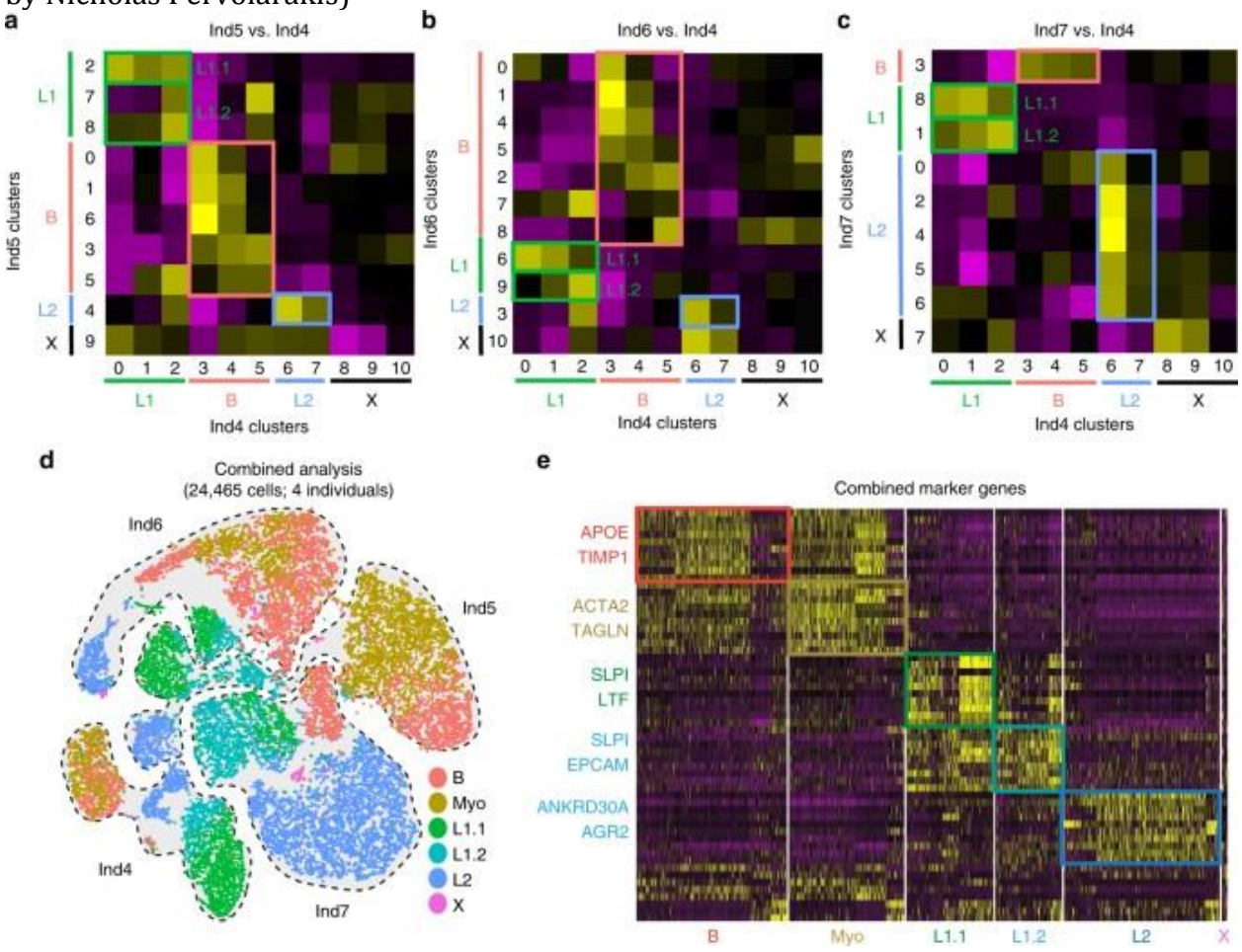
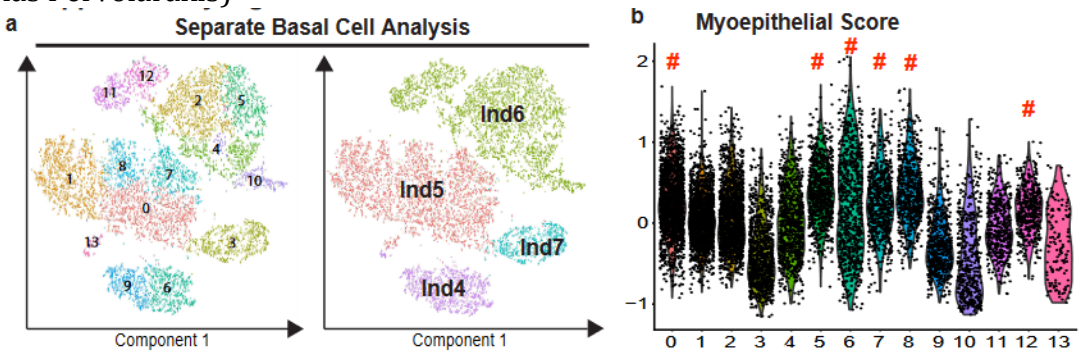


Figure 2.6. Combined basal cell only analysis and ingenuity pathway analysis (IPA). (a) Basal cell clusters (KRT14+) from all four droplet-enabled scRNAseq datasets were combined and analyzed using Seurat. tSNE projections and of cells belonging to the basal cell lineage across all individuals in a combined analysis, colored by cluster determination and individual library source. (b) Violin plots showing the gene scoring results for a curated Myoepithelial gene signature (see Supplementary Data 2) was used to stratify regular basal cells from myoepithelial clusters(marked by #). (c) Summary of individual cluster matches and final cluster assignments as indicated in "Cell State" column. Basal cell populations were separately analyzed and then scored using a myoepithelial signature gene list, resulting in the final cell state determinations of Basal (B), Myoepithelial (Myo), Luminal1.1 (L1.1), Luminal1.2 (L1.2), Luminal2 (L2), and Unclassified (X). (d) Heatmap showing log-scaled p-value of enrichment for IPA annotated pathways, processed via comparison of IPA expression enrichment analysis on marker genes for each cluster. (Panel a-d, generated by Nicholas Pervolarakis)



c

Clusters from Individual Analysis

Ind4	Ind5	Ind6	Ind7	Cell State
3	0,1,6	0,1,4,5	3	
4	3,5	2,7	3	Basal
5	-	8	3	
0	2	6	7,8	Luminal1.1 (L1.1)
1	2	6	7,8	
2	7,8	9	1	Luminal1.2 (L1.2)
6	4	3	0,2,4,5	Luminal2 (L2)
7	4	3	6	
8,9,10	9	10	-	Unclassified (X)

Clusters from separate Basal Analysis

1,2,3,4,9,10,11,13	Basal (B)
0,5,6,7,8,12	Myoepithelial (Myo)

d

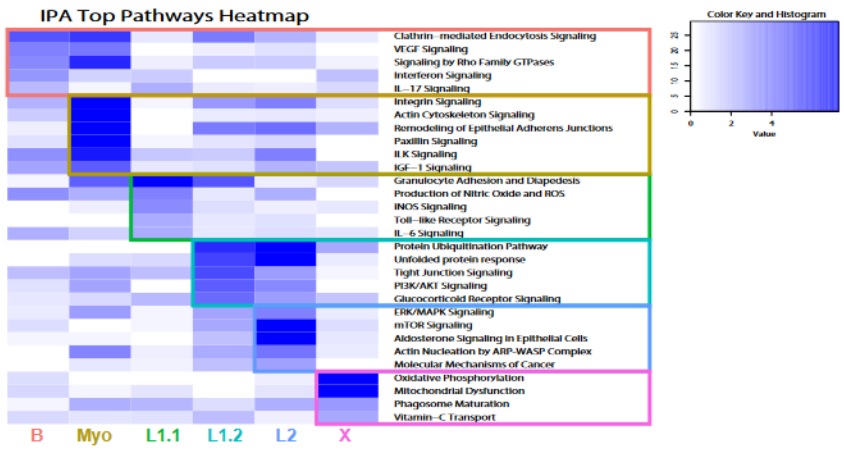


Figure 2.7. Characterization and spatial integration of basal cell states. **a** Immunofluorescence analysis of ZEB1 protein expression (red) in combination with basal marker KRT14 (green) and DNA stain using DAPI (blue) within tissue sections from primary human reduction mammoplasty samples showing ZEB1 expression in a subpopulation of basal (KRT14+) cells. Scale bar = 15 µm. **b** Heatmap showing expression of genes previously shown to be up- (red) or down-regulated (blue) in a population of PROCR+ mammary stem cells show a correlation with ZEB1+ cells in scRNAseq. **c** Immunofluorescence analysis of TCF4 protein expression (red) in combination with basal marker SMA (green) and DNA stain using DAPI (blue) within tissue sections from primary human reduction mammoplasty samples revealed that TCF4 is expressed in a subpopulation of basal (SMA+) cells. Scale bar = 25 µm. **d** Violin plot for expression of KRT14 by cell state showing highest expression in the myoepithelial (Myo) cells. **e** KRT14 and KRT8 double immunostaining revealed highest expression of KRT14 in ductal basal cells, while lobular basal cells show more diverse KRT14 positivity. Scale bar = 75 µm. See Supplemental Figure 2.4 for violin plots displaying selected myoepithelial gene expression and identification of KRT8/KRT14 double-positive cells.

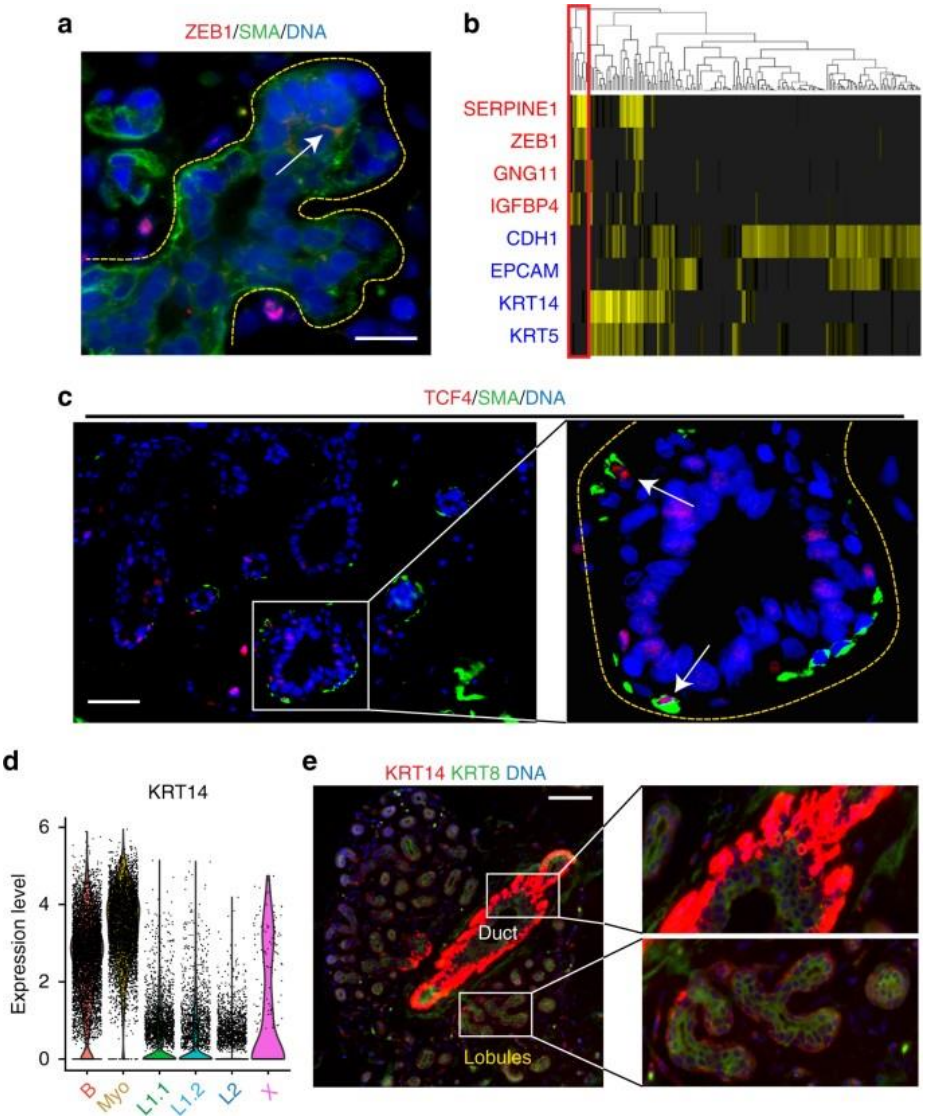


Figure 2.8. Expanded characterization of cellular heterogeneity within the basal compartment. (a) Violin plots of the expression of ACTA2, MYLK, TAGLN, and MYL9 in the combined analyses of the droplet-enabled scRNAseq data grouped by final cluster determination. (b) Correlated expression analysis of luminal marker KRT8 and basal marker KRT14 from scRNAseq data revealed a significant number of double-positive cells. (c) Combined immunostaining for KRT8 and KRT14 showing rare foci of double-positive cells in the luminal cell layer of lobular regions. Scale bar = 50 µm.

