UNIVERSITY OF CALIFORNIA,
IRVINE

Single cell sequencing analysis reveals mammary epithelial cell diversity and regulation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational, and Systems Biology

by

Nicholas Pervolarakis

Dissertation Committee:
Assistant Professor Kai Kessenbrock, Chair
Professor Ali Mortazavi
Professor Xing Dai
Professor Anand Ganesan
Assistant Professor Vivek Swarup

2021

DEDICATION

To

my parents,

for their constant support throughout my life,

my brother,

for always being there for me

You can never know everything, and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyway.
.

(Robert Jordan,
Winter's Heart)

There is no struggle too vast, no odds too overwhelming, for even should we fail - should we fall - we will know that we have lived.

(Steven Erikson,
Toll the Hounds)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Kai Kessenbrock, for his mentorship and guidance throughout my time in the lab as his student. I would not be the scientist I am without his supervision, and am better off for it through and through. Dr. Kessenbrock really is the mastermind behind the success of the projects that I have worked on, and really drove the science forward with his insight and motivation.

I would like to thank the larger Kessenbrock-Lawson lab group for their feedback and friendship over the years. We were fortunate to have had the opportunity to all work together in the way that we did, not nearly as many graduate students were as lucky to have such a great group. It was not always easy, but I think we all benefited from each other in ways measurable and immeasurable. Specifically, I would like to thank Ryan Davis and Kerrigan Blake for their contributions to my research on basically every project that my hands touched, as well as for their friendship during what is in general a pretty stressful experience for all involved.

I would additionally like to thank Dr. Gary Huffnagle, my undergraduate mentor at the University of Michigan, Ann Arbor. I can truly say I would not be on the path I am currently on if not for his excellent mentorship and forethought into the future of what a young scientist will need to be most useful to their lab. He encouraged me to seek the computational side of research and helped to design a course curriculum to best supplement this idea when no formal one existed at the time.

I would like to thank my family, who's contribution to my life cannot be captured in a simple paragraph. I have always considered myself very lucky to have the parents that I do, who worked tirelessly to foster a love of learning and reading in myself and my brother when we were children. It is this joy of discovery that I believe has carried me through to make the choice that I have regarding my schooling and life and will continue to do so. My brother Michael has also been a great influence on me, always making the time for me when I need an ear to listen and offer advice or some much needed grounding. For that, I thank you.

I would like to thank my girlfriend, Adele, who has been at my side through this experience. Her support has gotten me through many difficult times and I do not know how I would be had I not had her with me. Thank you dear.

I also thank Frontiers for permission to include material for Chapter One of my thesis, which was originally published in *Frontiers in Cell and Developmental Biology*. I thank Springer Nature for permission to include material for Chapter Two of my dissertation, which was originally published in *Nature Communications*. Finally, I thank Elsevier for permission to include material for Chapter Three of my thesis, originally published in *Cell Reports*. Financial support was provided by the University of California, Irvine and NIH/NIBIB Grant 5T32EB009418-07.

VITA
Nicholas Pervolarakis

EDUCATION:

Ph.D. - Mathematical, Computational, and Systems Biology *2015-Present*
University of California, Irvine

B.S. – Microbiology *2011-2015*
University of Michigan, Ann Arbor

RESEARCH EXPERIENCE

Ph.D. Graduate Student Researcher, Kessenbrock Lab, UCI *2016-Present*
Understanding breast cancer initiation in single cell resolution through multi-omic integration

Scientific collaborator, Single Cell ATAC-seq, 10x Genomics *2018*
Collaboration with industry to generate pilot mouse mammary epithelium open chromatin atlas

Data Analyst, NIH/NCI U54-Center for Cancer Systems Biology *2016-2019*
"Complexity, Cooperation and Community in Cancer"
Project: Single Cell Heterogeneity in Wnt and Metabolism in Xenograft Colon Tumors Xenografts.

Data Analyst, Dr. Gary Huffnagle's Lab, University of Michigan *2013-2015*
Exploring microbial community perturbations in pulmonary and gut tissue through 16srRNA sequencing

FUNDING AND AWARDS:

NIH T32 Training Grant *2016-2018*
Mathematical, Computational and Systems Biology

CCBS Opportunity Award *2016-2017*
Collaboration with Waterman Lab at UCI. "Single Cell Heterogeneity in Wnt and Metabolism in Xenograft Colon Tumors Xenografts."

CCBS Opportunity Award *2017-2018*

Collaboration with Mortazavi Lab at UCI. “Doublet Cell Sequencing to Elucidate Stem Cell Niche in Mouse Epithelium”

CCBS Opportunity Award *2018-2019*

Collaboration with Villalta Lab at UCI. “Characterization of macrophage heterogeneity in dystrophic muscle”

PUBLICATIONS:

1. Nguyen, Quy H*., **Pervolarakis, Nicholas*** et al. "Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity." **Nature communications** 9.1 (2018): 1-12. ***Contributed equally**

2. **Pervolarakis, Nicholas**, et al. "Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Regulators of Mammary Epithelial Cell Identity." ***Cell Reports*** 33.3 (2020): 108273.

3. Alshetaiwi, Hamad, **Nicholas Pervolarakis**, Laura Lynn McIntyre, Dennis Ma, Quy Nguyen, Jan Akara Rath, Kevin Nee et al. "Defining the emergence of myeloid-derived suppressor cells in breast cancer using single-cell transcriptomics." ***Science Immunology*** 5, no. 44 (2020).

4. Lawson, Devon A., Kai Kessenbrock, Ryan T. Davis, **Nicholas Pervolarakis**, and Zena Werb. "Tumour heterogeneity and metastasis at single-cell resolution." ***Nature Cell Biology*** 20, no. 12 (2018): 1349. **F1000 Recommended**.

5. Kessenbrock, Kai, Quy Nguyen, **Nicholas Pervolarakis**, and Kevin Nee. "Experimental Considerations for Single Cell RNA Sequencing Approaches." ***Frontiers in Cell and Developmental Biology*** 6 (2018): 108

6. Nusbaum, David J., Fengzhu Sun, Jie Ren, Zifan Zhu, Natalie Ramsy, **Nicholas Pervolarakis**, Sachin Kunde et al. "Gut microbial and metabolomic profiles after fecal microbiota transplantation in pediatric ulcerative colitis patients." ***FEMS Microbiology Ecology*** 94, no. 9 (2018): fiy133.

7. Kessenbrock, Kai, Prestina Smith, Sander Christiaan Steenbeek, **Nicholas Pervolarakis**, Raj Kumar, Yasuhiro Minami, Andrei Goga, Lindsay Hinck, and Zena Werb. "Diverse regulation of mammary epithelial growth and branching morphogenesis through noncanonical Wnt signaling." ***Proceedings of the National Academy of Sciences*** 114, no. 12 (2017): 3121-3126.

Conferences and Workshops:

Statistical Modeling of Big Data *2016*
University of California, Irvine

Next Generation Sequencing Data Analysis *2016*
University of California, Irvine

Center for Cell Circuits Computational Genomics Workshop *2016*
Broad Institute, MA

Nature Masterclass in Scientific Writing and Publishing *2017*
University of California, Irvine

Pilot Projects for a Human Cell Atlas Retreat *2018*
Asilomar, CA

TEACHING:

Using the High Performance Computing Cluster for Research *2017*
Biophysics and Systems Biology Seminar, UCI.

- Taught workshop to give students the tools to interact with the High Performance Computing Cluster (HPC) at UCI, organize and manage data, and write scripts for queue submission

Introduction to Task Arrays *2018*
Bioinformatics Support Group, UCI

- Led seminar on the theory and application of task arrays for massively parallel batch job submission for cluster computing

Single Cell RNA-seq Analysis Computational Lab *2018*
Cancer Systems Biology Short Course

- Organized and taught 4 hour workshop introducing single cell RNA-seq data, technical challenges, and analysis considerations and workflow

Getting Started with Single Cell RNA-seq *2018*
Bioinformatics Support Group, UCI

- Presented recommendations and for preliminary single cell RNA-seq analysis for first time users and new datasets

ABSTRACT OF THE DISSERTATION

Single cell sequencing analysis reveals mammary epithelial cell diversity and regulation
by
Nicholas Pervolarakis
Doctor of Philosophy in Mathematical, Computational, and Systems Biology
University of California, Irvine, 2021
Assistant Professor Kai Kessenbrock, Chair

The mammary epithelial system is a heterogeneous cellular compartment thought to be comprised of two major cell types, basal and luminal respectively, that are in flux throughout an individual's life and require a stem cell compartment to maintain. Questions remain about the origin and nature of these stem cells, and how the constituent components of the gland interplay in order to maintain a healthy tissue or ultimately responds to cancer present. Through the usage of single cell microfluidic based experimental tools, we have been able to explore this relevant heterogeneity with single cell RNA sequencing (scRNA-seq) in human and mouse and single cell ATAC sequencing (scATAC-seq) in mouse. We highlight in both species the different basal and luminal cell types, with the stratification of the luminal compartment into hormone response or secretory cells. Additional relevant cell states present within the secretory luminal cell type manifest with tissue relevant consequences. Using scRNA-seq in human we present three major epithelial cell types, one basal and two distinct luminal referred to as L1 (Secretory) and L2 (Hormone Responsive). After applying pseudotemporal reconstruction, it is shown that the three populations interconnect in a developmental lineage with basal cells branching into the two luminal end points. In mouse, a similar three epithelial cell type structure is highlighted in the mammary gland with both scRNA-seq and scATAC-seq in an integrated analysis. The secretory luminal compartment is additionally stratified into luminal progenitor and lactation-committed progenitors, with distinct regulatory features underpinning each cell state through both cis and trans acting elements. Taken together, these results emphasize newly discovered heterogeneity in the luminal compartment of the mammary gland, challenging of previously held definitions of the mammary stem cell, and the underlying regulation of cell state.

Chapter 1: INTRODUCTION

Background and Motivation

Breast cancer is currently the most commonly diagnosed form of cancer in the united states, and possesses the second highest death rate caused by the cancer behind only lung/bronchus cancers [https://gis.cdc.gov/Cancer/USCS/DataViz.html]. The disease itself is not a monolith, with a multitude of molecular subtypes that have been previously described through pathological and expression based studies[1]. These subtypes, namely Luminal A, Luminal B, Her2, Triple Negative (basal like), Normal-like, and Claudin-low, exhibit distinct expression patterns and regulation of hormone receptors[2]. Although prognosis remains good for patients that are identified at an early stage, late stage survival is still quite poor. This serves to highlight the critical importance of early detection for patients, and with that comes the need for a better understanding of what cell types and states are represented in the epithelial compartment of the mammary gland. A thorough description of the heterogeneity present in adult epithelial tissue, the cells that ultimately develop into breast cancer, under normal homeostasis will provide stronger insight into what trajectory that cell takes as it develops into a disease state. With this, scientist and clinicians can better stratify patients and give better risk assessments throughout the lifetime of an individual.