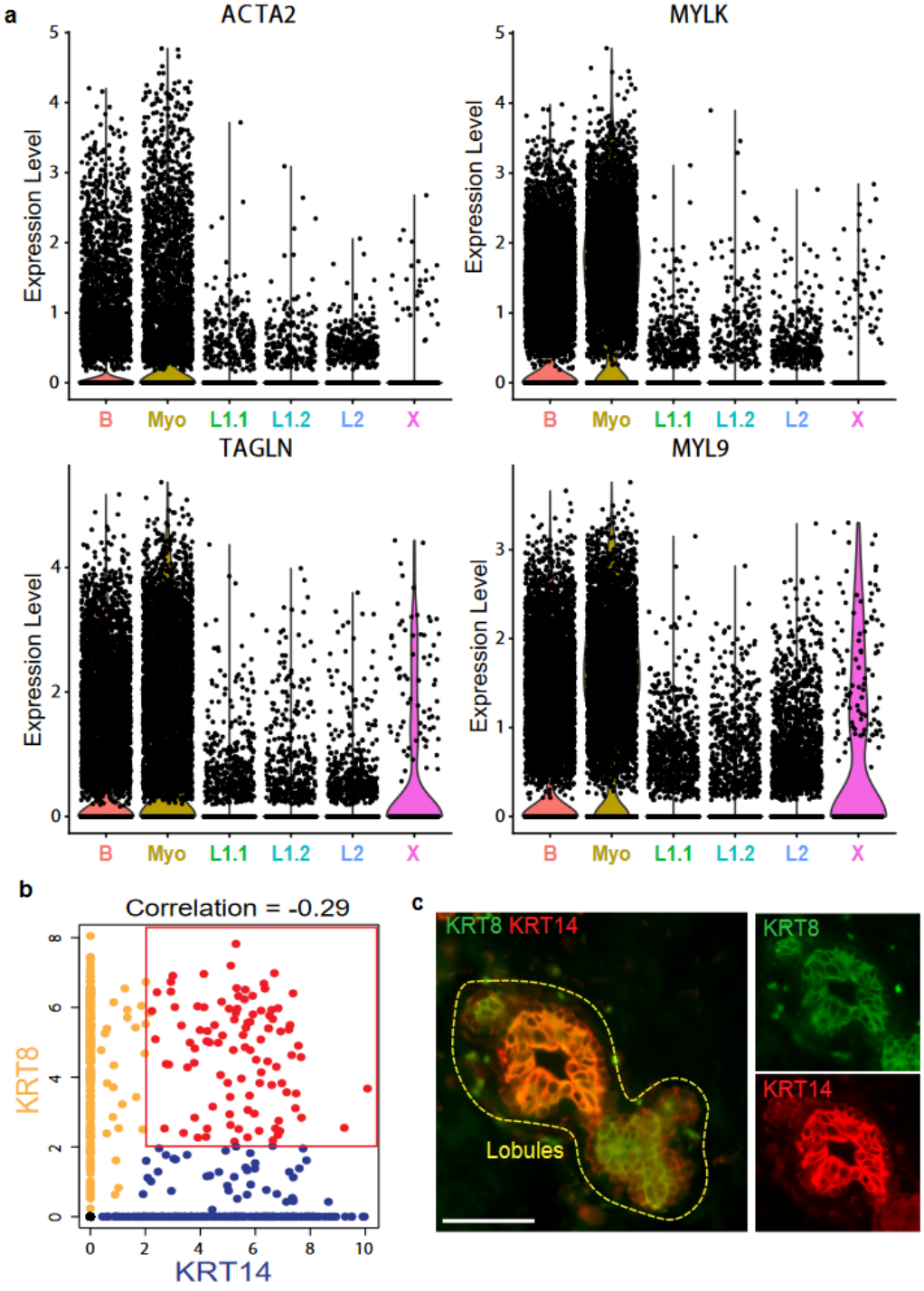


Figure 2.9. Validation and spatial integration of two distinct luminal cell types. **a** Immunofluorescence analysis of NY-BR-1 protein expression (green) in combination with basal marker SLPI (red) and DNA stain using DAPI (blue) within tissue sections from primary human reduction mammoplasty samples revealed that NY-BR-1 and SLPI are markers for distinct luminal subpopulations. **b**–**e** Immunofluorescence analysis of NY-BR-1 and SLPI (red) protein expression with: hormone receptors for estrogen receptor (**b**), progesterone (**c**), and androgen (**d**) and proliferation marker Ki67 **e** in green. **f** Summary of hormone receptor and proliferation marker expression in L1 and L2 cells. **g** Violin plot showing expression of KRT8 in the luminal subpopulations, higher expression is seen in the luminal L1.1 and L1.2 subpopulation. **h** Sample frame for detection of KRT8 protein content from individual cells using single-cell Western blot following detection using microarray scanner. **i** Population summary showing cell number per fluorescence intensity confirmed bimodal distribution of KRT8 expression on the protein level. See Supplemental Figure 2.5 for violin plots displaying expression of relevant hormone receptors as well as proliferation and luminal progenitor markers. All scale bars = 25 µm. (Panel a-f, generated by Dennis Ma, panel h-I, generated by Kai Kessenbrock)

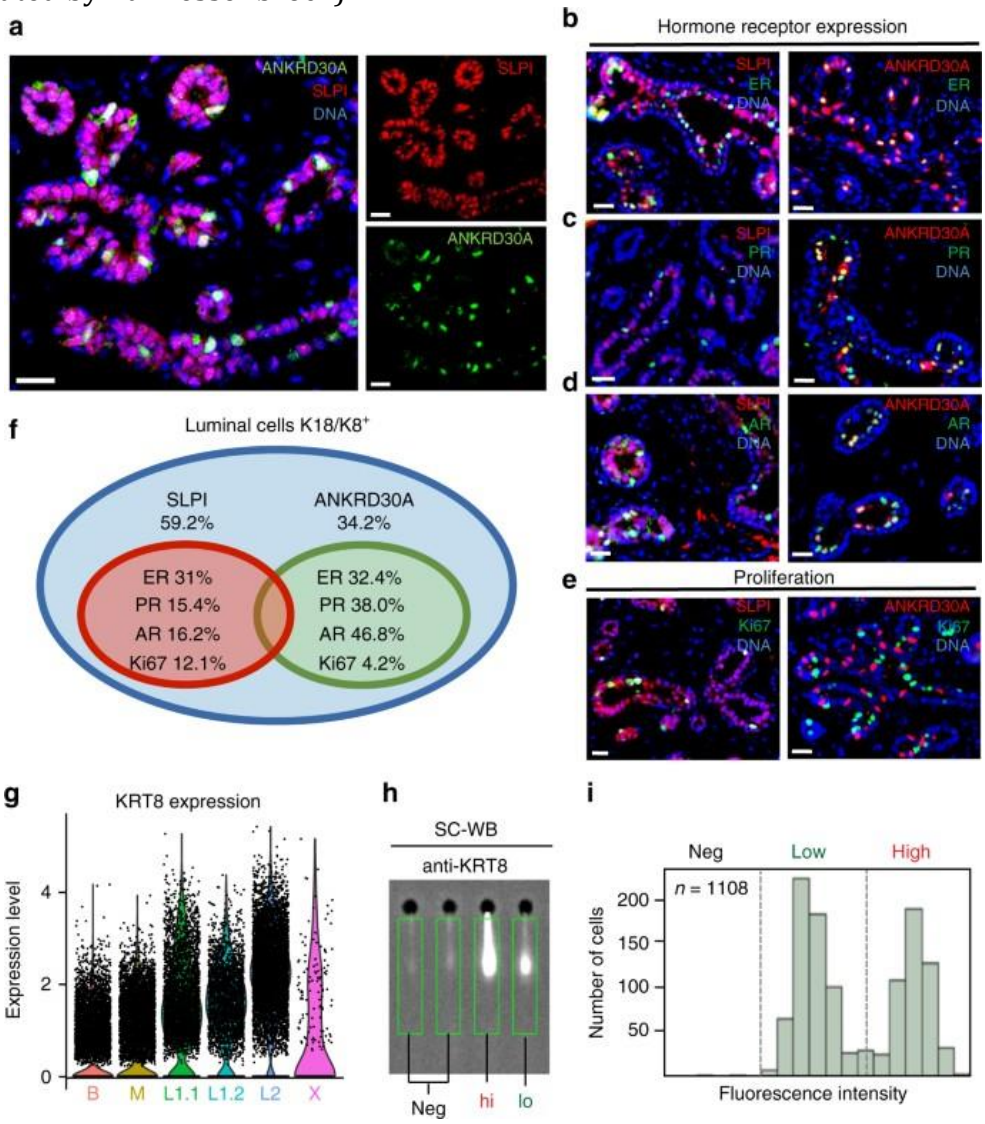


Figure 2.10. Expression patterns for selected genes of interest in combined analysis of droplet-enabled scRNAseq datasets. (a) Violin plots illustrating the expression patterns of hormone receptors, cell cycle genes MKI67 and CDKN1B (p27) and gene scoring for G2M gene signature (b), and luminal progenitor genes (c) grouped by final cluster determination.

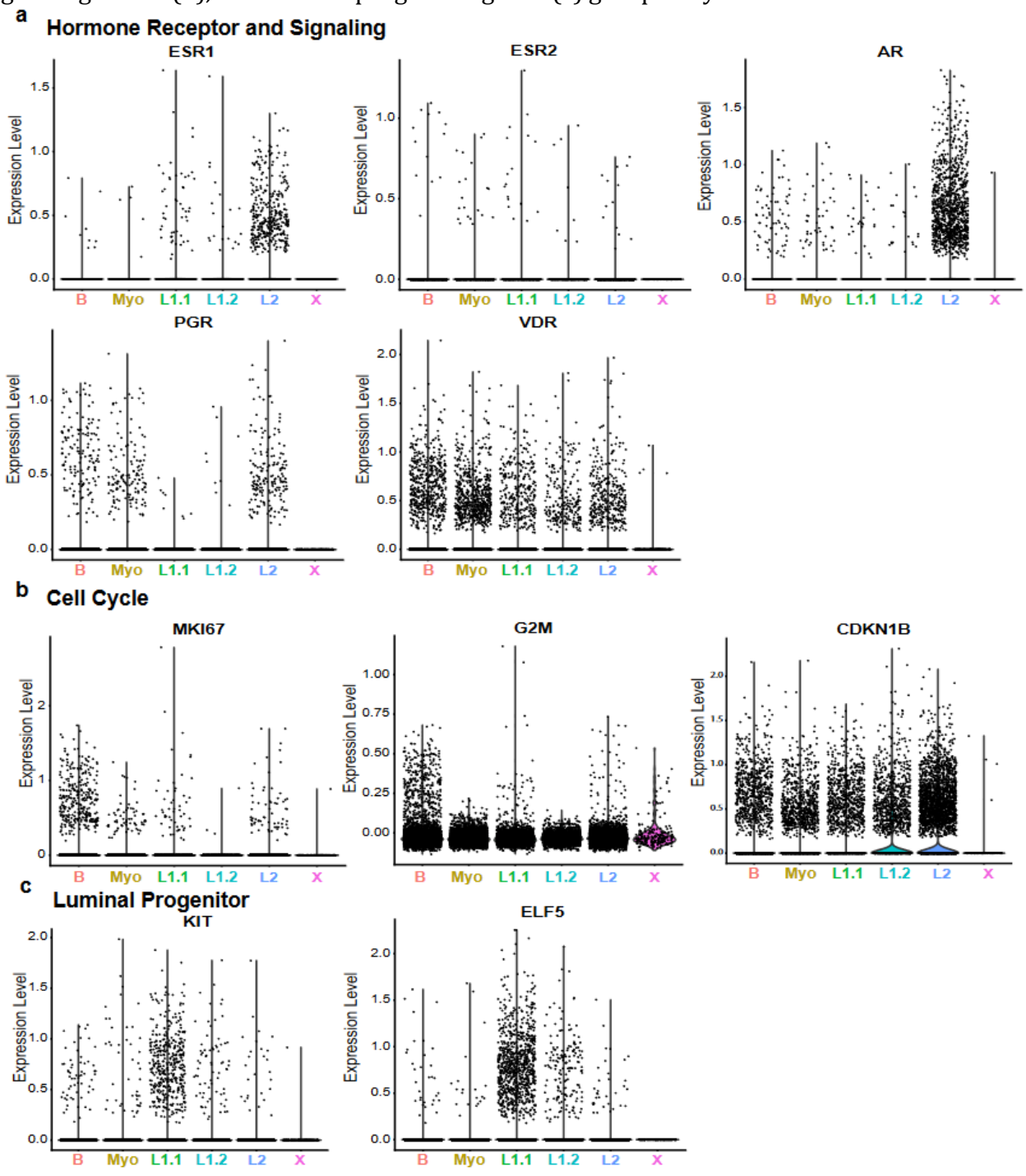


Figure 2.11. Reconstruction of differentiation and relation of cell states to breast cancer subtypes. **a** Monocle-generated pseudotemporal trajectory of a subsampled population of cells ($n = 4000$) from four individuals analyzed using droplet-mediated scRNAseq is shown colored by cell state designation. **b** Pseudotime is shown colored in a gradient from dark to light blue and start of pseudotime is indicated. See Supplemental Figure 6 for summary list of discovered cell states, Monocle analysis of microfluidics-enabled scRNAseq results, and gene scoring for breast cancer subtypes. (Panel a-b, generated by Kerrigan Blake)

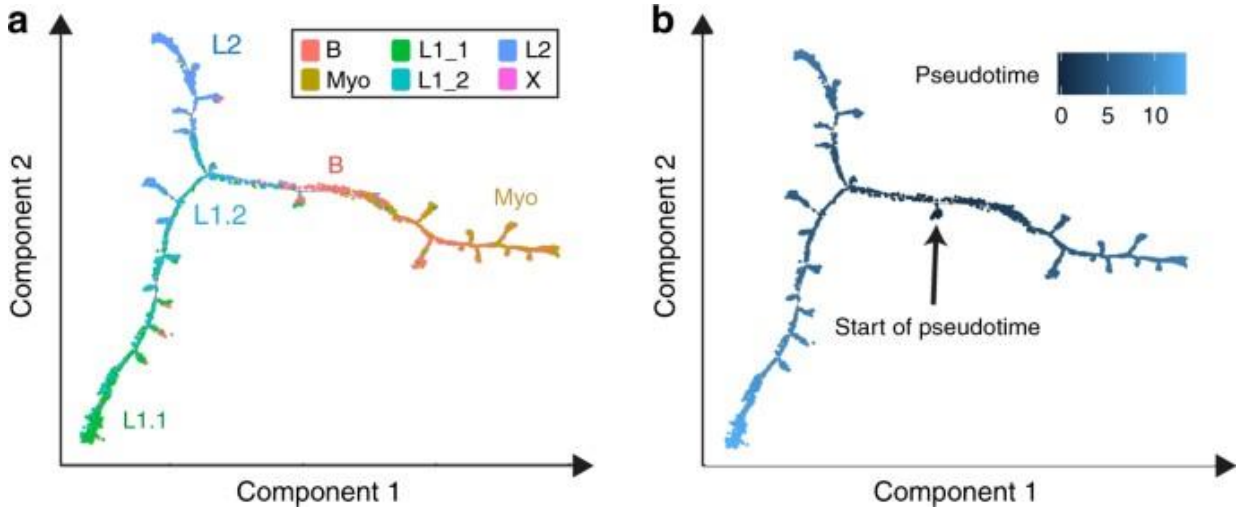


Figure 2.12. Reconstructing breast epithelial lineage hierarchies their relation to breast cancer. (a) Pseudotemporal analysis of microfluidics-enabled scRNAseq results using Monocle2 based on a set of 183 Seurat identified marker genes suggest a differentiation trajectory from ZEB1+ progenitor cells (green) bifurcating into basal (red) and luminal (blue) differentiated cells. (b) Selected marker genes are shown as dot plots displayed as expression level over pseudotime. (c) Relation of cell states identified in droplet-enabled scRNAseq analysis to different breast cancer subtypes is shown as violin plots displaying gene scoring results for a cells on gene lists derived from breast cancer subtypes, namely Metabric Luminal A (LumA), Metabric Luminal B (LumB), triple-negative breast cancer (TNBC) mesenchymal-like, and TNBC-Basal1. (Panel a-b generated by Kerrigan Blake, panel c, generated by Nicholas Pervolarakis)