The human mammary gland is comprised of a series of branching epithelial ducts and lobular structures embedded in an adipose rich tissue[3]. These epithelial structures exist as a bilayer with the inner component of cells referred to as luminal cells which surround the lumen of the ducts and lobules and perform a secretory function in the lactating gland, and basal cells which in turn surround these luminal cells and have a contractive ability thought to aide in the flow of milk through the tissue[4]. The mammary gland is one that even in adulthood requires the ability to undergo drastic changes in structure and function, in the cases of natural monthly hormonal cycling, pregnancy, lactation, and subsequent involution. It is thought that this plasticity is achievable through the existence of an epithelial stem cell compartment that is constantly maintained in the gland, referred to as the mammary stem cell (MaSC). Many studies have been performed in the hopes of identifying the MaSC in human, but many cell surface marker based flow cytometry assays have fallen short in clearly defining the ultimate cells of interest[5].

The mouse mammary gland has also served as a critical model system for the study of the dynamics of the tissue, and how the scientific community can better test hypotheses that working with human samples cannot achieve. In mouse, studies with the intent of finding the MaSC have been more fruitful, where through the transplantation of specific individual cells of the basal lineage researchers have been able to reconstitute a fully branched gland[6,7]. In the first of these two landmark studies, researchers transplanted isolated single Lin^{-} CD29^{hi} CD24^{+} cells into cleared mammary fat pads and were able to show the development of a healthy, fully functional gland that can produce milk. In the second by Stingl et al., individual transplanted mouse basal cells marked by $\text{CD45}^{-}\text{Ter119}^{-}\text{CD31}^{-}\text{CD140a}^{-}\text{CD24}^{\text{med}}\text{CD49f}^{\text{high}}$ were isolated through limiting dilutions. These mammary repopulating units (MRU) were able to reconstruct a fully branch gland when injected into a cleared mouse mammary fat pad at a 1 in 60 or 1 and 90 acceptance rate depending on the mouse model in question. This research is not without debate, and it has to be acknowledged that there is a difference between the capacity that a cell has in the context of a cleared gland to reconstitute its structure vs in an adult tissue where the gland is already established and does not have to serve such a plastic role. In addition to transplantation assays, lineage tracing experiments in mouse have been employed to further investigate the potential of different cell lineages and their role in the developing gland[8]. Specifically, Wang et al. expand upon the isolation strategy described above for transplantation assays and further delineate the $\mathrm{Lin^{-}}$ $\mathrm{CD29^{hi}}$ $\mathrm{CD24^{+}}$ compartment into *Procr*$^{+}$ and *Procr*$^{-}$ cell populations. The *Procr*$^{+}$ population is then interrogated using a *tdTomato*$^{+}$ / *GFP*$^{+}$ reporter line specific to the basal lineage that shows the multipotent potential of basal cells in this compartment both through development as well as adult maintenance of the gland. More recently, a comprehensive review paper by Fu et al. 2020 highlighted the state of the field in terms of mammary gland differentiation hierarchy[9]. Focusing on their summarized discussion of the progenitor cells in the adult gland, one result that seemed consistent from previous studies is the notion of the accumulation of CD61+ positive cells (thought to be indicative of the luminal progenitor population) when factors such as Gata3 and Elf5 are ablated in cells and then permitted to reconstitute the gland[9]. Taken together, these results still do not paint the full picture of the epithelial cell hierarchy in the mouse gland, let alone human.

An underlying problem of these studies is the inability to survey the full spectrum of cell types and states present in the gland in an unbiased fashion. Different types of cells in one study as separated by FACS markers x, y, and z might be lumped in the next studies gating strategy. For expression based assays like microarray or RNA-seq, you are limited by the purity of your sort to analyze what cell types are expressing what genes relevant to the biological question and important differences can and will be lost through the blending of different heterogeneous signatures present in the underlying cells sampled that would have otherwise offered important insight and hypotheses to test. It is with this in mind that we look to the burgeoning field of single cell based genomic assays to delve deeper into the true biology at play.

WHY SINGLE CELL OMICS TECHNOLOGIES ARE THE TOOL FOR THE JOB

Single cell based omics techniques have exploded in popularity for many of the same reasons that we tout the expansion of sequencing in general and its similarly profound reduction in cost per unit of information generated. We stand on the shoulders of many fields, with biochemistry, molecular biology, material science, and microfluidics foremost among them that contribute to what is possible today. The two modalities that have lent themselves best to single cell resolution have been RNA-sequencing (the capture and sequencing of mRNA molecules of a sample)[10] and ATAC-sequencing (the capture and sequencing of regions of open and accessible chromatin within the cells of a sample)[11]. Individually, scRNA-seq can provide a snapshot of what a cell is “thinking” by capturing the mRNA that is actively transcribed and ready to be turned into protein to ultimately perform the functions that the cell needs[12]. For scATAC-seq, a snapshot is captured of what the cell has the potential to be thinking and provide insight into what features of the cells regulatory machinery are at work in the cell of questions current state such as transcription factor binding and novel enhancer regions[13].

Combined together within a single study, these tools allow for the gain of insight not only into what cell types and states are present in a sample of interest, but additionally learn what regulatory underpinnings that better explain the observations made at the mRNA level[14–16]. Single cell omics assays create the opportunity to isolate an individual cell and perform the same traditional molecular assays thousands of times over for any sample of interest. The resultant dataset is then the opportunity for the direct identification of cellular heterogeneity within, without having to deal with complicated and limited deconvolution algorithms from an otherwise bulk dataset[17].

Single cell technologies are not a perfect tool, and as with any fledgling technology, it has its own suite of caveats associated with the generation and analysis of its data. Highlighted in Figure 1.1, we will now delve into greater detail regarding single cell library preparation and experimental design considerations.

CELL DISSOCIATION AND SINGLE-CELL PREPARATION

The process of single-cell preparation is arguably the greatest source of unwanted technical variation and batch effects in any single-cell study[18]. 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 therefore 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[19] have the potential to 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 tissue specimen, through cellular aggregates and clusters, and finally eluting a solution containing close to 100% single cells, which will be ideal for scRNAseq applications. In addition, 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 and efficient tissue dissociation and single-cell preparation[20,21]. Ultimately, determining a "best practices" dissociation strategy through heuristic optimization will be critical for downstream single-cell library quality.

Cell Type Enrichment

There are various methods for isolating specific cell populations or removal of unwanted populations that should be optimized for any specific tissues type. Manual isolation utilizing magnetic beads or gradient purification are potential methods for removal of 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[22,23]. However, these methods are not without drawbacks, since they can introduce artificial stress on cells and change their expression profile[24]. 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[25,26]. Flow cytometry-isolated cells are exposed to 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[24,27,28].

Quality Control

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 a number of readouts. A useful first metric can be acquired using imaging of viability such as using the Countess platform (Thermo Fisher Scientific). 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 in the appropriate frequency. Finally, an additional metric on RNA quality can be acquired using the RNA integrity number (RIN) method[29].

SINGLE-CELL TRANSCRIPTOMIC PLATFORM

Protocols for transcriptome analysis have advanced rapidly, resulting in several robust methods which range in cell and mRNA capture strategy, barcoding, throughput, and level of automation[12,30]. Selection of the optimal approach depends largely on the research question. Recent high-throughput protocols for scRNAseq 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[31–35]. The newest of these protocols utilize beads functionalized with oligonucleotide primers, which each contain a universal PCR priming site, a cell-specific barcode, an mRNA capture sequence, and Unique Molecular Identifiers (UMI). Individual cells are captured in wells or droplets with a single bead. Cell-specific barcode are similar within a droplet but unique UMI sequence on the primer allows for individual transcripts within a cell to be counted. This provides a quantitative readout of the number of transcripts of each gene detected in a cell, thereby reducing the effects of amplification duplicates that occur with earlier technologies[36,37]. 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[12,34]. High-throughput approaches typically only achieve ~10% transcriptome coverage, relative to ~40% for full-length scRNAseq protocols that use Switching Mechanism at 5' End of RNA Template (SMART) chemistry[38,39]. 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 sensitivity for detecting low-expressed genes[40].

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)[38]. 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[38,39]. cDNAs are then amplified by

PCR and libraries are prepared for sequencing using standard protocols. Although there have been several large scale projects utilizing these protocols, because they are manual in nature and utilize larger microliter reaction volumes, they limit the number of cells that can be processed at reasonable cost.

Another area of ongoing debate is how to determine how many cells one should be analyzed to reach sufficient statistical power. Several methods have been developed using power analysis statistics, such as Scotty or web-based tools , but one must estimate the number and expected frequencies of cell populations present in the sample, and such information is often not available. Therefore, these decisions are usually made based on logistical restraints (i.e., the number of cells available), financial considerations, or re-iterative experiments where an initial sample of cells is sequenced to get a sense for overall population structure, and then increasing numbers of cells are sequenced until one is satisfied that all the main populations have been identified.

SINGLE NUCLEI ISOLATION AND SEQUENCING

Single-cell RNA sequencing methods are optimal when cells can be harvested intact and viable[41]. However, certain cell types (e.g., neurons, adipocytes), are not amenable to standard organ dissociation protocols, since enzymatic and mechanical forces easily disrupt the cytoplasmic contents[42]. In these cases, an option could be to isolate intact nuclei for single-nucleus RNAseq (snRNAseq)[41–46]. To prepare single nuclei, cells are lysed with detergent and dounce homogenized to expel cytoplasmic contents and nuclei from the cellular membrane[43], which may avoid transcriptomic changes[24]. Nuclei can then be purified by flow cytometry or gradient centrifugation[41,43,47]. When cell-type specific nuclear proteins exist, they can be used for nuclei isolation from specific cell types using antibody labeling[42,45].

Single-nucleus RNAseq (snRNAseq) is not only amenable for difficult to isolate cell types, but can also be used for archived tissues such as flash-frozen clinical samples. Individual nuclei isolated from frozen adult mouse and human brain tissues have been successfully sequenced, demonstrating that snRNAseq has sufficient resolution to identify many different cell types from frozen and post-mortem tissue[41]. With the rapid development of many applications for snRNAseq, nuclei are amenable to other studies not easily done by scRNAseq.

An important question remains: To what degree is the nuclear transcriptome representative of the whole cell? Recent studies have demonstrated that many transcripts of cell http://scotty.genetics.utah.edu/ 2http://satijalab.org/howmanycells and nucleus are equally represented and that nuclear RNA represents an important and significant population of transcripts that contribute greatly to the overall diversity of transcripts[48,49]. Comparative studies of scRNAseq and snRNAseq in neural progenitor cells have also demonstrated that genes are expressed in equal proportion between whole cell and nuclei[41]. Nanogrid single-cell and nuclei RNA sequencing studies in the same breast cancer lines found that overall copy number, expression level, and abundance had a high (rs = 0.95) Spearman's correlation[50]. Similarly, the transcriptomes of single cells and nuclei of 3T3 cells have also demonstrated high correlation (Pearson, r = 0.87)[42]. Together these results suggest that nuclei and cells have highly correlated relative gene expression.