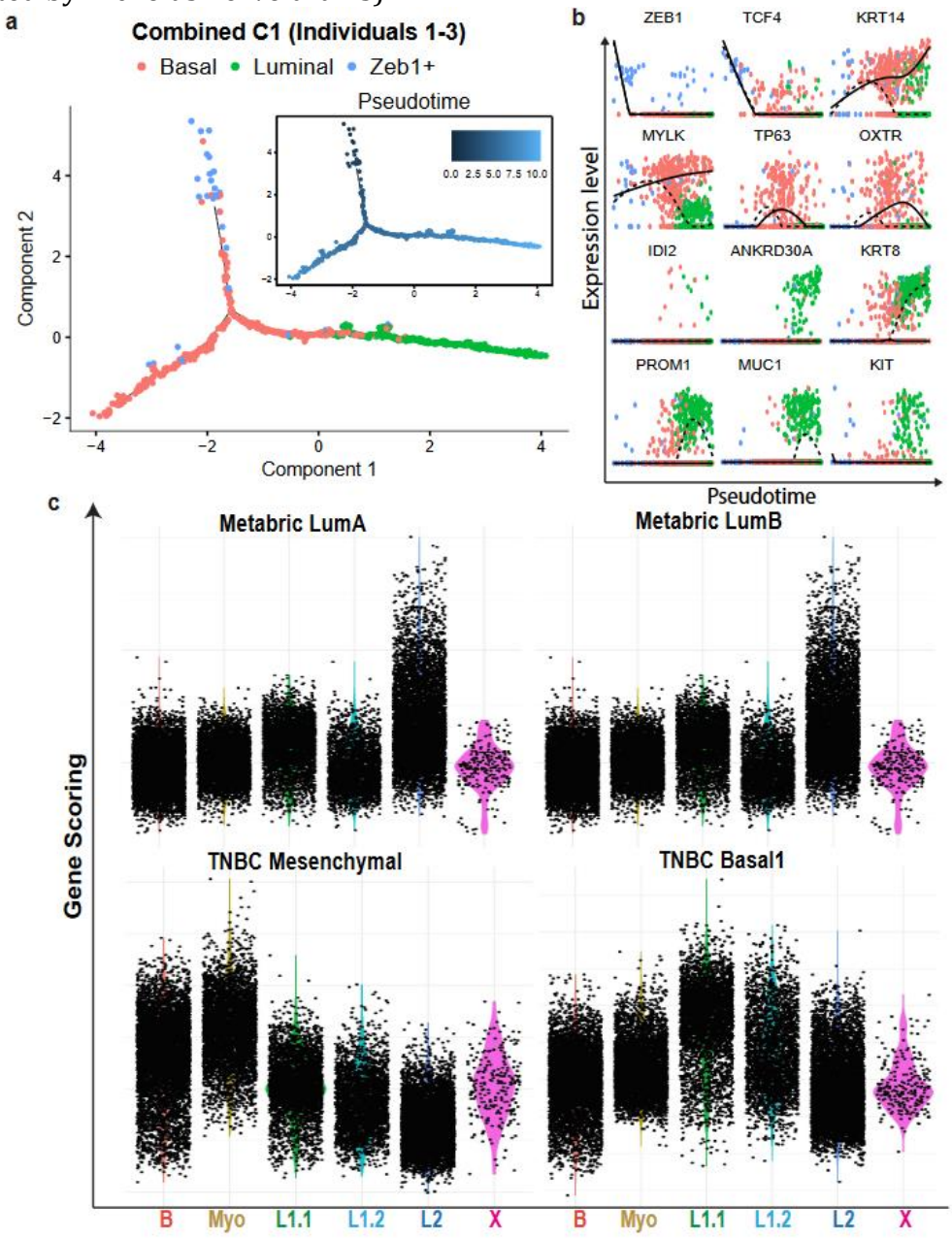
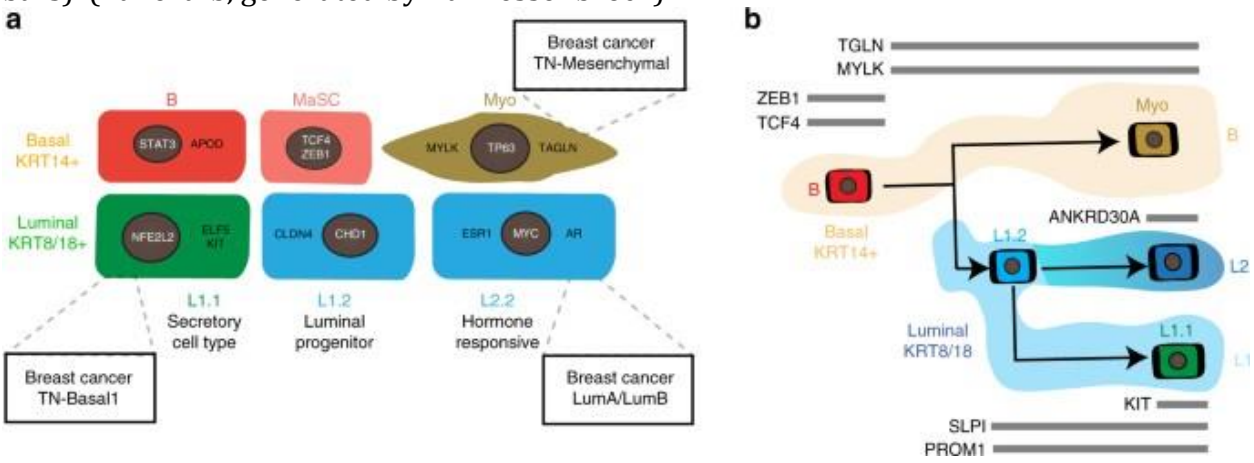


Figure 2.13. Proposed cellular heterogeneity and lineage hierarchies within the human breast. **a** Schematic summary of discovered cell states within the basal and luminal compartment of the human breast epithelium with proposed function, key transcription factors (in white), selected markers (in black), and similarities to breast cancer subtypes indicated in boxes. **b** Proposed model summarizing the lineage hierarchies within the breast epithelium based on one continuous differentiation trajectory from basal stem cells to three distinct differentiated cell types with overlaid marker genes of interest shown (black on gray bars). (Panel a-b, generated by Kai Kessenbrock)



CHAPTER 3: Aberrant changes in breast epithelial homeostasis of BRCA1 mutation carriers revealed by single-cell RNA sequencing

Incorporates components from manuscript:

Aberrant changes in the breast epithelium homeostasis induced by BRCA1 mutation revealed by single-cell RNA sequencing

Quy H. Nguyen, Yanwen Gong, Suoqin Jin, Grace Hernandez, & Kai Kessenbrock.

ABSTRACT

Breast cancer is one of the most prevalent forms of cancer in women worldwide, with especially poor prognosis for the most aggressive basal-like subtype and late-stage cancer diagnosis. Pro-tumorigenic mutations promote destabilization of cellular homeostasis and cancer formation. For example, mutations in the BRCA1 gene drastically increase the probability of developing breast cancer in women, particularly the most aggressive basal-like breast cancer. There is no targeted therapy for basal-like breast cancer and no primary prevention therapy for women with BRCA1 mutations.

INTRODUCTION AND RESULTS

A cohort of BRCA1+/+ reduction mammoplasties and BRCA1+/mut samples from age- and ethnicity-matched, post-pubertal, and pre-menopausal females were collected and processed to viable organoids for cryo-preservation. We performed scRNAseq on dissociated and purified breast epithelial cells isolated from surrounding stromal cells using flow cytometry-based on differential expression of CD49f and EpCAM. Isolated epithelial cells were subjected to scRNAseq by droplet-based sequencing methods (Figure 3.1a). All six individuals sequenced had similar genes and UMI detected in their data set (Figure 3.2a). We performed combined unbiased clustering with all individuals to identify the main cell types within the breast epithelium. Using known epithelial markers, we could annotate clusters based on their expression of KRT14, KRT18, SPLI, and ANKRD30A (Figure 3.2b-d). This analysis identified three major cell types and two additional cell states (Figure 3.1b), as previously described (Nguyen et al., 2018a). To confirm each cluster's identities, we identified genes that were significantly upregulated in each cell state cluster (Figure 3.1c).

We find that similar gene signatures to published markers for each of the five cell states. We detect canonical basal and luminal markers in our clusters. The distribution of BRCA1+/+ and BRCA1+/mut cells were skewed for some of the clusters, namely Luminal 1.1 and Luminal 1.2 (Figure 3.1c, bottom). Cluster Luminal 1.2 has an expansion of BRCA1+/mut cells while the opposite trend is seen in cluster Luminal 1.1. Luminal 1.2 is a potential transitional state that basal cells transition to a more differentiated luminal cell state. Plotting these cells with Monocle and overlaying RNA velocity vectors highlight their differentiation trajectory. With Basal cells as the start of pseudotime, we get a trajectory that differentiates into three main cell types that include: Myo, Luminal 1, and Luminal 2 (Figure 3.1d). Moreover, enrichment of Luminal 1.2 cells fall in the branching point before Luminal 1.1 and Luminal 2. These transitional cells, Luminal 1.2, are enriched for BRCA1+/mut cells compared to BRCA1+/+ cells (Figure 3.1e). These transitional cells have an upregulation of genes previously indicated in progenitor cells, such as ALDH1A3 (Figure 3.1f).

We identified additional genes with aberrant expression, specifically in the BRCA1+/mut cells, compared to the BRCA1+/+ cells. KRT23 expression in BRCA1+/mut cells was overall higher, and we detected expression in Luminal 2 cells while none were detected in the BRCA1+/+ cells (Figure 3.3c). Immunofluorescence analysis of KRT23 reveals that BRCA1+/mut patient samples had foci of higher expression compared to surrounding tissues (approximately 17% of cells had higher expression), while BRCA1+/+ tissue sections had relative consistent expression (approximately 1% had higher expression) (Figure 3.1g, Figure 3.4).

Next, we asked what other aberrant changes we see in BRCA1+/mut cells and what could lead to such changes. We performed single-cell energy calculation on all cells and

overlay their values on the tSNE plots faceted by BRCA1 status (Figure 3.3a). We see higher single-cell energy values for BRCA1+/mut cells; more specifically, we see an accumulation of high energy cells in the Luminal 1.2 cluster. Overall BRCA1+/mut cells have higher single-cell energy as compared to BRCA1+/+ (Figure 3.3b); this is also noticeable on the individual bases (Figure 3.5a-b). The increase in energy between BRCA1+/+ and BRCA1+/mut is very noticeable within the Basal and Luminal 1.2 compartment (Figure 3.5c), possibly explaining the expansion of Luminal 1.2 cells in the BRCA1+/mut carriers.

To characterize further changes induced by BRCA1 mutation, we analyzed the expression of KRT19. KRT19 is an exclusive marker for luminal cells, but we detect the expression of KRT19 in our basal and myoepithelial populations (Figure 3.3d). Immunofluorescence analysis of KRT19 in patient samples reveals expression in the basal layer (Figure 3.5e, Figure 3.6). Interestingly, this provides insight into the deregulation of cellular identities with a BRCA1 mutation, suggesting that BRCA1 mutation leads to dysregulation of cellular identities and limitation, thereby driving the Luminal 1.2 population's expansion.

DISCUSSION

BRCA1 mutation is known to cause triple-negative breast cancer, with little understanding of breast cancer's origin. Here, our results show that BRCA1 mutation carriers have dysregulation in cellular identities and an increase in energy potential leading to an expansion and accumulation of cells in the Luminal 1.2 state. Our work reinforces the idea that basal cells differentiate into the other cell type lineages, including myoepithelial and the

two luminal types. The BRCA1 mutation decreased the cellular lineage specification and increases the stemness potential of these cells, thereby causing pre-neoplastic changes.

METHODS

Acquisition of tissue samples

De-identified reduction mammoplasty and mastectomy samples were acquired from Cooperative Human Tissue Network (CHTN). Samples were washed in PBS (Corning 21-031-CV) and mechanically dissociated using a razor blade. Dissociated samples were digested overnight in DMEM (Corning 10-013-CV) with Collagenase Type I, 2 mg/mL (Life Technologies 17100-017). Viable organoids were separated using differential centrifugation at 250 rcf and viably frozen in 50% FBS (Omega Scientific FB-12), 40% DMEM, and 10% DMSO (Sigma-Aldrich D8418) by volume.

Single-Cell RNA sequencing

Viable organoids were thawed and washed using DMEM with 10% FBS, and digested with 0.05% trypsin (Corning 25-052-CI). Single-cell suspension was treated with 10 kunitz/uL DNase (Sigma Aldrich D4263-5VL) in PBS. Cells were stained for FACS using fluorescently labeled antibodies for CD31 (eBiosciences 48-0319-42), CD45 (eBiosciences 48-9459-42), EpCAM (eBiosciences 50-9326-42), CD49f (eBiosciences 12-0495-82), SytoxBlue (Life Technologies S34857).

Sorted cells were washed and resuspended with 0.04% BSA at a concentration of approximately 1,000 cells/μl. Final cell suspension was counted on the Countess II

Automated Cell Counter (Thermo Fisher AMQAX1000). Library generation for 10x Genomics v2 chemistry were performed following the Chromium Single Cell 3' Reagents Kits v2 User Guide: CG00052 Rev B. Loading of cells on to 10x Genomics Chromium chips were done to target a capture of approximately 10,000 cells per samples.

Quantification of cDNA libraries was performed using Qubit dsDNA HS Assay Kit (Life Technologies Q32851) and high-sensitivity DNA chips (Agilent. 5067-4626). Quantification of final cDNA libraries was performed using KAPA qPCR (Kapa Biosystems KK4824). Sequencing of single-cell libraries was performed on the Illumina HiSeq4000 platform to achieve an average of 50,000 reads per cell.

Processing of scRNAseq data

Sequencing data was demultiplexing and converted to individual samples fastq files for alignment. Alignment of 10x single-cell libraries was completed utilizing 10x Genomics Cell Ranger 1.3.1. Each library was aligned to the indexed human GRCh38 reference genome provided by 10x Genomics using Cell Ranger Count function. Cell Ranger Aggr function was used to normalize the number of confidently mapped reads per cells across all the libraries from different individuals.

Cluster identification using Seurat

The Seurat pipeline (version 2.3.1) was used for dimensionality reduction and clustering of scRNAseq data. In brief, the combined count matrix data was loaded into R (version 3.4) scaled by a size factor of 10,000 and subsequently log transformed. Gene

expression cutoffs were at a minimum of 200 and a maximum of 6000 genes per cell for each dataset. Cells with greater than 20% mitochondrial genes were filtered. Individual epithelial and stromal libraries were analyzed to create cell type labels based on known marker gene expression. Seurat's canonical correlation analysis (CCA), was then used to group cell types from disparate patients after integration of the datasets. For tSNE projection and clustering analysis, we used the first 15 canonical correlation components. Specific markers for each cell type that was annotated were determined using the "FindAllMarkers" function. For the epithelial subset analysis, cell states were clustered as described above and clusters were classified using gene scoring according to the previously described cell states(Nguyen et al., 2018a), namely basal, myoepithelial, L1.1, L1.2, and L2. For gene scoring analysis, we analyzed the gene signature transcriptional prevalence in cell types using Seurat's "AddModuleScore" function. Differential gene expression analysis was performed for each of the cell types, comparing the transcriptome of cells from BRCA1+/+ and BRCA1+/mut cells using the "FindMarkers" function, using the wilcox rank sum test.

Immunofluorescence analysis

Tissues were fixed in 4% formaldehyde overnight and stored in 70% ethanol until processing. Tissues were dehydrated in solutions of increasing concentrations of ethanol, cleared with xylene, and embedded in paraffin. Slides of 5-μm sections were prepared using a Leica SM2010 R Sliding Microtome (Leica Biosystems, Wetzlar, Germany). Slides were heated at 65°C overnight, followed by two 5-min incubations in Histo-Clear (National Diagnostics, Cat. No. HS-200, Atlanta, Georgia, USA) for paraffin removal. Tissues were rehydrated with solutions of decreasing concentrations of ethanol, washed in PBS, and

subjected to antigen retrieval with 10 mM citric acid buffer (0.05% Tween 20, pH 6.0). Tissues were blocked in blocking solution (0.1 % Tween 20 and 10% Goat Serum in PBS) for 1 hr at room temperature, incubated with primary antibodies prepared in blocking solution at 4°C overnight. Slides were then washed in PBS, incubated with secondary antibodies diluted in PBS for 1 hr at room temperature. Slides were mounted with VECTASHIELD Antifade Mounting Medium with DAPI (Vector Laboratories, Cat. No. H-1200, Burlingame, California, USA) and images were taken with the BZ-X700 Keyence fluorescent microscope. For quantification of staining (e.g. KRT23 staining), we manually counted highly positive cells, and all luminal cells, that reside in the luminal population that has a signal around nuclei (DAPI).

Figure 3.1. Identification of aberrant luminal population by scRNAseq. **a**. Overview of scRNAseq approach for collect and processing of human breast tissues into single cells for FACS isolation followed by scRNAseq using 10x Genomics gene expression platform. **b**. Combined tSNE plot of 6 individuals analyzed using Seurat. Major cell types and states are outlined. **c**. Heatmap showing top markers genes from each cell type shows a distinct expression of canonical markers for each. Pie charts show the composition of each cell type. **d**. Monocle plot colored based on cell states and overlaid with RNA velocity projections. **e**. Monocle plot colored based on BRCA1 mutation status with major cell type highlighted. **f**. Gene expression of KRT23 and ALDH1A3 plotted along pseudo time, with basal cells at the beginning and ending in luminal cells. **g**. Immunofluorescence analysis of KRT23 in BRCA1+/+ and BRCA1+/mut patients samples. Expression of KRT23 high cells is quantified. (Panel d-e, generated by Yanwen Gong and Kai Kessenbrock)

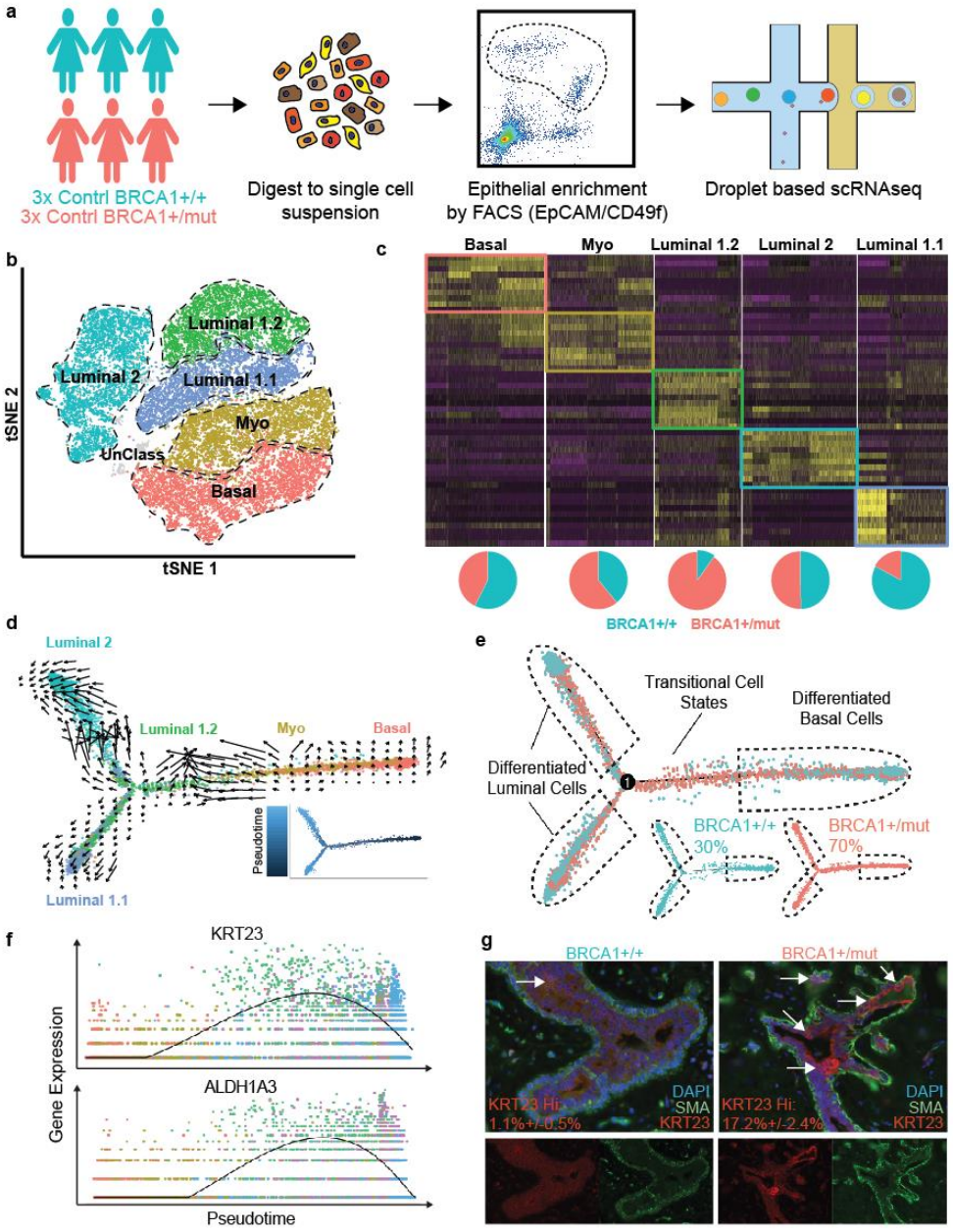


Figure 3.2. Cluster analysis and cell states assignment. **a**. QC metric violin plots for number of genes, numbers of UMI and percent mitochondrial genes. **b**. Feature plots for canonical markers for basal and luminal cell types. **c**. tSNE plot of unbiased clustering before cell states assignment. **d**. Violin plots for scored values based on each cell state compared to reference data.

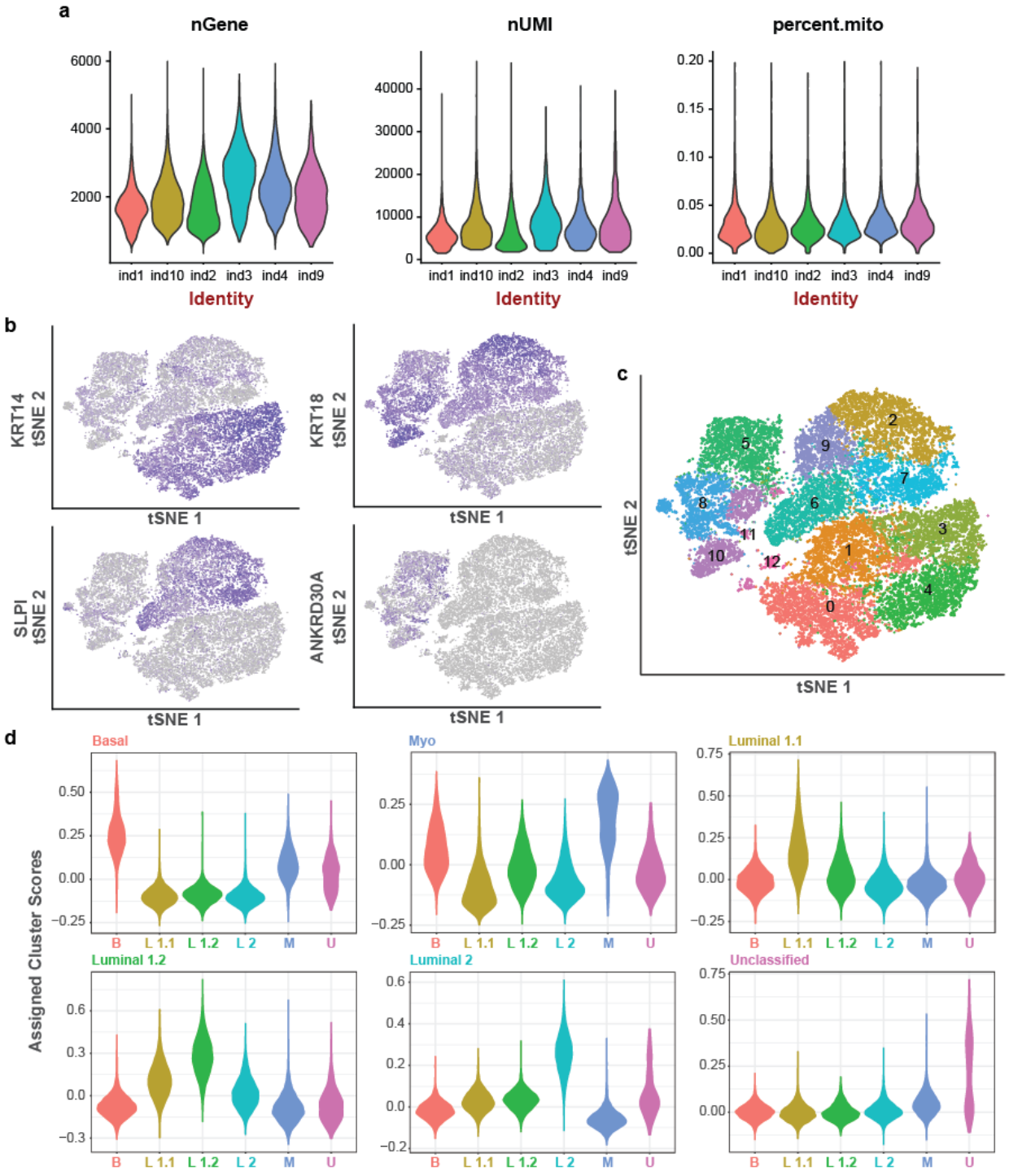


Figure 3.3. Characterization of the luminal progenitor cell state. **a**. tSNE plots with either BRCA1+/+ or BRCA1+/mut cells, colored based on scEnergy calculations. **b**. Overall scEnergy values for either BRCA1+/+ and BRCA1+/mut cells. **c-d**. Violin plots for expression of KRT23 and KRT19, plotted for each cell state. **e**. Immunofluorescence analysis of KRT19 in BRCA1+/+ and BRCA1+/mut patients samples. **f**. Proposed model summarizing the dysregulation in cellular identities in relation to scEnergy. (Panel a-b, generated by Suoqin Jin, panel f, generated by Kai Kessenbrock)

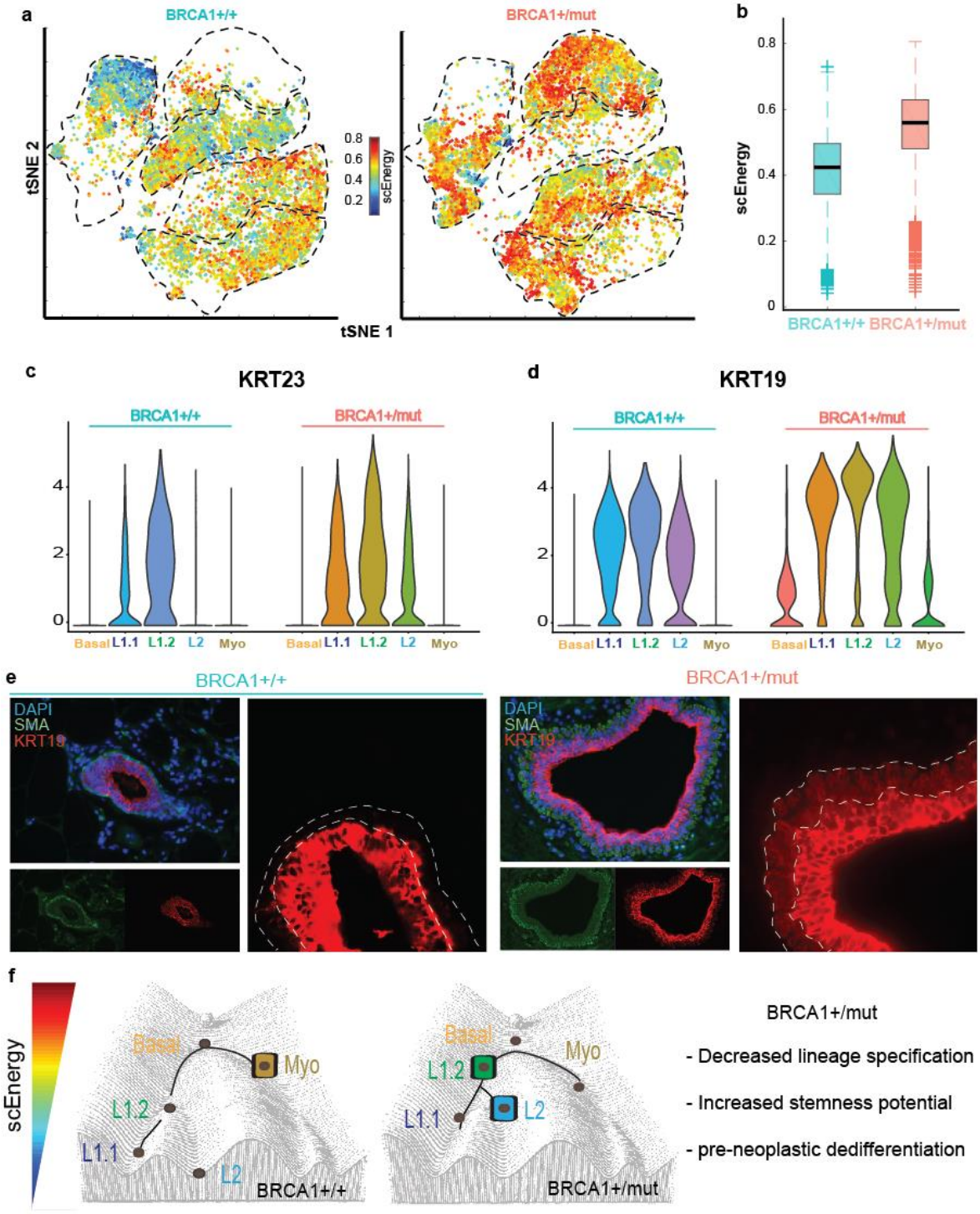


Figure 3.4. Expanded validation of KRT23 expression in BRCA1 individuals. Immunostaining for KRT23 in 4 additional samples, 2 from BRCA1+/+ and 2 from BRCA1+/mut.