Despite the similarities between single-cell and nuclei transcriptomic profiles there remain notable differences. Not surprisingly, nuclear transcriptomes are enriched for several types of nuclear RNAs (ncRNAs)[41–44,50]. Since ncRNAs are only polyadenylated in the nucleus, snRNAseq provides a feasible strategy to capture the heterogeneity of ncRNA transcription in single-cell resolution[44]. In addition, nuclear transcriptomes are enriched for lncRNAs and nuclear-function genes[50]. Another difference between cell and nuclear RNAseq is the higher abundance of intronic sequences in snRNAseq, which ranged between 10–40% of mapped reads[41,42,50]. These features need to be accounted for when comparing datasets from cellular versus nuclear transcriptome analyses. In conclusion, snRNAseq has emerged as a promising avenue for profiling archived samples or cell types that are hard to viably isolate from tissues.

SINGLE-CELL LIBRARY SEQUENCING

The next critical part of designing single-cell workflows is to align the analysis pipeline with the respective NGS platform and sequencing depth. It is important to confirm that the chemistry used for library construction is compatible with the sequencing technology. Currently, there are two major outputs for libraries from scRNAseq: full-length transcript or 3' -end counted libraries, which each require different read depths[51]. Full-length transcript libraries are typically sequenced at a depth of 10^6 reads per cell, but may still yield important biological information at as low as 5×10^4 reads per cell[10]. For specific applications such as alternative splicing analysis on the single-cell level, much higher sequencing depth up to $15– 25 \times 10^6$ reads per cell is necessary. On the other hand, 3' -end counting libraries are sequenced at much lower depth of around 10^4 or 10^5 reads per cells[51]. Reaching the optimal sequencing depth can be an iterative process and may require multiple rounds of optimization. Sequencing saturation can be estimated by plotting down-sampled sequencing depth in mean reads per cell (e.g., 10X Genomics Cell Ranger).

STUDY DESIGN AND DATA ANALYSIS

In the following section, we highlight several key considerations from a data analysis perspective for adequately designing a successful scRNAseq study. As mentioned, many single-cell technologies can be greatly affected by technical variation, and without proper study design the results can be difficult to interpret. One critical aspect of this is the separation of batch and condition. Batch refers to a library that was singularly generated in a contained workflow (i.e., harvesting tissue specimen, disassociating into single-cell suspension, and generating scRNAseq library). Condition refers to a biological state or experimental treatment that is being analyzed in the study. Technical variation can be difficult to separate from relevant biological variation when conditions are interrogated individually. To help correct for this, the generation of replicates (biological or technical) whenever possible is strongly recommended.

In addition to replicates, an option is to mix samples and conditions within a batch, such that they can be treated without confounding each other[52]. One example is the Demuxlet workflow, where samples from genetically distinct individuals can be processed within the same library generation protocol and sequenced together[53]. Prior to library generation, genotyping of distinct samples is performed and subsequently used in conjunction with the scRNAseq library to demultiplex the mixed cell sample into the samples of origin. In situations where genetically identical samples are used, or genotypic data is not readily available, cellular hashing can be employed[54]. This involves oligo-tagged antibodies specific to each sample in the study and then pooling and generating the scRNAseq library from the sample mixture. The antibodies labeled with unique barcodes can be traced back to its sample of origin[54].

Efforts can be made computationally to mitigate batch-to batch variation. Batch effects are not unique to scRNAseq data, but the assumptions made by correction algorithms are not always appropriate for the bimodality of gene expression in zero-inflated scRNAseq data. Here, we highlight recent analytical frameworks that may be used to correct for this phenomenon. A recently developed approach by Haghverdi et al. (2018)[55] builds a mixed nearest neighbor model for cells between datasets or samples that does not require known or equal proportions of cell types between data sets. In addition, the widely used Seurat pipeline for scRNAseq analysis has implemented a workflow to allow for not only multiple batches but can integrate even different data modalities by first learning a set of "anchors" between the data groupings in question using Canonical Correlation Analysis (CCA), then projecting query datasets onto a reference set to achieve a coembedding of the data as well as an adjusted gene expression matrix of the features used for the above process[56]. Finally, the single-cell batch correction framework MAST[57] models the positive expression mean and the over-the-background and calculates a fraction of detected genes per cell and uses this as a covariate that is independent of a previously specified control set of genes. Together, these methods serve as recent examples to handle batch-to-batch variation computationally, resulting in improved dimensionality reduction and clustering for meaningful scRNAseq data analysis.

Beyond accounting for technical variation, a common question that researchers address is the relatedness of described cell populations through the lens of a differentiation processes. The key assumption of pipelines that seek to address this is that the tissue sample analyzed using scRNAseq contains cell types/states that represent not only the ends of a differentiation process, but also stem/progenitor cells and transitional cell states along the path of differentiation. Common analysis suites that seek to reconstruct these differentiation trajectories are Monocle[58], TSCAN[59], and CellTree[60]. Each use different methods, but their goal is to visualize differentiation trajectories and identify expression signatures that change through pseudotime.

CONCLUSION

To fully harness the potential of single-cell analysis tools to decipher complex biological systems on the level of individual cells, careful study design and rigorous optimization of every step along the experimental procedure are mandatory. Here, we delineate a step-wise experimental approach for optimizing tissue handling, cell dissociation and enrichment, single-cell platform selection, library sequencing, and data analysis for designing single-cell workflows. A move toward standardized and automated processing of tissues will minimize changes introduced by tissue handling that may obscure biologically relevant transcriptomic profiles. For tissues that are problematic to dissociate into high-quality and viable single-cell suspensions, snRNAseq offers a solution to this problem, and can be used to achieve uniform extraction and sequencing of multiple cell types for cross comparison. Numerous computational frameworks are currently emerging that help mitigate batch effects to separate biological variation from unwanted technical variation.

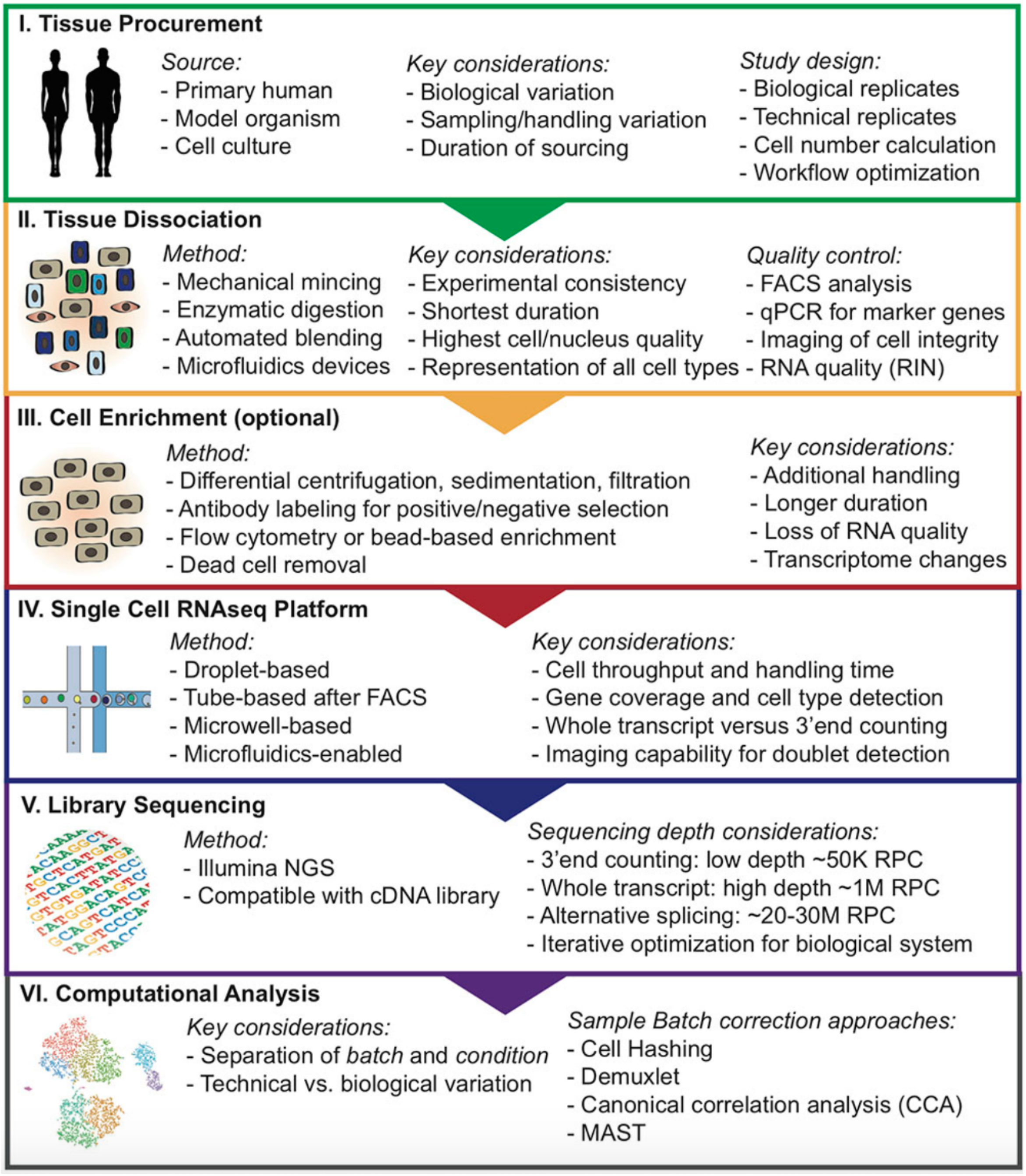

I. Tissue Procurement
Source:
- Primary human
- Model organism
- Cell culture
Key considerations:
- Biological variation
- Sampling/handling variation
- Duration of sourcing
Study design:
- Biological replicates
- Technical replicates
- Cell number calculation
- Workflow optimization
II. Tissue Dissociation
Method:
- Mechanical mincing
- Enzymatic digestion
- Automated blending
- Microfluidics devices
Key considerations:
- Experimental consistency
- Shortest duration
- Highest cell/nucleus quality
- Representation of all cell types
Quality control:
- FACS analysis
- qPCR for marker genes
- Imaging of cell integrity
- RNA quality (RIN)
III. Cell Enrichment (optional)
Method:
- Differential centrifugation, sedimentation, filtration
- Antibody labeling for positive/negative selection
- Flow cytometry or bead-based enrichment
- Dead cell removal
Key considerations:
- Additional handling
- Longer duration
- Loss of RNA quality
- Transcriptome changes
IV. Single Cell RNAseq Platform
Method:
- Droplet-based
- Tube-based after FACS
- Microwell-based
- Microfluidics-enabled
Key considerations:
- Cell throughput and handling time
- Gene coverage and cell type detection
- Whole transcript versus 3'end counting
- Imaging capability for doublet detection
V. Library Sequencing
Method:
- Illumina NGS
- Compatible with cDNA library
Sequencing depth considerations:
- 3'end counting: low depth ~50K RPC
- Whole transcript: high depth ~1M RPC
- Alternative splicing: ~20-30M RPC
- Iterative optimization for biological system
VI. Computational Analysis
Key considerations:
- Separation of batch and condition
- Technical vs. biological variation
Sample Batch correction approaches:
- Cell Hashing
- Demuxlet
- Canonical correlation analysis (CCA)
- MAST

FIGURE 1.1. **Overview of step-wise approach to designing single-cell analysis workflows**. RNA integrity number (RIN); Reads per cell (RPC).