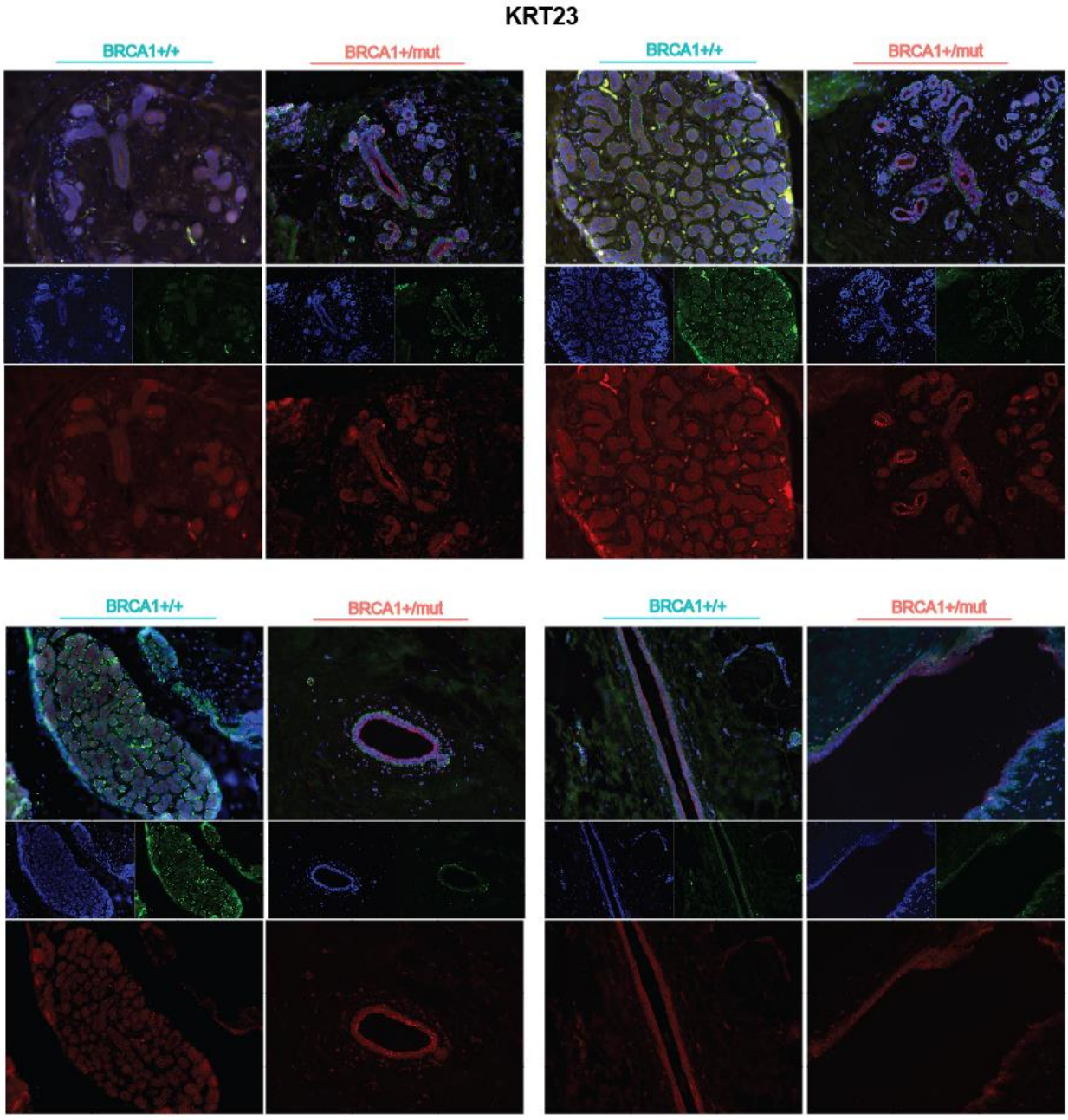


Figure 3.5. scEnergy analysis of cell state and BRCA1 mutation status. **a**. Box plot showing scEnergy value for each individual, separated by BRCA1 mutation status. **b**. Box plots showing scEnergy values each individual across the different cell states. **c**. Box plot showing combined scEnergy value for all individuals in either BRCA1+/+ or BRCA1+/mut for each cell state. **d**. Splice ration plotted for BRCA1+/+ and BRCA1+/mut cells. **e**. Correlation plot for splice ratio between BRCA1+/+ and BRCA1+/mut cells. (Panel a-c, generated by Suoqin Jin, panel d-e, generated by Yanwen Gong)

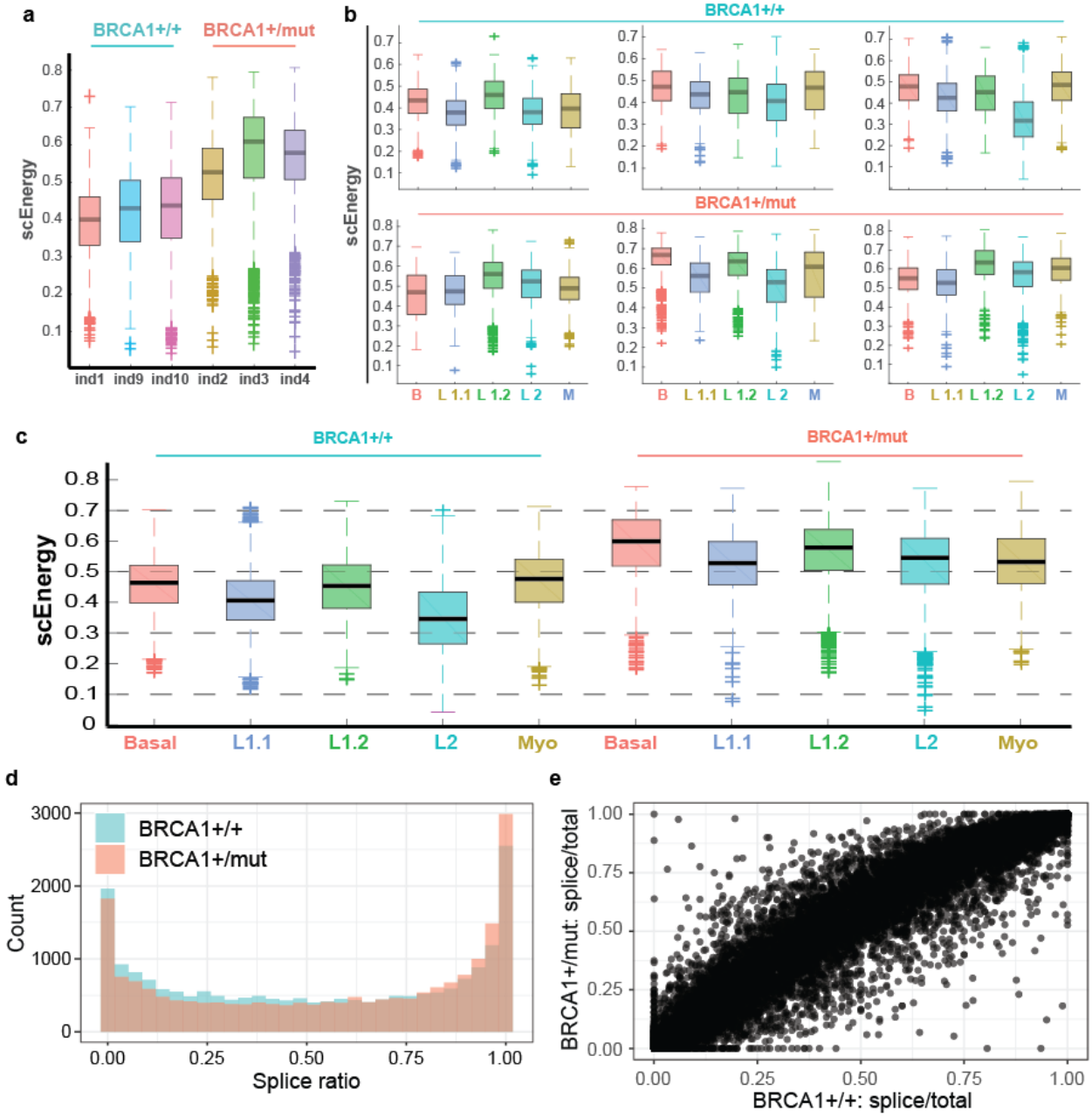
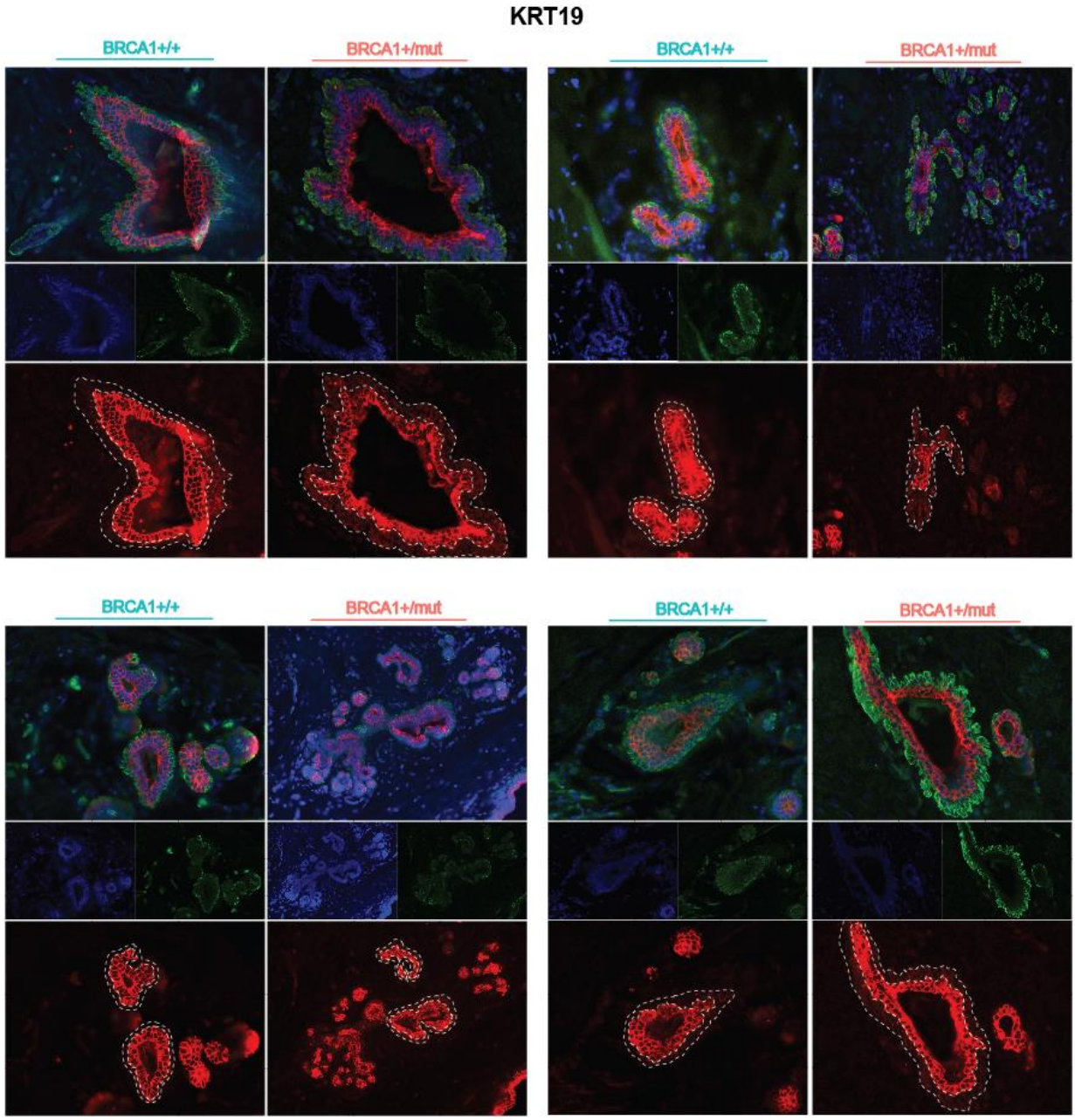


Figure 3.6. Expanded validation of KRT19 expression in BRCA1 individuals. Immunostaining for KRT19 in 4 additional samples, 2 from BRCA1+/+ and 2 from BRCA1+/mut.



CHAPTER 4: Flow cytometry induced changes within mammary epithelial cells revealed by single-cell RNA sequencing

Incorporates components from manuscript:

FACS isolation-induced changes within mammary epithelial cells revealed by single-cell sequencing

Quy H. Nguyen, Grace Hernandez, & Kai Kessenbrock.

INTRODUCTION

Cellular heterogeneity within a population is critical for most tissues to maintain their identity and maintain their cellular functions (Hassiotou and Geddes, 2013; Van Keymeulen et al., 2011). Dysregulation in these pathways can be detrimental to the overall function or identity of the tissues. Understanding how an organ function on a cellular level provides valuable insights for research and biomedical applications. Previous attempts at cellular profiling of tissue systems occur on the bulk level. This work provides an overall view of the cellular processes that take place without the distinction of individual cellular identities. Recent advancement of sequencing technology has allowed for profiling at the single-cell level (Pollen et al., 2014). Single-cell sequencing is a highly sensitive method for profiling cellular heterogeneity and can be applied to most tissue systems. Most current single-cell sequencing technologies require that the users provide a clean single-cell suspension for processing(Nguyen et al., 2018b). Tissue dissociation is one of the first steps in most protocols for single-cell sequencing. It has been shown that certain tissue dissociation protocols will induce changes in gene expression and lead to a false identification of differentially expressed genes (Van Den Brink et al., 2017). Cell isolation typically follows the dissociation of the tissues. This step is critical for obtaining a clean sample that can be used to produce high-quality data. Previous works on the epithelial system have utilized enzymatic dissociation and FACS isolation strategy to generate single-cell data and isolate cell types of interest for functional assays (Nguyen et al., 2018a). These assays involve isolating cell types of interest, typically epithelial cells, for functional characterization. It is important to keep the molecular and cellular identities of these cells from changing to properly evaluate their role in biological processes. Here we show that traditional FACS

isolation of certain cell types will lead to gene expression and other biological changes. These changes affect the cells underlining abilities to proliferate and grow.

RESULTS

FACS isolation is a commonly used method for isolating single cells of interest from a population. This typically involves staining cells with antibodies against markers of interest or dyes that are fluorescence. Then cells are then processed through a FACS machine, where they are fed through tubing at high pressure and exposed to high shear forces and rapid decompression. Towards the end, the detectors pick up which fluorescence markers the cells are stained with and apply a change to the cells so the deflector can sort the cells into the appropriate vessels. Our experiment tested two traditionally available cell cytometers and the MACSQuant Tyto cell sorter (Fig. 1f). We also assessed the sorting capabilities of the MACSQuant Tyto cell sorter. This platform performs the same underlying functions as traditional cell sorter, where they detect fluorescence markers on the cells to isolate the population of interest. There are several noticeable differences in their design. First, the MACSQuant Tyto utilized a sterile cartridge for cell handling. The sterile cartridge ensures a safe environment for operation and sample handling, with the tradeoff of only enriching for one population of interest at a time. This cartridge acts as the fluidics engine for the system, with the benefit of maintaining constant low pressure on the cells, approximately 1.5 psi (Figure 1f). In comparison, traditional cell sorter exerts 20-25 psi (Fig. 1f), and even higher when utilizing a smaller diameter nozzle.

Epithelial heterogeneity in the breast is a field of study that still draws interest, especially due to the still elusive stem or progenitor cell populations in the epithelium.

Recently, new single-cell sequencing approaches have been applied to answer this question, and many functional studies have attempted to indicate whether there is a multipotent stem cell or their lineage-specific progenitor cells. These studies typically involve isolating specific epithelial cells from the tissues while excluding other stromal cells that could confound the results. We applied a similar approach for isolating single epithelial cells from a mouse mammary gland and subjected them to single-cell sequencing and functional assays (Fig. 1a). Mammary gland 4 were pooled from several mice and dissociated to single cells using enzymatic dissociation. To isolate just epithelial cells, we stain with CD49f and EpCAM. We then isolated both basal and luminal cells either with traditional FACS isolation or with MACSQuant Tyto. Both systems generated a pure population of epithelial cells marked by CD49f and EpCAM. A portion of cells from each sort was used to generate a 3' gene expression libraries, while the rest were platted in a mammosphere formation assay (Fig. 2b).

Unbiased clustering of cells from all six experiments identifies three major cell types previously reported, one basal and two luminal clusters (Fig 1b). We detected the expression of canonical markers corresponding to each of these cell types (Fig 1c-d). When comparing each cell type from traditionally sorted cells to cells sorted on the MACSQuant Tyto, canonical cell type markers are similar between the two conditions. This indicates that neither system plays a major role in altering cellular identities in the short term. We next looked into the quality control metrics for these libraries. Cells isolated utilizing the gentler MACSQuant Tyto cell sorter have higher genes and UMI detected (Fig.1e). This indicates that higher pressure and rapid decompression of traditional FACS isolators could negatively affect gene capture and detection when applied for single-cell profiling.

Next, we wanted to identify any signatures that can arise from isolating cells using these two methods. We scored the expression of genes that are known to be involved in stress response pathways. While FACS isolated cells did score higher for these genes, the differences are relatively minor (Fig 2a). This could be because the isolated cells were profiled within 30 minutes of isolation. This relatively short time might not have provided the cells with the opportunity to affect changes in their response to stress following cell isolation.

As previously described, to understand the long terms effect of cell isolation by cytometry, we set up a mammosphere formation assay with cells isolated from both systems. These cells were plated into Matrigel droplet and grown for seven days to assess their proliferation and growth potential (Fig. 2b). The MACSQuant Tyto sorted condition was denser and had more spheres; on average approximately, 1.5x more spheres (Fig. 2c-d).

DISCUSSION

Our results show that single-cell isolation induces changes at the biological level and alters gene capture when applied to single-cell profiling methods. Even though the alteration in the stress response pathway's expression is relatively small for epithelial cells, this effect could be amplified when applied to more sensitive cell types. Epithelial cells are relatively robust and handled extensively and still retain their identities. Such an application that could benefit from this gentler sorter would be nuclei isolation for profiling. It is known that excessive shear force could weaken and even disrupt the nuclear membrane leading to nuclear content loss. The high pressure and forces exerted by traditional FACS isolators

could severely compromise the nuclear membrane leading to lower gene capture when utilized with single-cell profiling methods.

Our finding not only shows that this stress results in lower gene capture, but it also induces long term biological changes that affect cell proliferation and growth. Some transient stress responses typically are upregulated right after induction and subside once the cells have time to recover and grow. Our mammosphere formation assay provides the cells with optimal growth conditions and time to proliferate and expand. Even then, we see differences in sphere formation capabilities between the two sort conditions. This could have implications in the various experimental setup used to identify stem and progenitor cells. Methods for cell isolations vary greatly from FACS purification of cells to no direct cell sorting. These various methods for obtaining cells for transplant assay could alter the cell proliferation and growth potential and skew functional results. Taken together, this work highlight the importance of single-cell isolation and their effect on cellular identities.

METHODS

Acquisition and processing of tissue samples

10 weeks old female FVB/NJ mice are from Jackson Laboratory (Stock Number: 001800) were utilized for these experiments. All experiments have been approved and abide by regulatory guidelines of the International Animal Care and Use Committee (IACUC) of the University of California, Irvine.

Mammary gland 4 were isolated and pooled from these mice for downstream processing. Glands were minced into 1mm pieces and dissociated with 2 mg/mL collagenase

type IV for 1 hr at 37 C while shaking. Digested organoids were collected and further dissociated to single cell using 0.05% trypsin. Cells were stained for CD49f, EpCAM, CD31, CD45, Ter119, and SytoBlue dye.

Single-Cell Isolation

The pool of cells was split into two, one for FACS isolation and one for MACSQuant Tyto sorting. FACS isolated cells were sorted using either the BD Astrios with 100 um nozzle or the BD Aria Fusion with 100 um nozzle. Single epithelial cells from all sorting conditions were isolated based on CD49f and EpCAM.

Single-Cell RNA sequencing

Sorted cells were washed in PBS with 0.04% BSA and resuspended at a concentration of ~1000 cells/µl. Library generation for 10× Genomics v2 chemistry was performed following the Chromium Single Cell 3' Reagents Kits User Guide: Chromium Single Cell 3' Reagents Kits v2 User Guide: CG00052 Rev B.

Quantification of cDNA libraries was performed using Qubit dsDNA HS Assay Kit (Life Technologies Q32851) and high-sensitivity DNA chips (Agilent. 5067-4626). Quantification of library construction was performed using KAPA qPCR (Kapa Biosystems KK4824). For microfluidics-enabled scRNAseq libraries, we generally multiplexed 96 cells per lane on an Illumina HiSeq2500 resulting in a calculated depth of ~1.6 million reads per cell (Illumina Rapid PE kit v2 402-4002 and Rapid SBS kit v2 FC 401-4022). For droplet-enabled scRNAseq, we used the Illumina HiSeq4000 platform to achieve an average of 50,000 reads per cell.

Processing of scRNAseq data

After demultiplexing sequencing libraries to individual cell FASTQ files (observed average read depth per cell was found to be ~1.6 Million reads), each library was aligned to an indexed GRCh38 RefSeq genome using RSEM version 1.2.1242, and bowtie2 version 2.2.3 with the following options enabled: rsem-calculate- expression -p $CORES—bowtie2—paired-end -output- genome-bam. Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values were quantified and concatenated into a resulting gene expression matrix for each library, which was then loaded into R for subsequent computational analysis. For quality control filtering, we generally excluded libraries with less than 900 genes detected. In addition, genes that were not detected in at least 3 of the cells after this trimming were also removed from further analysis. Alignment of 3' end counting libraries from droplet-enabled scRNAseq analyses was completed utilizing 10× Genomics Cell Ranger 1.3.1. Each library was aligned to an indexed GRCh38 genome using Cell Ranger Count. Cell Ranger Aggr function was used to normalize the number of confidently mapped reads per cells across the libraries from different individuals utilizing 10× v2 chemistry.

Cluster identification using Seurat

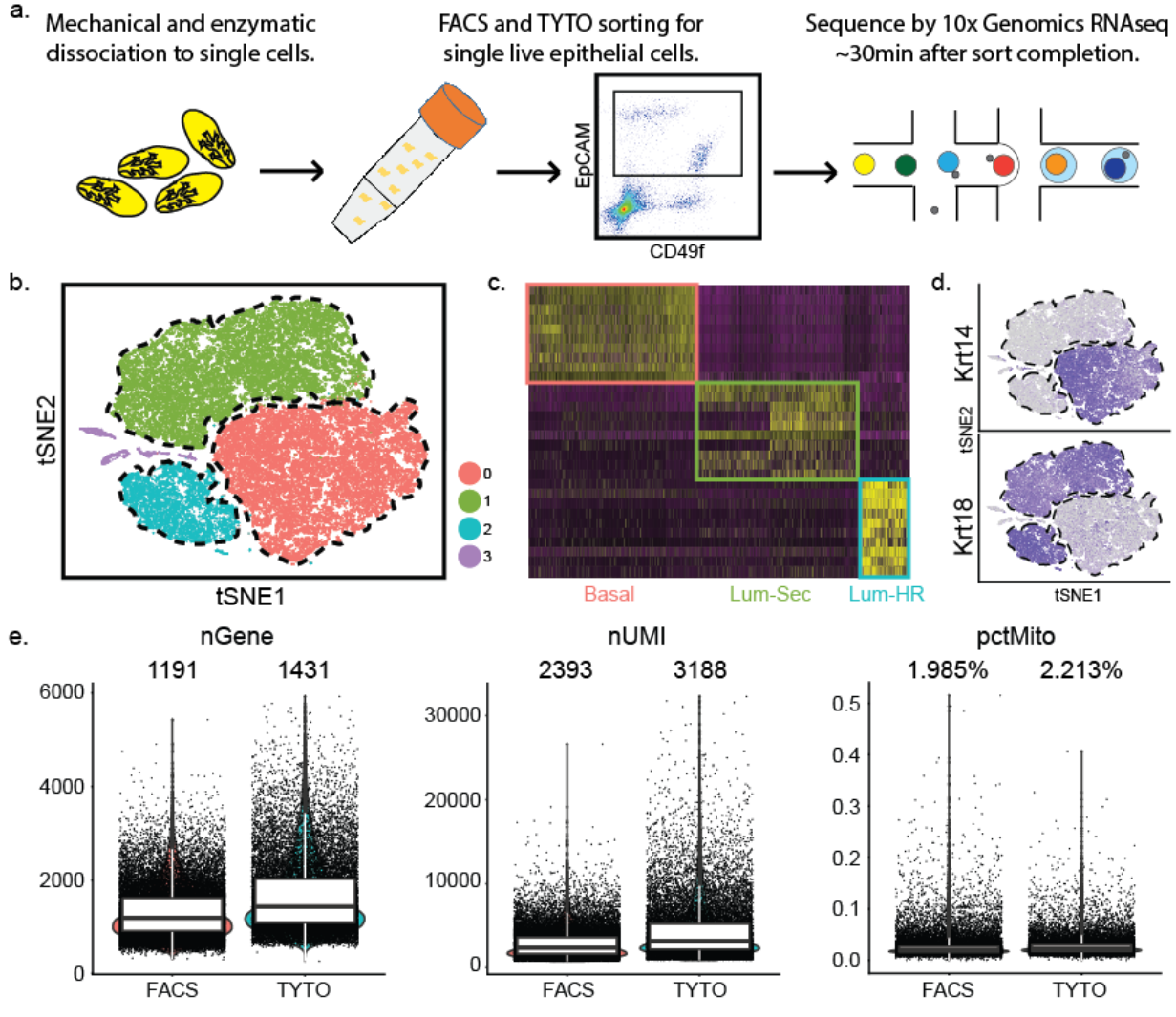
For cluster identification, we utilized the Seurat pipeline. Data was read into R as a counts matrix and transformed into log-space. Due to the difference in gene detection across the two platforms, differences in chemistry for the library prep, as well as sequencing depth per cell, a minimum cutoff of 200 and a maximum cut-off of 6500 genes per cell for this dataset was used. In addition, cells with a percentage of total reads that aligned to the mitochondrial genome (referred to as percent mito) greater than 60% were removed, since

increased detection of mitochondrial genes can be associated with cells undergoing stress and cell death.

Mammosphere assay

Cells acquired after sorting on both platforms were suspended in DMEM/F12 and mixed 1:1 with Matrigel (Corning: 354230). Cells were plated in the center of individual wells on a 12-well cell culture plate. Droplets were allowed to solidify on a heat block at 37C, then incubated in a CO2 incubator for 10 minutes before additional growth media was added. Mammospheres were cultured in a humidified 37C cell culture incubator with 5% CO2 for a total of 7 days. Images were taken using a Keyence BZX710 microscope for analyzing sphere formation.