Portions of this chapter were reprinted and adapted with permission from:

Nguyen, Quy H., et al. "Experimental considerations for single-cell RNA sequencing approaches." Frontiers in cell and developmental biology 6 (2018): 108.

KK outlined concept and overview of review. QN, NP, and KN wrote the manuscript. KK and QN designed and prepared the figures.

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

Introduction

Breast cancer is a highly heterogeneous disease that is subtyped based on tissue morphology and molecular signatures[2]. 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[1], and more recently up to ten subtypes have been described[61]. Each subtype is speculated to arise from a different cell of origin[62]; 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. Throughout the duct and lobular system, the breast epithelium is composed of two known cell types, 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[62]. Two landmark papers published in 2006 identified a functionally distinct subpopulation of basal epithelial cells that harbors stem cell capacity and is capable of reconstituting a fully developed mammary epithelial network when transplanted into the cleared mammary fat pads of mice[6,7]. Moreover, a subpopulation of luminal progenitor cells identified by high expression of KIT as well as a subpopulation of mature luminal cells have been identified using flow cytometry (FACS) isolation strategies[63,64]. Interestingly, based on comparative bulk expression analyses, these luminal progenitors may have increased propensity to give rise to triple negative breast cancers in patients with mutations in the BRCA1 gene[65]. 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)[10,66]. 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[66] or brain[10] as well as in various cancers including melanoma[38], glioblastoma[37], and within circulating tumor cells from patients with pancreatic cancer[67].

The goal of the present study is 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[68]. This molecular census can shed light on lineage relationships and differentiation trajectories in the human system and how it relates 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 and 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[40]. Basal and luminal cells were separately loaded onto the Fluidigm C1 microfluidics-enabled scRNAseq platform (Fig. 2.1a). Capture efficiency was monitored by microscopic imaging to exclude doublets and debris from further analysis. 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.

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[12]. This analysis identified three very distinct clusters of cells (Fig. 2.1b), indicating that the breast epithelium is composed of three main cell types. We then explored the genes that are significantly up-regulated within each cluster (Fig. 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. 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. Comparing these signatures to previously published microarray expression analyses of FACS-isolated human breast epithelial cells[65,69], 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)[6–8], we examined the basal cell cluster in more detail. Particularly intriguing was the observation of a subset with increased expression of mesenchymal and stem cell markers *ZEB1*[70] and *TCF4* (Fig. 2.1d). Interestingly, previous work established a direct link between mesenchymal gene expression signatures and MaSC capacity[71], 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 (10x Genomics Chromium)[35]. 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

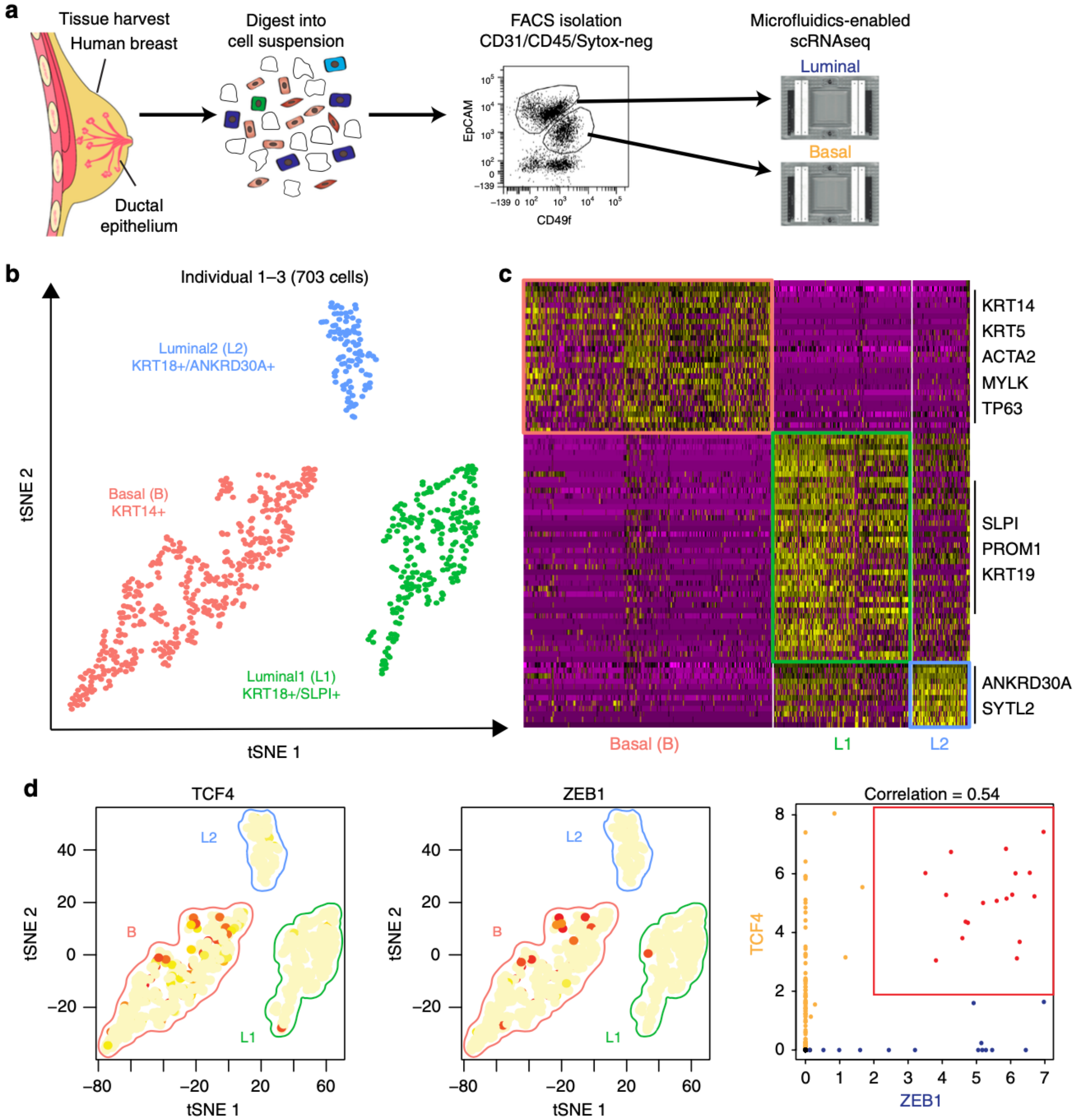


Fig. 2.1 Identification of three major epithelial cell types and their markers using scRNAseq. a Overview of scRNAseq approach using primary human breast tissue samples that were processed into single cell suspension, followed by FACS isolation of basal (CD49f-hi, EPCAM+) and luminal (CD49f+, EPCAM- 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.

cells per sample (Fig. 2.2a). We sequenced a total of 24,646 cells from four individuals (Ind4-7) at an average ~60,000 reads per cell.

After quality control filtering to remove cells with low gene detection (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*+) (Fig. 2.2b). 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 nature 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 (Fig. 2.2c). 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[72]. 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 (Fig. 2.1d), were lowly detected and therefore not interpretable in droplet-enabled scRNAseq, which is likely due to lower coverage compared to the microfluidics-enabled platform[73].

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 (Fig. 2.2). Like Ind4, the other individuals possessed three main cell clusters clearly corresponding to cell types Basal, L1, and L2, and eight to ten subclusters (Fig. 2.3a–c). The number of subclusters per cell type varied across the individuals with Ind5 comprising five Basal, three L1 and one L2 clusters, Ind6 containing seven Basal, three L1 and one L2 clusters, and Ind7 comprising one Basal, three L1 and five L2 clusters (Fig. 2.3a–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[38]. Using the marker gene signatures for each of the 11 clusters (0–10) detected in Ind4 (Fig. 2.2b, c), we performed pairwise gene scoring analyses to find matches for every distinct cluster identified in Ind5–7 (Fig. 2.3a–c). Comparing Ind4 to Ind5–7

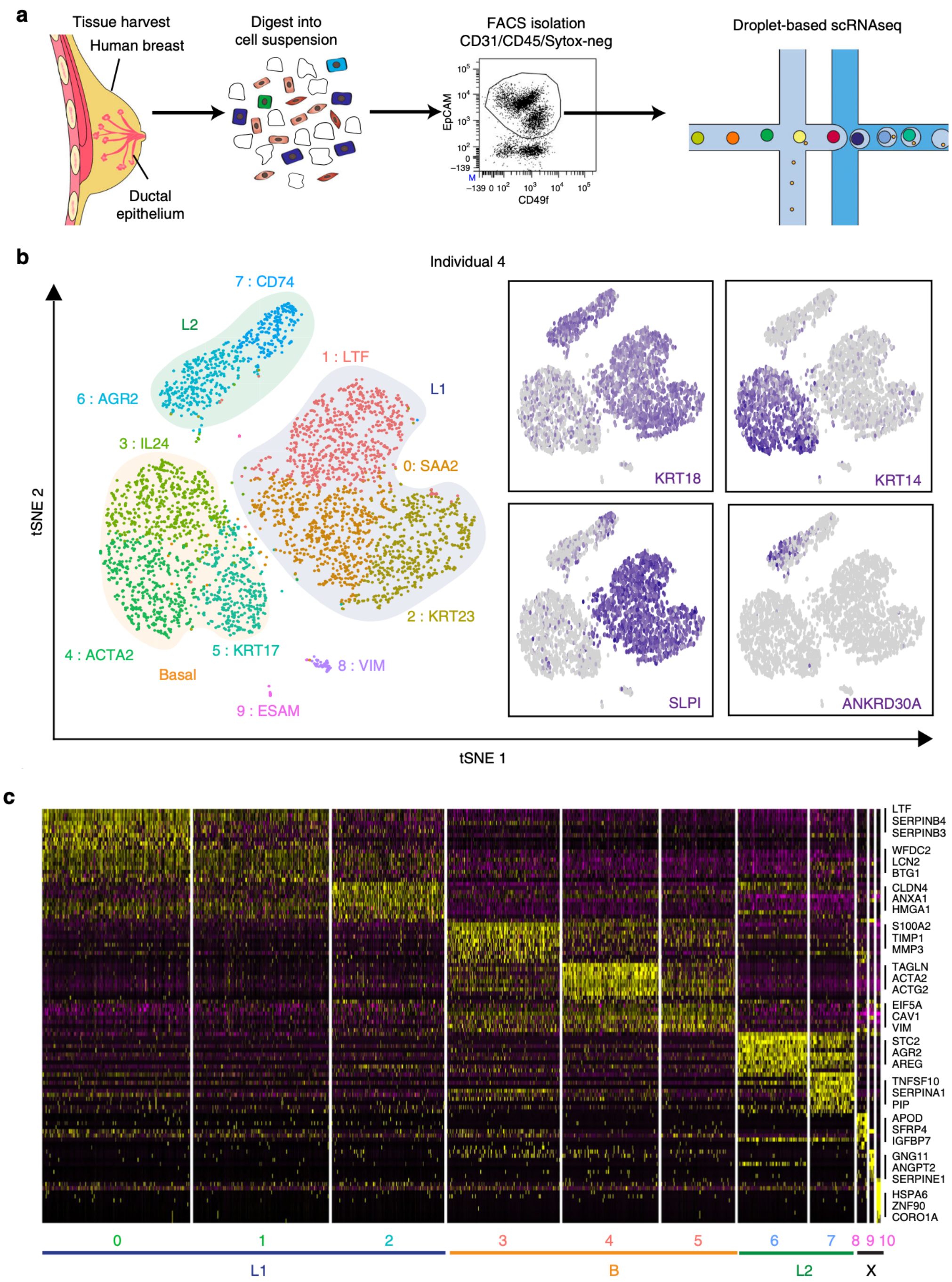


Fig. 2.2 High throughput droplet-mediated scRNAseq reveals additional epithelial cell states. a Overview for droplet-enabled scRNAseq approach as described above; basal and luminal epithelial cells were sorted together and subjected to combined scRNAseq analysis using the droplet-based scRNAseq. **b** 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 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.

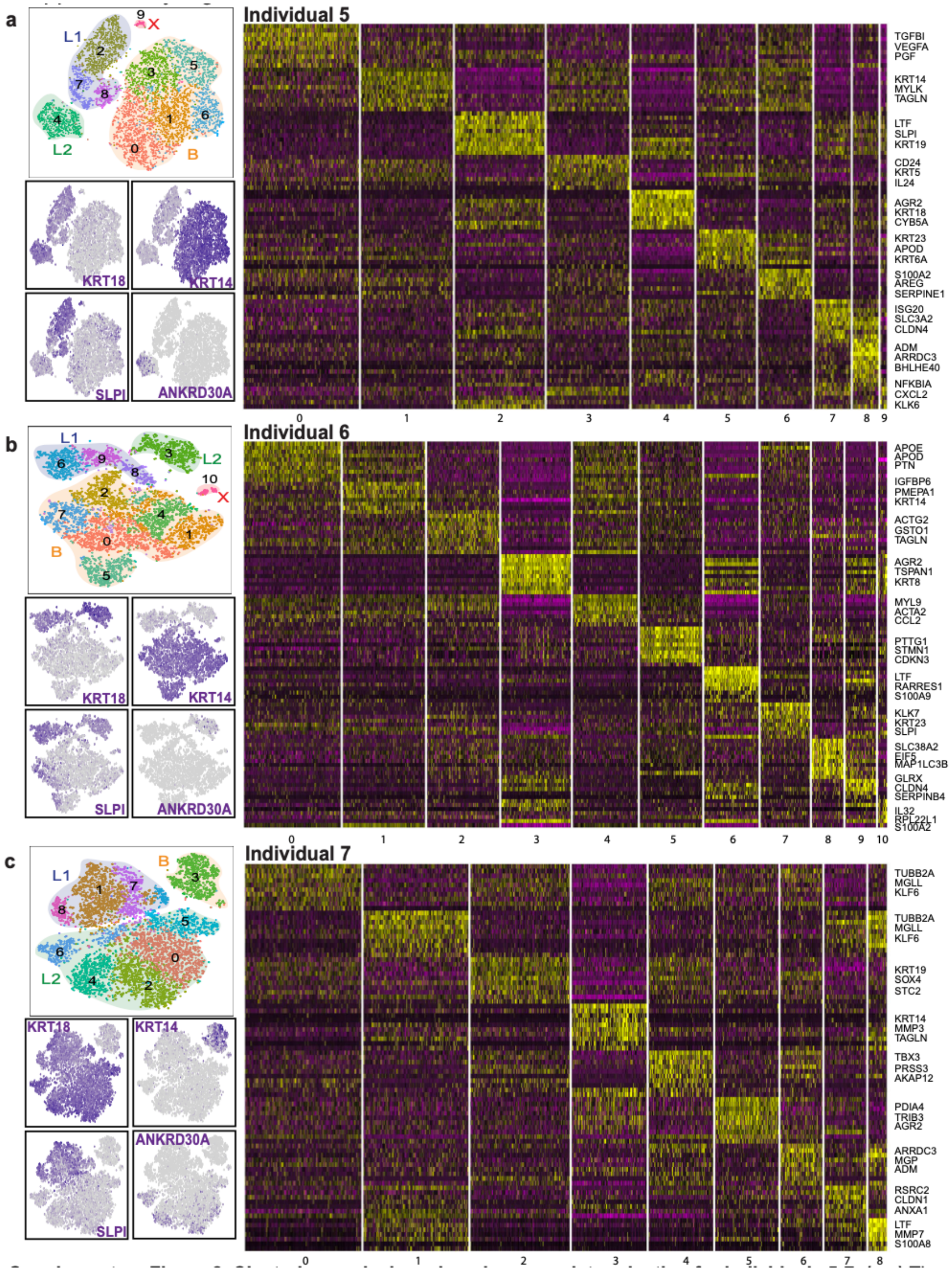


Fig. 2.3. Clustering analysis and marker gene determination for individuals 5-7. (a-c) The individual data matrices for Individual 5 **(a)**, 6 **(b)**, and 7 **(c)** were analyzed using Seurat and their initial cluster determinations are displayed using tSNE projection. Feature plots of characteristic 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 is displayed highlighting selected marker genes for each cluster.

showed that the main cell types (Basal, L1, L2) readily match up across all individuals (Fig 2.4a–c). In addition, it revealed that the 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 found to be more homogeneous, and therefore these clusters were combined to a single L2 population. Comparing basal subclusters between individuals suggested that there are at least two generalizable cell states within basal cells (Fig 2.4a–c). To further explore this, we performed a separate Seurat analysis using combined basal cells from all four individuals (Fig 2.5a). 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[74] to stratify basal cells into either a "Basal" or "Myoepithelial" grouping (Fig 2.5b, 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 Fig 2.5c. These designations were used to perform a combined Seurat analysis of all 24,465 cells from four individuals (Fig 2.4d), 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 are generalizable across all four individuals (Fig 2.4e).

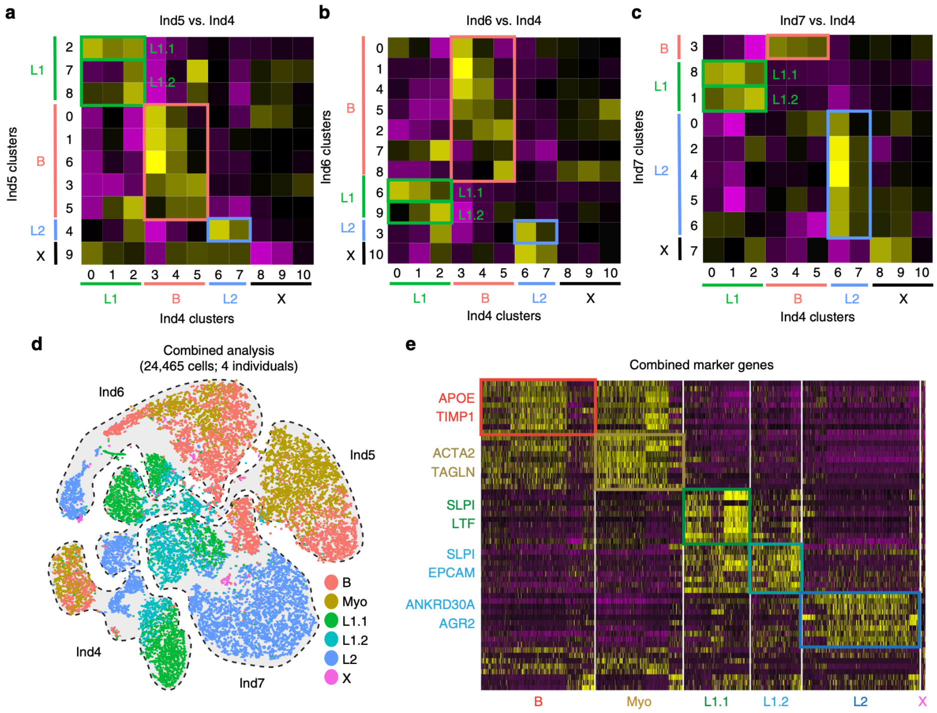
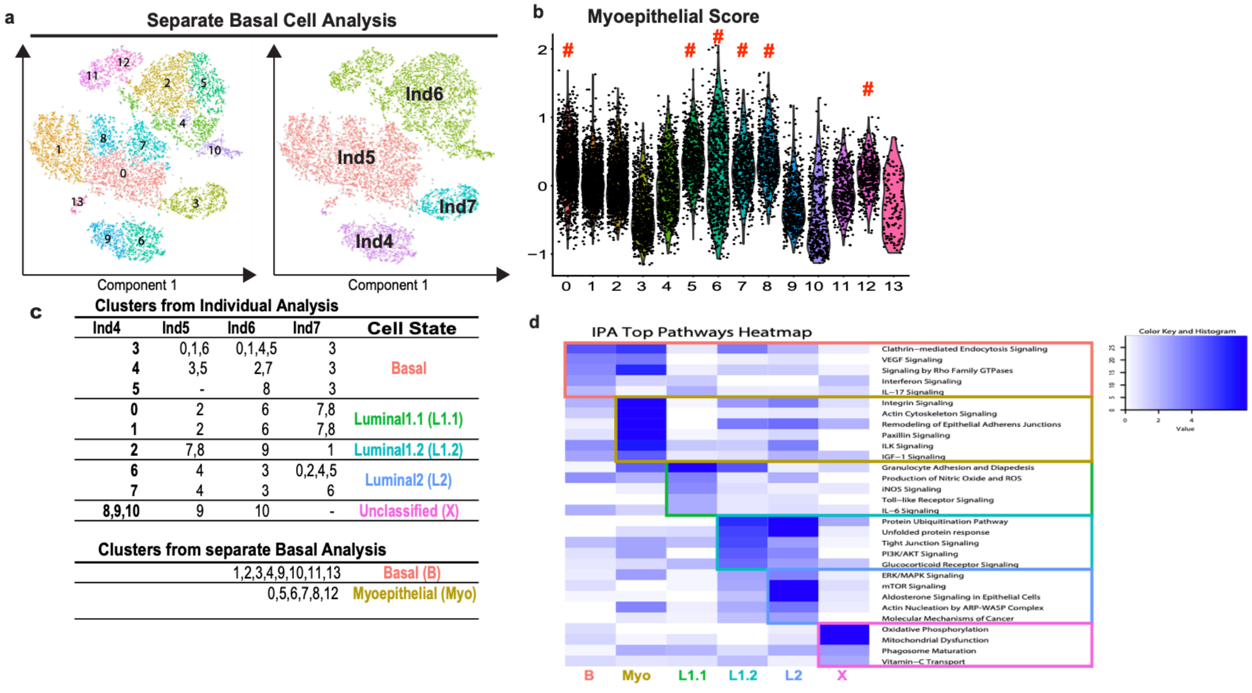


Fig 2.4 Combined droplet based RNAseq data to identify generalizable cell types and states. **a–c** Heatmaps showing gene scoring results using marker genes for Ind4 clusters (0–10; on bottom of heatmap) in all clusters from Ind5 (**a**), Ind6 (**b**), and Ind7 (**c**). Individual-specific cluster IDs are shown in different colors on the right and bottom, and cell type IDs for Basal (b), L1, L2, X are indicated on for every cluster. Data 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 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).