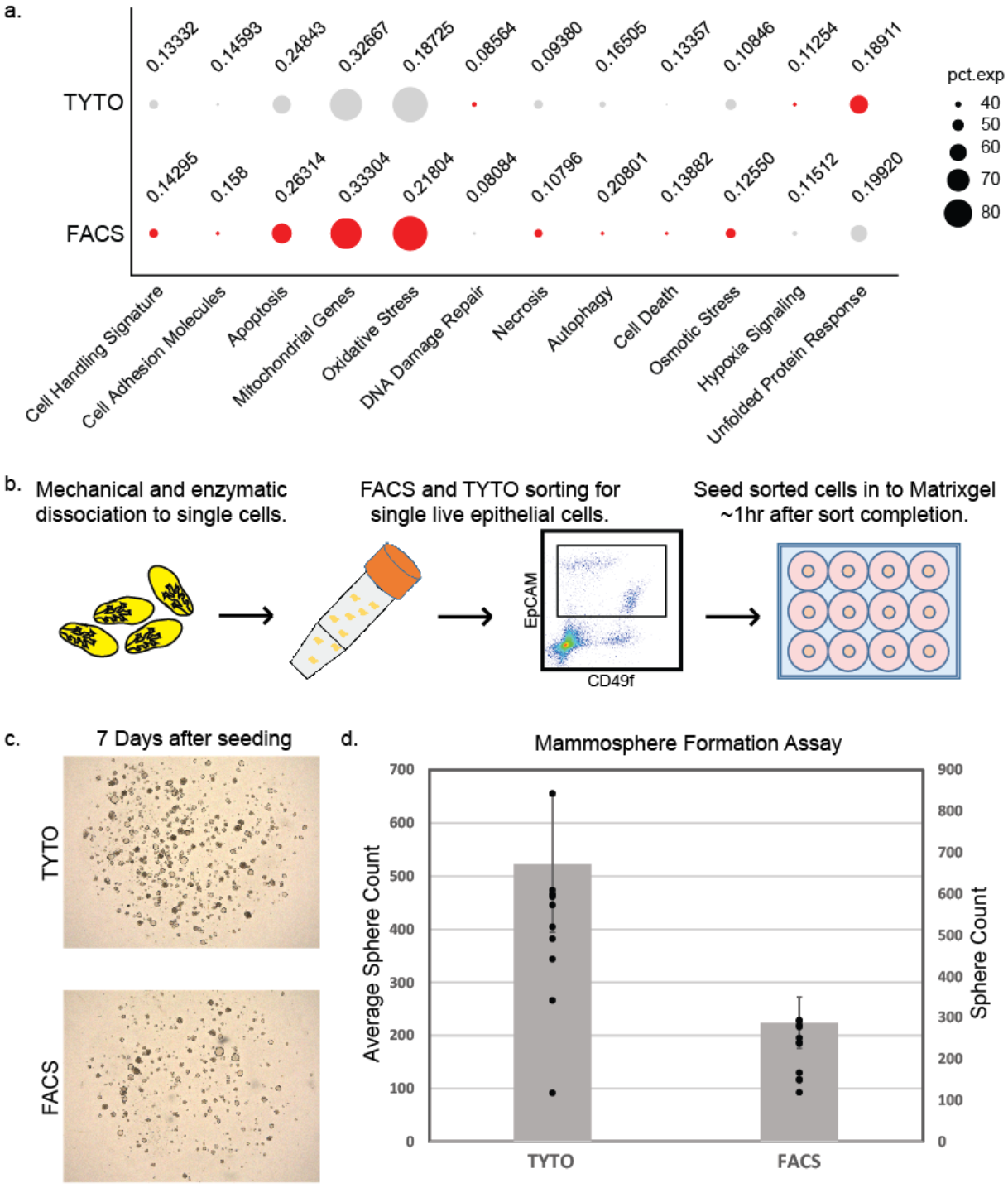
Figure 4.1. Single-cell analysis of isolated epithelial cells reveals protocol induced differences. **a**. Overview for single-cell dissociation from mouse mammary glands and isolation for scRNAseq. **b**. Combined tSNE plot with cell type's outline, basal and two luminal cell types. **c**. Heatmap of top markers genes showing distinct clustering of major cell types. **d**. Feature plot of canonical basal and luminal markers. **e**. Violin plot for QC metrics for FACS and Tyto isolated cells. **f**. Summary table for isolation methods applied.



f.

	BD Astrois	BD Aria Fusion	MB TYTO
Number of Samples	1	2	3
Nozzle Size	100 uM	100 uM	25x50 uM Channel
Pressure	25 psi	20 psi	1.5 psi
Cells Event Rate (cells/sec)	2000	500	2000
Trigger Rate (events/sec)	3200	3000	1000
Target %	5	10-13	10-13
Sort Rate	77	60	200

Figure 4.2. Mammosphere formation assay reveals functional differences in isolated cells. **a**. Split dot plot showing scored values for genes involved in stress response pathway. **b**. Overview of mammosphere formation assay with single cells isolated by FACS or Tyto. **c**. Representative bright-field images of Matrigel culture of isolated cells after 7 days of culture. **d**. Box plot of sphere counted from the assay.



CHAPTER 5: Summary and Conclusions

Overview

This is a remarkable time in scientific discovery and advancement in biomedical technologies. Innovations and developments that occurred in the past several years have led to an exponential growth in our ability to gather new data and increase our understanding of the world around us. This major advancement in biological sciences is driven largely by the development and adoption of big data in omics, such as genomics, transcriptomic, proteomics, metabolomics, and so much more. Our challenge is integrating these different data and interpreting the results to better understand biological processes and interactions that govern our lives. From there, we can understand the mechanism that drives disease progression and advance diagnostic tools and treatment to better human lives. An ongoing grand effort to provide a foundation for scientific and medical advancement is the Human Cell Atlas. This, like the Human Genome Project, is an international collaboration that spans many different disciplines and involves many experts in their fields. Hopefully, this will one day provide the foundations for scientists and medical personnel to better understand disease progression and innovate better treatment.

Breast cancer is consistently listed as the cancer with the most new cases every year. Advancement in early detection and medical procedures has dramatically improved outcomes for those diagnosed in the early stage of the disease, and identification of novel biomarkers and mutations that drive cancer initiation has also been implemented to improve patients' outcomes. It appears that the best outcome for all potential breast cancer patients is early detection and prevention. Utilizing known breast cancer biomarkers, people

can be screened for their risk and take appropriate actions to prevent or manage this disease. A good example of this is the screening for mutations in the BRCA genes. Individuals who have a family history of breast cancer and are carriers of a mutation in their BRCA genes typically opt for a prophylactic mastectomy to protect them from developing breast cancer. This example of early detection and intervention highlights the importance of understanding how the disease initially forms and progress. Although we understand the role of BRCA in suppressing tumor formation, we do not fully understand all the implications of having a mutation in the BRCA genes, nor do we have an effective way of identifying and targeting just the cancer cell of origin. First, we must create a baseline reference for normal breast development to identify any changes that may arise due to genetic variation like mutations in the BRCA genes.

Chapter 2

In chapter 2 of this work, I explore the normal breast epithelium's cellular heterogeneity through the published work; profiling human breast epithelial cells using single-cell RNA sequencing identifies cell diversity, Nature Communication 2018. I applied single-cell gene expression sequencing methods to profile over 25,000 epithelial cells from 7 individuals. This lead to the identification of three major cell types, one basal and two luminal populations. When projecting these major cell types on a monocle pseudotime trajectory, we revealed a continuous hierarchy that connects the luminal lineage to the basal cells implicating their origin.

Many previous works have described the breast at the anatomical and cellular level. In short, the breast is composed of an epithelial ductal network that grows into adipose rich

tissue. Identification of cellular heterogeneity within the breast epithelium previously relied on the separation of cell types based on cell surface markers and typically involves profiling on the bulk level. While these efforts have greatly advanced our understating of the epithelial system of the breast, we are still left with just an overview. High-resolution profiling is required to identify true heterogeneity on the single-cell level. This work provides a high resolution of insight and allows us to identify and profile three major cell types in the breast epithelium. Studying the two luminal cell types as separate population provides insight into their role in the breast. The luminal 1 population is the more secretory of the two, and the luminal 2 population has a higher level of hormone receptors, indicating that they may be more responsive to hormones during changes. Reconstructing the differentiation hierarchy using pseudotime analysis resulted in a projection of basal cells transitioning into the luminal lineage. While this pseudotime projection is created from the transcriptional profile of representative cells taken at one point in time, it complements the notion that there may be stem cells that reside in the basal population. There may also be a lineage-specific progenitor that is responsible for the upkeep of each of the two cell types. The luminal 1 population is believed to contain progenitor cells that help maintain the luminal lineage; this population falls earlier on the pseudotime projection. This concept has been eluded to by various work that characterizes a population referred to as the luminal progenitor cells (Lim et al., 2009). This work provides the foundation for further isolation and characterization of these major cell types. It is a resource that provides a profile of the composition of the normal breast epithelium during homeostasis. This resource can be used to identify changes that occur during cancer initiation or any abnormal changes. Although this work provides a profile of the normal breast epithelium, we do not know whether there are multipotent stem

cells responsible for maintaining the whole epithelial ductal network, or that role is filled by lineage-specific progenitor cells. Since breast cancer is such a prevalent disease, interrogating this work to identifying the potential cancer cell of origin reveals the three cell types correspond closely to various breast cancer subtypes. This work supports the notion that breast cancer arises from the breast epithelium but does not delineate the specific cell types responsible for specific breast cancer subtypes.

Chapter 3

Chapter 3 builds on the normal epithelial cell atlas by identifying changes that arise during early cancer initiation. In the context of BRCA1 mutations carrier, we identified dysregulation in the cellular identities in epithelial cells. This dysregulation leads to an expansion of the luminal progenitor population, luminal 1.2. Higher scEnergy in these cell states, especially in the BRCA1 mutation carrier cells, supports the notion that these cells are more deregulated and have a higher potential to lose their cellular control and proliferate uncontrollably. Staining for markers associated with stem and progenitor cells reveals an increased expression in this luminal progenitor population.

Histological and molecular profiling efforts have generated a wealth of knowledge for the various breast cancer subtypes. There are five main subtypes, Luminal A, Luminal B, Triple-Negative, HER2-positive, and Normal-like. Those studies provide valuable insight into the pathological characteristics of these cancer subtypes. We now understand how this cancer progress, and how each subtype response to cancer therapies. Although this knowledge is useful in the diagnosis and treatment of cancer patients, a deeper cellular understanding is required to improve patients' care and overall survival.

Previous work has made great strides in correlating mutations in the BRCA genes with breast cancer development. It is known that up to 70% of women who carry a mutation in their BRCA1 gene will develop breast cancer at some point (Kerr and Ashworth, 2001; Orban and Olah, 2003). This information is currently used in the health care system to help individuals at higher risk prevent and manage disease onset. Most individuals who are BRCA1 mutation carriers and develop breast cancer develops a subtype known as TNBC (Stevens et al., 2013). Previous work analyzing breast cancer cases has shown a high correlation between BRCA1 mutation and TNBC with a larger tumor burden. This work, in conjunction with studies into BRCA1 role in DNA damage response and tumor-suppressing mechanism, highlights the importance of identifying the cellular response during development when BRCA1 is mutated. This work identifies an expansion of the luminal progenitor population in individuals with a mutation in their BRCA1 gene. Our previous work correlates luminal cells' expression pattern to best match those of TNBC/Basal-like breast cancer. This work supports the notion that BRCA1 mutations disrupt cellular homeostasis and lead to a massive expansion of the luminal progenitor cells. If left unchecked, these luminal progenitor cells could ultimately lead to cancerous growth. Dysregulation in cellular identities is also evident in their expression of canonical basal and luminal markers. This idea supports the concept that with the loss of BRCA1 functions, the cells accrue mutations that eventually disrupt their normal homeostasis and cellular identities. This dysregulation eventually leads to a population that no longer responds to regulating cellular signals.

This work provides a complete profile of normal breast epithelial cells and corresponding epithelial cells from individuals with mutations in the BRCA1 gene. It is a

valuable resource for identifying changes that arise before cancer progression. The role of BRCA1 mutations in other cell types remains unclear. BRCA1 mutation does disrupt other cell type homeostasis, but its role in driving cancer initiation and progression through cellular interaction remains elusive. The basal cells and other stromal components play an important role in maintaining and supporting the luminal cells during normal homeostasis (Dong-Le Bourhis et al., 1997; van Roozendaal et al., 1992); disruption in this communication and support with BRCA1 mutation could be a driving factor in BRCA1 mediated breast cancer.

Chapter 4

Chapter 4 highlights the importance of proper cell handing for profiling or functional assays. It is well known that various cell types respond to cellular stress differently. Proper handling and processing of samples will ensure high-quality data and reproducible results. We find that commonly used FACS isolation methods could alter cellular contents and have lasting biological changes.

FACS isolation is a widely used method used by many for purifying cells of interest. This process involves labeling the cells with antibodies or dyes so the machine can identify the various markers by fluorescence. FACS isolation utilizes the concept of single-cell separation and profiling. Traditional FACS isolation systems require high pressure and rapid decompression to achieve high speed and high-efficiency sorting cells. We see that cells subjected to those conditions yielded lower gene capture when profiled by single-cell gene expression sequencing. In comparison, cells that are sorter with a constant lower pressure have higher gene recovery after. Previous work has shown that dissociation protocol used

to break down tissues can induce changes to the gene expression profile and upregulate the stress response gene (Van Den Brink et al., 2017). This step is where isolation protocols need to be optimized to reduce artificial changes during sample handling. This work highlights the importance of a proper understanding of the changes that this step can induce. It is critical to consider this when handling more sensitive cell types such as nuclei. The long-term changes are seen by the mammosphere formation assay highlight an important point about how FACS isolation could affect experimental results, especially those involving isolating stem and progenitor cells for transplant. The changes induced by FACS isolation on cell type beyond mouse mammary epithelial cells are still unclear. Every cell types handle and respond to this stress differently.

Future directions and conclusions

Although we created an atlas to investigate the heterogeneity within the normal breast epithelium, we still need to integrate any findings into the breast cellular environment as a whole. Single-cell RNA sequencing used in this work is extremely powerful and can tell us what populations are in the epithelium but cannot tell us their spatial localization. The microenvironment is essential in supporting this structure, and the cell to cell interactions helps maintain homeostasis. Understanding the spatial localization of these cells and their neighbors would provide us more insight into the role of stromal components and other unknown sub-states. This is especially true for spatially identifying the luminal progenitor cells in BRCA1+/mut samples; this would help us with additional insight into possible interactions with other cell types that could promote differentiation and growth. In a clinical context, this could generate novel histological markers that can

then be used to identify patients who are at high risk of cancer formation. As eluded to, the microenvironment of any system is critical to its upkeep. This work investigates cellular heterogeneity within the breast epithelium. Other stromal components such as fibroblast, pericyte, adipocyte, immune, and endothelial cells, interacts with the epithelial system daily and play a big role in everything from maintaining homeostasis to promoting normal growth and cancer initiation. Understanding this microenvironment signaling would provide a complete picture of normal breast development and cancer initiation.

To complement our knowledge of what cells are there and their localization, we also need a complete understanding of their biological roles. When looking at these cell states, we need to go beyond just knowing that luminal cells produce milk and basal cells surround them. We currently lack an understanding of the role each sub-state plays in the system as a whole. Understanding this would provide better diagnostic tools for the detection of cancer and treatments. We see an expansion of the luminal progenitor population in the BRCA1+/mut samples, but we also need to know the origin of this cell state and what causes this accumulation.

There are many ongoing efforts to delineate a complete understanding of the normal breast epithelium across a large cohort of individuals. This work provides that first step to understanding the complete spectrum of heterogeneity and provides a first attempt to understanding the changes that arise during early cancer progression. This process is critical for identifying novel biomarkers and therapeutics to improve patients' overall survival.

REFERENCES:

Ali, H.R. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biol. *15*.

Anders, C., and Carey, L.A. (2008). Understanding and treating triple-negative breast cancer. Oncology (Williston Park). *22*, 1233-9; discussion 1239-40, 1243.

Bach, K., Pensa, S., Grzelak, M., Hadfield, J., Adams, D.J., Marioni, J.C., and Khaled, W.T. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. Nat. Commun. *8*.