Clusters from Individual Analysis

Ind4	Ind5	Ind6	Ind7	Cell State
3	0,1,6	0,1,4,5	3	Basal
4	3,5	2,7	3	
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)

Fig. 2.5. 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 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.

To learn more about the biology underlying these cell states, we used Ingenuity Pathway Analysis (IPA) to identify distinct signaling pathways (Fig 2.5d), and interrogated for transcription factor consensus sites using the Enrichr tool[75]. These analyses revealed that the Myo state might be controlled by the transcription factors TP63 and PPARγ, and 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[76], 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 levels of PI3K/AKT 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 as well as 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 (Fig. 2.1d), showed that this protein is indeed expressed in a small fraction of basal epithelial cells (Fig. 2.6a). 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[8]. Comparison of published gene expression signatures of ProtCR+ MaSCs with the ZEB1+ population identified here showed striking similarity (Fig. 2.6b), suggesting that the ZEB1+ basal cells may represent a population of human MaSCs. In addition, staining for TCF4, revealed a comparable staining pattern to ZEB1 within the basal (smooth muscle actin-positive) compartment (Fig. 2.6c). These findings show that the cell state characterized by ZEB1 and TCF4 expression exists within the basal compartment in intact breast tissue.

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 population with particularly high expression in the Myo cell state (Fig. 2.6d). 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 (Fig. 2.6e). Myo cells also expressed high levels of the definitive myoepithelial marker ACTA2, as well as other genes associated with smooth muscle

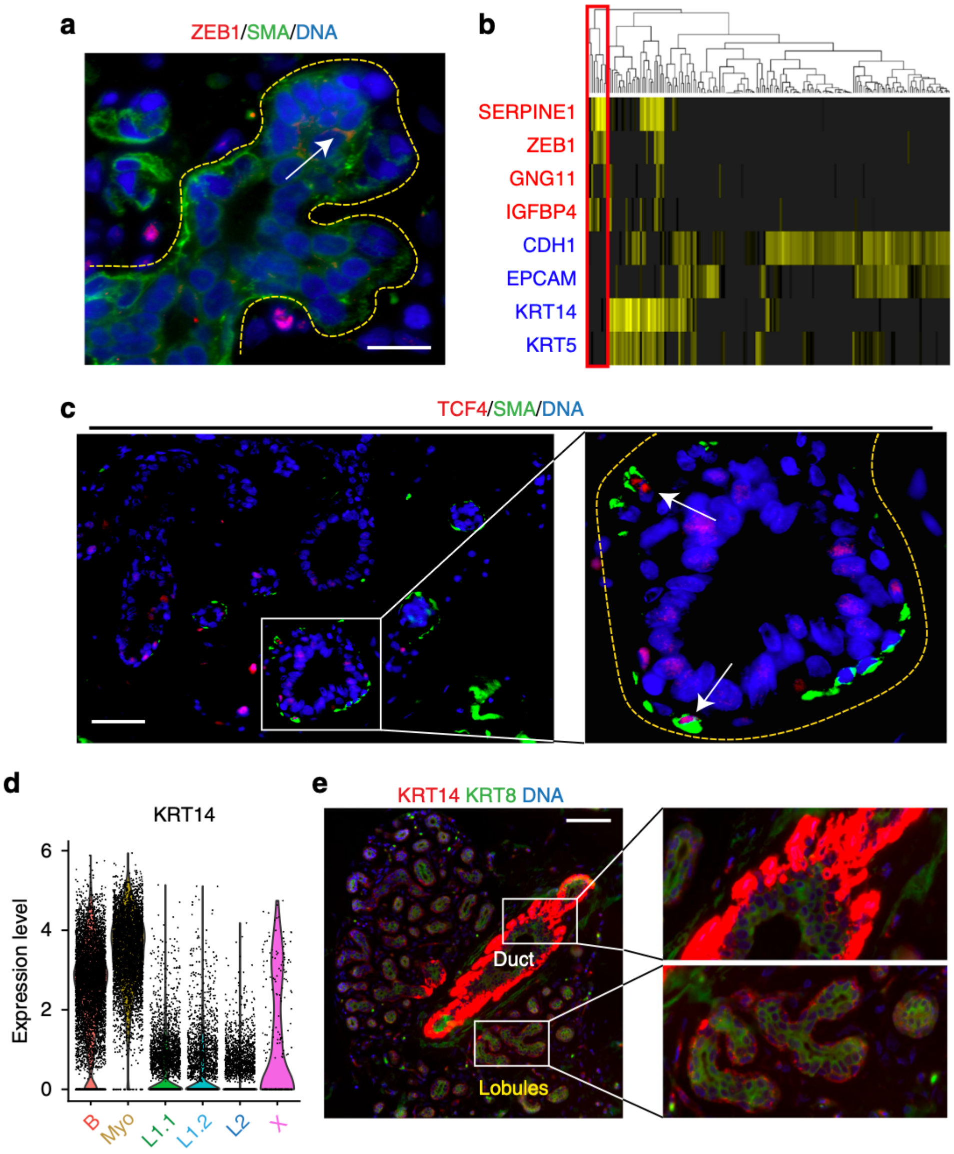


Fig. 2.6 Characterization and spatial integration of basal cell states. a Immunofluorescence analysis of ZEB1 protein expression (red) in combination with basal marker KRT14 (green) and DNA stain using DAPI (blue) within tissue sections from primary human reduction mammoplasty samples showing ZEB1 expression 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 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.

differentiation and function in other tissues such as MYLK, MYL9, and TAGLN/Transgelin[77].

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 (Fig. 2.7a). 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 (Fig. 2.7b). Although this cell state has been previously observed in mouse fetal MaSCs[78], our work revealed that this state exists in the human tissue in adult homeostasis.

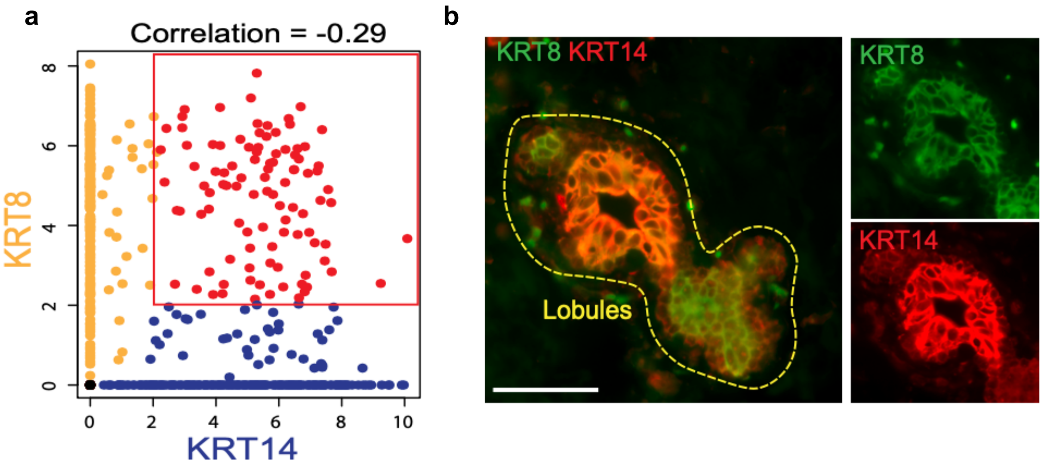


Fig. 2.7. Expanded characterization of cellular heterogeneity within the basal compartment. (a) Correlated expression analysis of luminal marker KRT8 and basal marker KRT14 from scRNAseq data revealed a significant number of double positive cells. **(b)** Combined immunostaining for KRT8 and KRT14 showing rare foci of double positive cells in the luminal cell layer of lobular regions. Scale bar = 50 µm.

The scRNAseq analyses revealed that the luminal compartment harbors two discrete epithelial cell types (L1, L2). To determine if L1 and L2 correspond to 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 (Fig. 2.8a). 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 markedly lower percentage of hormone receptor expression (Fig. 2.8b–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 (Fig. 2.8c).

L2 was also characterized by higher levels of *KRT8* than L1 (Fig. 2.8g). 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 (Fig. 2.8h–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) that 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. Based on their expression signatures, L1 may be committed to secretory function, while L2 likely functions as a hormone-sensing unit of the breast epithelium.

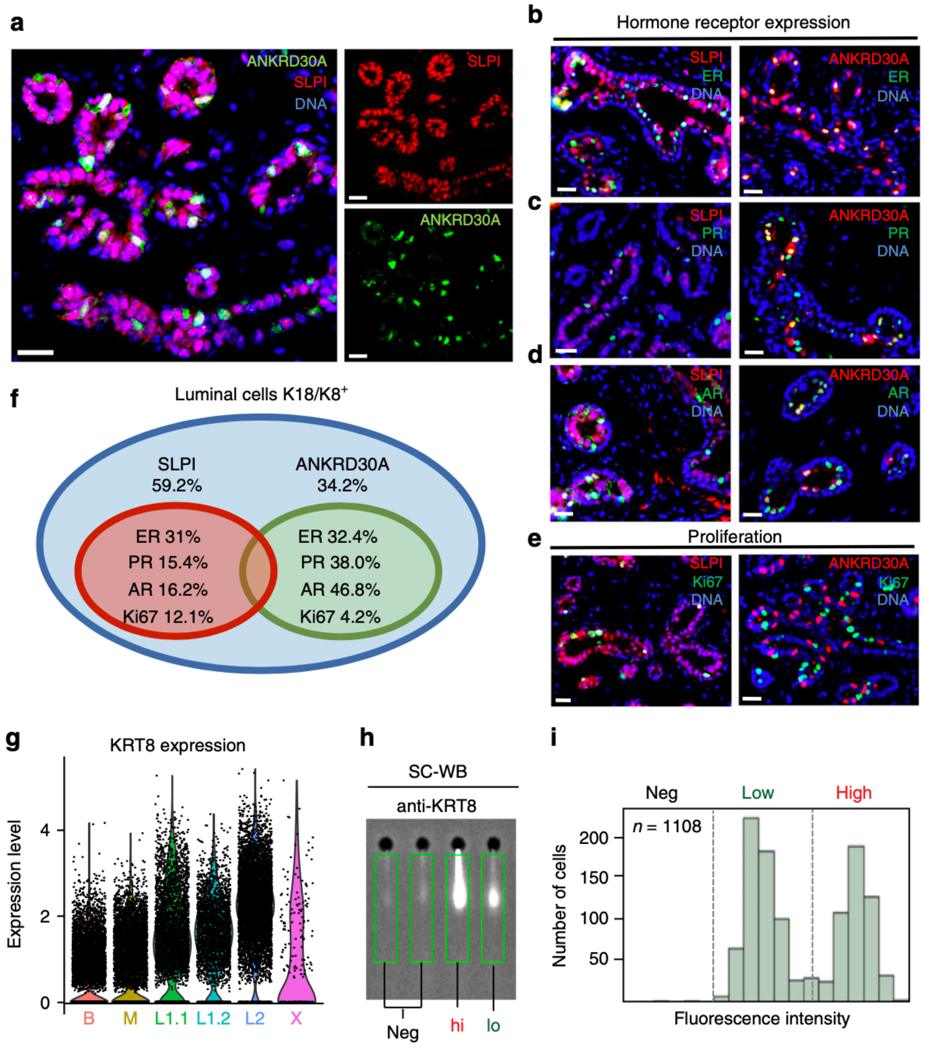


Fig. 2.8 Validation and spatial integration of two distinct luminal cell types. a Immunofluorescence analysis of NY-BR-1 protein expression (green) in combination with basal marker SLPI (red) and DNA stain using DAPI (blue) within tissue sections from primary human reduction mammoplasty samples revealed 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.

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[79]. 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 main cell types Basal, L1 and L2 (Fig. 2.9a). 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[6,7], we manually set the start of pseudotime within the basal cell type (Fig. 2.9b), 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, which have been previously reported as progenitor cell markers[63,64]. Our data instead suggests that L1.1 represents a second mature, differentiated luminal cell type rather than a luminal progenitor that is upstream of L2. These results are in line with previous models of mammary differentiation mediated by bi-potent stem/progenitor cells[62].

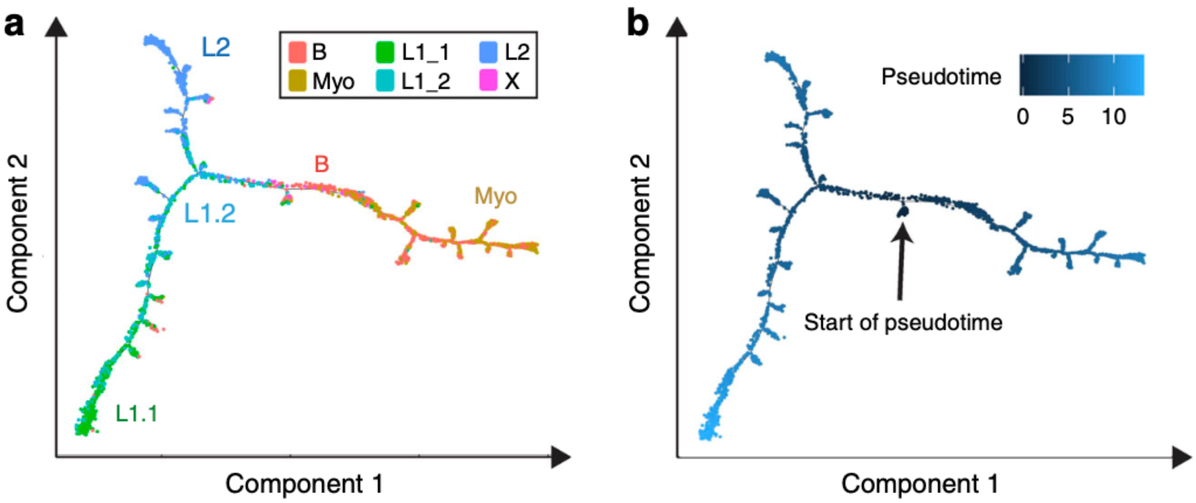


Fig. 2.9 Reconstruction of differentiation and relation of cell states to breast cancer subtypes. **a** Monocle-generated pseudotemporal trajectory of a subsampled population of cells (n = 4000) from four individuals analyzed using droplet-mediated scRNAseq is shown colored by cell state designation. **b** Pseudotime is shown colored in a gradient from dark to light blue and start of pseudotime is indicated.

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 the gene signatures of each population to gene signatures associated with each cancer subtype from the Metabric dataset[80]. This showed that both Luminal A and Luminal B subtypes of breast cancer are closely related to L2- type luminal cells (Fig. 2.10c, top), which is in line with previous gene signature analyses of FACS-enriched basal, luminal progenitor, and mature luminal cells[65]. 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)[81]. We found that Myo showed highest similarity to the mesenchymal-like subtype of TNBC, while the Basal1 class of TNBC yielded highest scores in the luminal L1.1 state (Fig. 2.10c, 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.

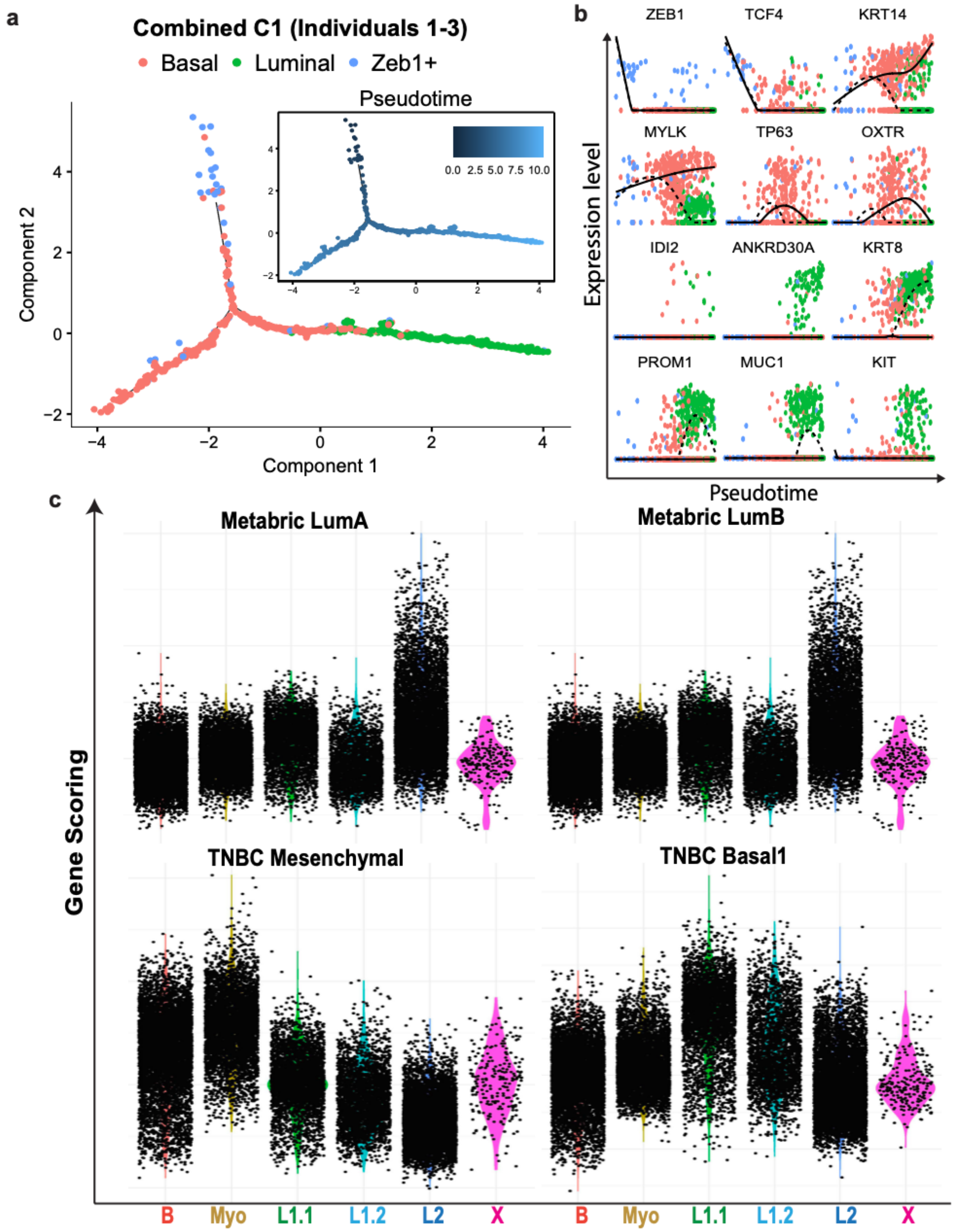


Fig. 2.10. Reconstructing breast epithelial lineage hierarchies their relation to breast cancer. (a) Pseudotemporal analysis of microfluidics-enabled scRNAseq results using Monocle2 based on a set of 183 Seurat identified marker genes suggest a differentiation trajectory from ZEB1+ progenitor cells (green) 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.

Discussion

The current state of knowledge in breast epithelial biology is largely based on population-level analyses of separated basal and luminal cells following bulk analyses of these distinct epithelial cell types[63]. While several distinct subpopulations of murine basal and luminal cells have been reported anecdotally[62], 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-menopause individuals for the first time allow 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[82], although this work referred to these populations as "basal", "luminal progenitor" and "mature luminal cells". Our spatial analyses showed that these three cell types inter- mingle within ducts and lobules, and appear to form functionally distinct lineages that contribute to different aspects of breast biology (summarized in Fig. 2.11a). The fact that all three cell types contained a fraction of proliferative cells suggests that each cell type may be maintained by cycling, lineage-restricted progenitor cell subpopulations 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, namely basal/myoepithelial, secretory L1, and hormone-responsive L2 cells (Fig. 2.11b). The lineage hierarchy likely starts with basal

MaSCs[6,7] that differentiate either into specified myoepithelial cells, or into a common luminal pro- genitor, 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[63,64]. 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.

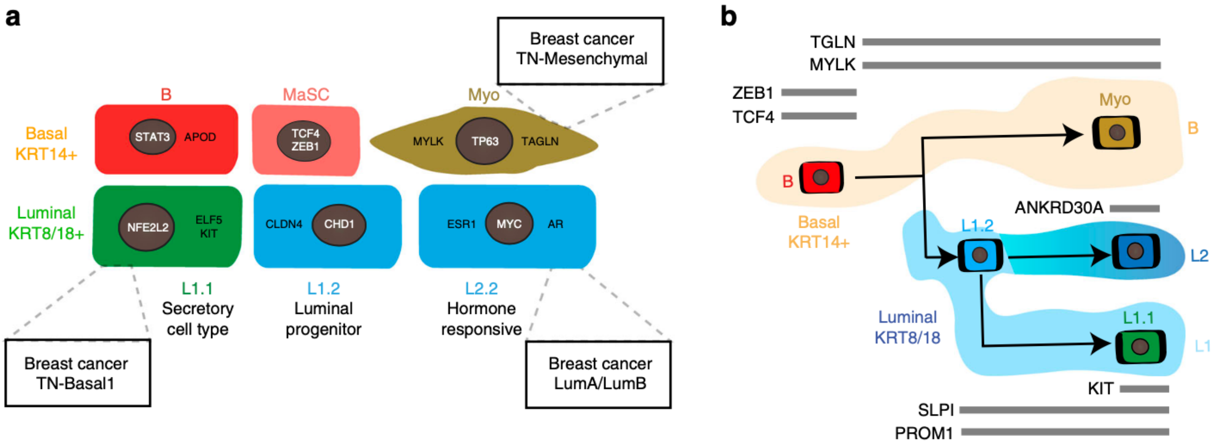


Fig. 2.11 Proposed cellular heterogeneity and lineage hierarchies within the human breast. a Schematic summary of discovered cell states within the basal and luminal compartment of the human breast epithelium with proposed function, key transcription factors (in white), selected markers (in black) and 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)

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 cells[83], or whether homeostasis is mediated through distinct uni-potent, lineage-restricted basal and luminal stem cells[84]. 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 uni-potent 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 clearly 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 thus 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[85], which is in line with previously reported similarities between a FACS-enriched population of mature luminal cells and the luminal-like breast cancer subtypes[65]. The fact that several L2 markers are independently known as breast cancer-associated antigens such as *SYTL2* and *ANKRD30A*[86], and that it shows highest expression of *CDKN1B/p27* as a marker for potential breast cancer cells of origin[87] 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 found to be 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[65]. Interestingly, those cell states containing subsets of proliferative cells, namely B, L1.1 and L2 are predominantly linked to breast cancer subtypes, which is line with previous reports showing an association of mammary epithelial cell proliferation in normal tissues with increased breast cancer risk[88].