Baldan, V., Griffiths, R., Hawkins, R.E., and Gilham, D.E. (2015). Efficient and reproducible generation of tumour-infiltrating lymphocytes for renal cell carcinoma. Br. J. Cancer *112*, 1510–1518.

Bianchini, G., Balko, J.M., Mayer, I.A., Sanders, M.E., and Gianni, L. (2016). Triple-negative breast cancer: Challenges and opportunities of a heterogeneous disease. Nat. Rev. Clin. Oncol. *13*, 674–690.

Børresen-Dale, A.-L., Geisler, S., Demeter, J., Botstein, D., Tibshirani, R., Hastie, T., Lønning, P.E., Deng, S., Nobel, A., Parker, J., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. *100*, 8418–8423.

Van Den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and Van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods *14*, 935–936.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Cancer, B., Service, M., and Nazionale, I. (2011). Triple-Negative Breast Cancer: An Unmet Medical Need. Oncologist *16*, 1–11.

Casey, A.E., Sinha, A., Singhania, R., Livingstone, J., Waterhouse, P., Tharmapalan, P., Cruickshank, J., Shehata, M., Drysdale, E., Fang, H., et al. (2018). Mammary molecular portraits reveal lineage-specific features and progenitor cell vulnerabilities. J. Cell Biol. *217*, 2951–2974.

Choudhury, S. (2013). Molecular profiling of human mammary gland links breast cancer risk to a p27(+) cell population with progenitor characteristics. Cell Stem Cell *13*.

Christaki, E., Opal, S.M., Keith, J.C., Kessimian, N., Palardy, J.E., Parejo, N.A., Tan, X.Y., Piche-Nicholas, N., Tchistiakova, L., Vlasuk, G.P., et al. (2011). A monoclonal antibody against RAGE alters gene expression and is protective in experimental models of sepsis and pneumococcal pneumonia. Shock *35*, 492–498.

Curtis, C. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*.

Davis, F.M., Lloyd-Lewis, B., Harris, O.B., Kozar, S., Winton, D.J., Muresan, L., and Watson, C.J. (2016). Single-cell lineage tracing in the mammary gland reveals stochastic clonal dispersion of stem/progenitor cell progeny. Nat. Commun. *7*, 13053.

Dimri, G., Band, H., and Band, V. (2005). Mammary epithelial cell transformation: Insights from cell culture and mouse models. Breast Cancer Res. *7*, 171–179.

Dong-Le Bourhis, X., Berthois, Y., Millot, G., Degeorges, A., Sylvi, M., Martin, P.M., and Calvo, F. (1997). Effect of stromal and epithelial cells derived from normal and tumorous breast tissue on the proliferation of human breast cancer cell lines in co-culture. Int. J. Cancer *71*, 42–48.

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

Eroles, P., Bosch, A., Alejandro Pérez-Fidalgo, J., and Lluch, A. (2012). Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. Cancer Treat. Rev. *38*, 698–707.

Ferguson, J.E., Schor, A.M., Howell, A., and Ferguson, M.W. (1992). Changes in the extracellular matrix of the normal human breast during the menstrual cycle. Cell Tissue Res. *268*, 167–177.

Filipczyk, A. (2015). Network plasticity of pluripotency transcription factors in embryonic stem cells. Nat. Cell. Biol. *17*.

Foulkes, W.D., Smith, I.E., and Reis-Filho, J.S. (2010). Triple-Negative Breast Cancer. N. Engl. J. Med. *363*, 1938–1948.

Fridriksdottir, A.J., Villadsen, R., Gudjonsson, T., and Petersen, O.W. (2005). Maintenance of cell type diversification in the human breast. J. Mammary Gland Biol. Neoplasia *10*, 61–74.

Fuchs, E., and Nowak, J.A. (2008). Building epithelial tissues from skin stem cells. Cold Spring Harb. Symp. Quant. Biol. *73*.

Glukhova, M., Koteliansky, V., Sastre, X., and Thiery, J.P. (1995). Adhesion systems in normal breast and in invasive breast carcinoma. Am. J. Pathol. *146*, 706–716.

Gudjonsson, T., Adriance, M.C., Sternlicht, M.D., Petersen, O.W., and Bissell, M.J. (2005). Myoepithelial cells: their origin and function in breast morphogenesis and neoplasia. J. Mammary Gland. Biol. Neoplasia. *10*.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091–1097.e17.

Harbeck, N., Gottschalk, N., Nitz, U., Liedtke, C., Gluz, O., and Pusztai, L. (2009). Triple-negative breast cancer--current status and future directions. Ann. Oncol. *20*, 1913–1927.

Harris, J. (2006). Socs2 and elf5 mediate prolactin-induced mammary gland development. Mol. Endocrinol. *20*.

Hassiotou, F., and Geddes, D. (2013). Anatomy of the human mammary gland: Current status of knowledge. Clin. Anat. *26*, 29–48.

Heath, J.R., Ribas, A., and Mischel, P.S. (2016). Single-cell analysis tools for drug discovery and development. Nat. Rev. Drug Discov. *15*, 204–216.

Hedenfalk, I.A. (2006). Gene expression profiling can distinguish tumor subclasses of breast carcinomas. Gene Expr. Profiling by Microarrays Clin. Implic. *98*, 132–161.

Hens, J.R., and Wysolmerski, J.J. (2005). Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland. Breast Cancer Res. *7*, 220–224.

Hinck, L., and Näthke, I. (2014). Changes in cell and tissue organization in cancer of the breast and colon. Curr. Opin. Cell Biol. *26*, 87–95.

Howard, B.A., and Gusterson, B.A. (2000). Human breast development. J. Mammary Gland Biol. Neoplasia *5*, 119–137.

Hu, G., Li, L., and Xu, W. (2017). Frontiers in Laboratory Medicine Extracellular matrix in mammary gland development and breast cancer progression. Front. Lab. Med. *1*, 36–39.

Huh, S.J. (2016). The proliferative activity of mammary epithelial cells in normal tissue predicts breast cancer risk in premenopausal women. Cancer Res. *76*.

Ilicic, T. (2016). Classification of low quality cells from single-cell RNA-seq data. Genome Biol. *17*.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Jena, M.K., and Janjanam, J. (2018). Role of extracellular matrix in breast cancer development: a brief update. F1000Research *7*, 274.

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.

Kaas, R., Verhoef, S., Wesseling, J., Rookus, M.A., Oldenburg, H.S.A., Peeters, M.J.V., and Rutgers, E.J.T. (2010). Prophylactic mastectomy in BRCA1 and BRCA2 mutation carriers: Very low risk for subsequent breast cancer. Ann. Surg. *251*, 488–492.

Kerr, P., and Ashworth, A. (2001). New complexities for BRCA1 and BRCA2. Curr. Biol. *11*, 668–676.

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.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kornbluth, J., and Hoover, R.G. (1989). Anti-HLA Class I Antibodies Alter Gene Expression in Human Natural Killer Cells. In Immunobiology of HLA, pp. 150–152.

Kuleshov, M. V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set

enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.

Lawson, D.A., Bhakta, N.R., Kessenbrock, K., Prummel, K.D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C.Y., et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature *526*, 131–135.

Lehmann, B.D. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. *121*.

Lescroart, F., Wang, X., Lin, X., Swedlund, B., Gargouri, S., Sànchez-Dànes, A., Moignard, V., Dubois, C., Paulissen, C., Kinston, S., et al. (2018). Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. Science (80-.). *359*, 1177–1181.

Lim, E. (2010). Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast Cancer Res. *12*.

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. Nat. Med. *15*, 907–913.

Love, S.M., and Barsky, S.H. (2004). Anatomy of the nipple and breast ducts revisited. Cancer *101*, 1947–1957.

Macias, H., and Hinck, L. (2012). Mammary gland development. Wiley Interdiscip. Rev. Dev. Biol. *1*, 533–557.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Margan, M.M., Jitariu, A.A., Cimpean, A.M., Nica, C., and Raica, M. (2016). Molecular Portrait

of the Normal Human Breast Tissue and Its Influence on Breast Carcinogenesis. J. Breast Cancer *19*, 99–111.

Meeson, A., Fuller, A., Breault, D.T., Owens, W.A., and Richardson, G.D. (2013). Optimised Protocols for the Identification of the Murine Cardiac Side Population. Stem Cell Rev. Reports *9*, 731–739.

Morel, A.P. (2017). A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability. Nat. Med. *23*.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–95.

Nguyen, Q.H., Pervolarakis, N., Blake, K., Ma, D., Davis, R.T., James, N., Phung, A.T., Willey, E., Kumar, R., Jabart, E., et al. (2018a). Profiling human breast epithelial cells using singlecell RNA sequencing identifies cell diversity. Nat. Commun. *9*, 1–12.

Nguyen, Q.H., Pervolarakis, N., Nee, K., and Kessenbrock, K. (2018b). Experimental Considerations for Single-Cell RNA Sequencing Approaches. *6*, 1–7.

Orban, T.I., and Olah, E. (2003). Emerging roles of BRCA1 alternative splicing. J. Clin. Pathol. - Mol. Pathol. *56*, 191–197.

Pal, B., Chen, Y., Vaillant, F., Jamieson, P., Gordon, L., Rios, A.C., Wilcox, S., Fu, N., Liu, K.H., Jackling, F.C., et al. (2017). Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. Nat. Commun. *8*.

Pandya, S., and Moore, R.G. (2011). Breast development and anatomy. Clin. Obstet. Gynecol. *54*, 91–95.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P.,

Nahed, B. V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (80-.). *344*, 8–13.

Pergamenschikov, A., Johnsen, H., van de Rijn, M., Zhu, S.X., Sørlie, T., Børresen-Dale, A.-L., Williams, C., Jeffrey, S.S., Botstein, D., Rees, C.A., et al. (2002). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Perou, C.M., Sørile, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., Ress, 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.

Petersen, O.W., and Polyak, K. (2010). Stem cells in the human breast. Cold Spring Harb. Perspect. Biol. *2*, a003160.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. *32*, 1053–1058.

Qiu, X., De Jesus, J., Pennell, M., Troiani, M., and Haun, J.B. (2014). Microfluidic device for mechanical dissociation of cancer cell aggregates into single cells. Lab Chip *15*, 339–350.

Radbruch, A., and Recktenwald, D. (1995). Detection and isolation of rare cells. Curr. Opin. Immunol. *7*, 270–273.

Ramsay, D.T., Kent, J.C., Hartmann, R.A., and Hartmann, P.E. (2005). Anatomy of the lactating human breast redefined with ultrasound imaging. J. Anat. *206*, 525–534.

Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I.,

Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. *30*, 777–782.

Robin, Y.M. (2013). Transgelin is a novel marker of smooth muscle differentiation that improves diagnostic accuracy of leiomyosarcomas: a comparative immunohistochemical reappraisal of myogenic markers in 900 soft tissue tumors. Mod. Pathol. *26*.

Robinson, G.W., Karpf, A.B., and Kratochwil, K. (1999). Regulation of mammary gland development by tissue interaction. J. Mammary Gland Biol. Neoplasia *4*, 9–19.

Romero-santacreu, L., Moreno, J., Perez-Ortin, J.E., and Alepuz, P. (2009). Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. Rna 1110–1120.

Rønnov-Jessen, L., Petersen, O.W., and Bissell, M.J. (1996). Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. Physiol. Rev. *76*, 69–125.

van Roozendaal, C.E., van Ooijen, B., Klijn, J.G., Claassen, C., Eggermont, A.M., Henzen-Logmans, S.C., and Foekens, J.A. (1992). Stromal influences on breast cancer cell growth. Br. J. Cancer *65*, 77–81.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017a). The Human Cell Atlas: From vision to reality. Nature *550*, 451–453.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017b). The human cell atlas: from vision to reality. Nature *550*.

Schedin, P., Mitrenga, T., McDaniel, S., and Kaeck, M. (2004). Mammary ECM composition and function are altered by reproductive state. Mol. Carcinog. *41*, 207–220.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S.,

Menzel, W., Granzow, M., and Ragg, T. (2006). The RIN: An RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. *7*, 1–14.

Seil, I. (2007). The differentiation antigen NY-BR-1 is a potential target for antibody-based therapies in breast cancer. Int. J. Cancer *120*.

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.

Shehata, M., Teschendorff, A., Sharp, G., Novcic, N., Russell, I.A., Avril, S., Prater, M., Eirew, P., Caldas, C., and Watson, C.J. (2012). Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. Breast Cancer Res. *14*.

Spike, B.T. (2012). A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. Cell Stem Cell *10*.

Sternlicht, M.D. (2006). Key stages in mammary gland development: the cues that regulate ductal branching morphogenesis. Breast Cancer Res. *8*, 201.

Stevens, K.N., Vachon, C.M., and Couch, F.J. (2013). Genetic susceptibility to triple-negative breast cancer. Cancer Res. *73*, 2025–2030.

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 Res. Treat. *67*.

Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell rnA-sequencing experiments. Nat. Methods *14*, 381–387.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to

mechanism. Nature *541*, 331–338.

Ting, D.T. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep. *8*.

Tirosh, I., Izar, B., Prakadan, S.M., Ii, M.H.W., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular exosystem of metastatic melanoma by single-cell RNA-seq. Science (80-.). *352*, 189–196.

Tobon, H., and Salazar, H. (1974). Ultrastructure of the human mammary gland. I. Development of the fetal gland throughout gestation. J. Clin. Endocrinol. Metab. *39*, 443–456.

Trapnell, C. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 371–375.

Tung, P.Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. Sci. Rep. *7*, 1–15.

Visvader, J.E., and Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: Current status and perspectives. Genes Dev. *28*, 1143–1158.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat. Biotechnol. *34*, 1145–1160.

Wang, D. (2015). Identification of multipotent mammary stem cells by protein C receptor expression. Nature *517*.

Will, B., and Steidl, U. (2010). Multi-parameter fluorescence-activated cell sorting and analysis of stem and progenitor cells in myeloid malignancies. Best Pract. Res. Clin. Haematol. *23*, 391–401.

Woodward, W.A. (2005). On mammary stem cells. J. Cell Sci. *118*, 3585–3594.

Xiong, L., Lee, H., Ishitani, M., and Zhu, J.K. (2002). Regulation of osmotic stress-responsive gene expression by the LOS6/ABA1 locus in Arabidopsis. J. Biol. Chem. *277*, 8588–8596.

Yang, X., Wang, H., and Jiao, B. (2017). Mammary gland stem cells and their application in breast cancer. Oncotarget *8*, 10675–10691.

Ye, X. (2015). Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. Nature *525*.

Yuan, G.C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., et al. (2017). Challenges and emerging directions in single-cell analysis. Genome Biol. *18*, 1–8.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 1–12.

Zhou, J., Chen, Q., Zou, Y., Zheng, S., and Chen, Y. (2019). Stem Cells and Cellular Origins of Mammary Gland: Updates in Rationale, Controversies, and Cancer Relevance. Stem Cells Int. *2019*, 4247168.