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 provide evidence for defined differentiation trajectories to maintain homeostasis in the adult human breast, as well as 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 possibly strategies for cancer prevention.

Methods

Origin of tissue samples.

Anonymous reduction mammoplasty samples were acquired from NCI Cooperative Human Tissue Network (CHTN) and from 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 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 1.2.12[89], 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 in both microfluidics- and droplet-enabled scRNAseq datasets, we utilized the Seurat pipeline[12]. 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[90]. 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 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 previously[38]. 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 Enrichr[75] 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 antimouse 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. 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 expressions values for all cells within each State, and generated a Pearson correlation matrix for these average gene expression value 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.

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 reported35,39. 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 as sorted by log-fold change were then used for downstream scoring

Reprinted with permission from:

Nguyen, Quy H., et al. "Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity." Nature communications 9.1 (2018): 1-12.

Author Contributions:

K.K., D.A.L. and Z.W. designed research and supervised research; Q.H.N., D.M., A.T.P., E.W., R.K., E.J., D.A.L. and K.K. performed research; J.R., S.A.K. and A.G. contributed new reagents/analytic tools and biospecimens; N.P., K.B., R.T.D., I.D. and K.K. performed bioinformatic analyses; Q.H.N., N.P., D.A.L., Z.W. and K.K. wrote the paper manuscript, and all authors discussed the results and provided comments and feedback.

Chapter 3: Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Regulators of Mammary Epithelial Cell Identity

INTRODUCTION

Breast cancer is a heterogeneous disease of at least six intrinsic subtypes, namely, the luminal A, luminal B, HER2-enriched, basal-like, normal breast, and claudin-low subtypes[2]. Breast cancer arises from the breast epithelium, which forms a ductal epithelial network consisting of an inner layer of luminal cells and an outer layer of basal/myoepithelial cells[91], with additional heterogeneity existing within these two cell layers. For example, a functionally distinct subpopulation of mammary stem cells may comprise a small subset of basal cells[6,7], as well as subpopulations of progenitors and mature, hormone-responsive cells defined within the luminal compartment[63]. Technological advances enable us to explore cellular heterogeneity without bias using single-cell RNA sequencing (scRNAseq)[10]. This approach was used to describe the cell types and states within the human[92] and mouse mammary epithelium[82,93,94] and generally yielded three main cell types, namely, basal cells (marked by Krt14); secretory luminal (L-Sec) cells, also called luminal progenitors (marked by Elf5); and mature, hormone-responsive luminal (L-HR) cells (marked by Prlr). Although it is known that these cell types change their transcriptional programs during pregnancy[93], it remains elusive whether additional cellular diversity exists under normal, adult homeostasis. Cellular identity is strongly influenced by the epigenetic wiring of the cell, which is not measurable by scRNA-seq. Instead, these features can interrogated by the assay for transposase-accessible chromatin using sequencing (ATAC-seq) to reconstruct cis/transregulatory elements associated with cellular identity in bulk assays[95] and at the level of single-cell ATAC-seq (scATAC-seq)[11]. This approach provided insights into the differentiation trajectories of the hematopoietic system[96,97] and has elucidated transcriptional regulators of developmental lineages of the fetal mammary gland both using bulk ATAC-seq[98] and scATAC-seq[99]. The goal of the present study is to elucidate the molecular underpinnings mediating cellular identity within the mouse mammary epithelium by integrating single-cell transcriptomics (scRNA-seq) and chromatin accessibility (scATAC-seq) profiling of mammary epithelial cells (MECs). Our combined scRNA-seq/ scATAC-seq analysis revealed luminal progenitor and lactation-committed cell states within the L-Sec cell type and identified cis/trans-regulatory elements associated with cellular identity and luminal differentiation states. Our work provides important insights into the spectrum of MEC identity under normal homeostasis and will serve as a resource to understand how the system changes in cancer.

RESULTS AND DISCUSSION

Single-Cell Chromatin Accessibility Reveals Luminal Epithelial Cell States in the Mouse Mammary Epithelium

Recent single-cell transcriptomics analyses revealed that the MEC system consists of three main cell types—namely, basal (marked by Krt14), L-Sec (marked by Elf5), and mature L-HR (marked by Prlr)—in both human and mammary glands[82,92–94] To determine whether additional cell states exist on an epigenetic level, we used massively parallel, droplet-enabled scATAC-seq analysis (10X Genomics Chromium) on MECs sorted from post-pubertal mice using flow cytometry. We subjected MECs to scATAC-seq analysis in three separate samples, profiling in total 23,338 individual cells (Figure 3.1A). After data processing using the Cell Ranger pipeline (10X Genomics), we performed unbiased clustering on all peaks using Seurat[12], which revealed 4 main clusters (0–3) of MECs and minor populations of contaminating stromal cells (Figure 3.1B; Figures S1A and S1B). To identify the genes accessible in each cell type, we generated a gene activity matrix to serve as pseudoexpression data[56]. This enabled us to identify basal cells (cluster 0; marked by Krt14), L-Sec (clusters 2 and 3; marked by Kit), and L-HR (cluster 1; marked by FoxA1) (Figure 3.1C). We also generated pseudobulk profiles to visualize differentially accessible genomic regions. Wnt10a was found to be specifically accessible in basal cells (Figure 3.1D), whereas Cldn3 displayed one major peak of high accessibility in all three clusters of luminal cells, which was essentially absent from the basal pseudobulk analysis (Figure 3.1E). Interestingly, we observed two distinct clusters within the L-Sec cell type (Figure 3.1C): cluster 2 (marked by Tm4sf1, encoding a tetraspanin transmembrane molecule involved in breast cancer metastasis through regulation of the phosphatidylinositol 3-kinase [PI3K] pathway[100]), and cluster 3 (marked by Rspo1, encoding a regulator of Wnt signaling, R-Spondin 1, that can mediate mammary stem cell renewal[101]). Cluster 3 also showed moderate accessibility of the basal marker gene Krt14 (Figure 3.1C), suggesting that this cell state within L-Sec shows similarity to basal cells, which could indicate a bipotent progenitor cell state that can

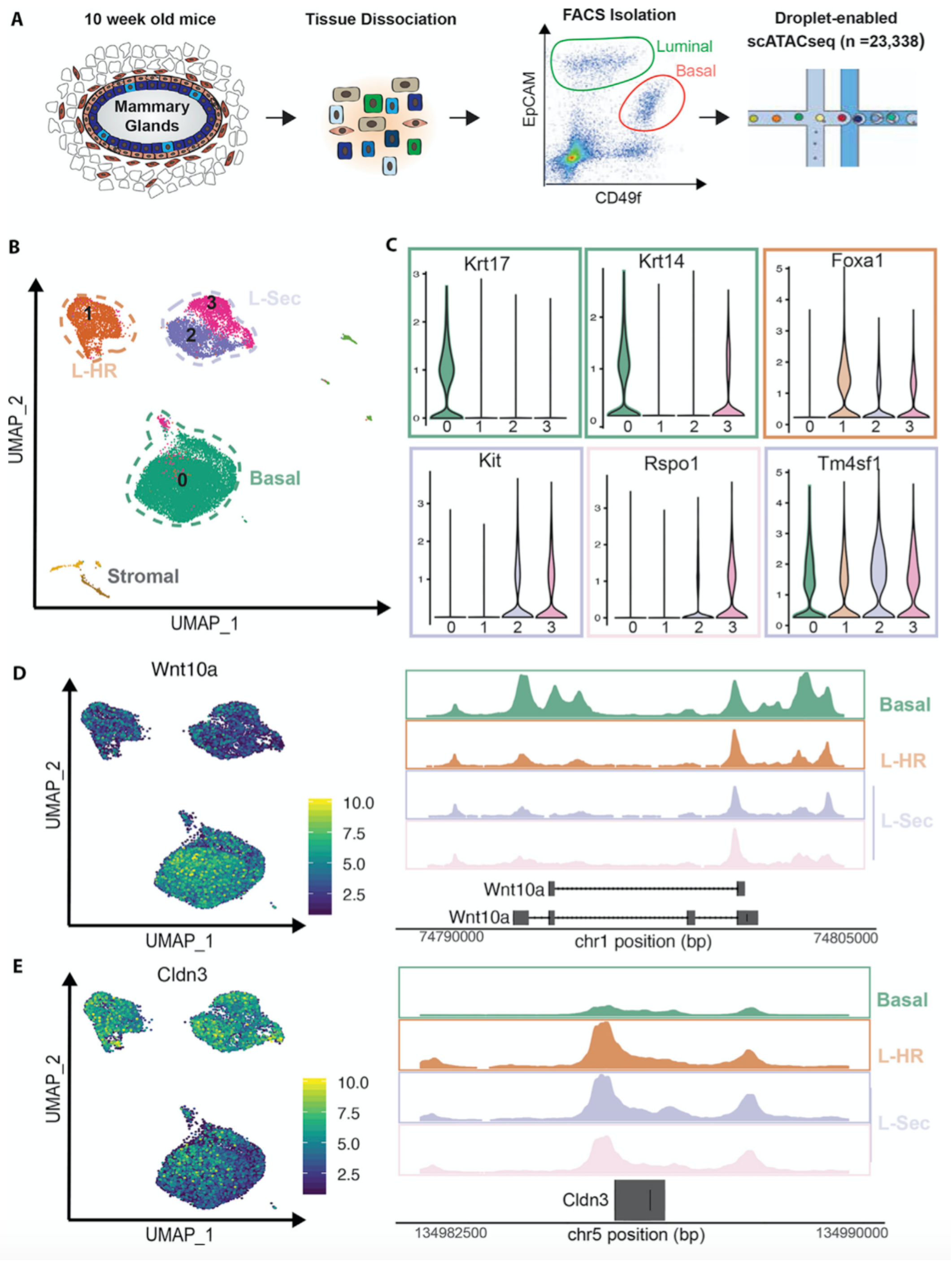


Figure 3.1. Single-Cell Chromatin Accessibility Profiling of MECs from Post-pubertal Mice Reveals Luminal Epithelial Cell States
(A) Schematic of the experimental workflow for scATAC-seq analysis. (B) UMAP visualization of scATAC-seq libraries, colored by Seurat clustering performed on an aggregated peak matrix. Cell types are outlined by dotted lines, with basal cells in green, hormone-responsive luminal (L-HR) cells in orange, and secretory luminal (L-Sec) cells in indigo. (C) Violin plots of Cicero-generated gene accessibility matrix-based marker genes of each cluster, with boxes colored by cell-type-specific accessibility. (D and E) UMAP of scATAC-seq analysis on the left, with cells colored by gene accessibility expression level of Wnt10a and Cldn3. Pseudobulk profiles of library fragments on the right, subset by cluster at genomic regions corresponding to Wnt10a and Cldn3.

differentiate into both basal and luminal lineages or a transitory luminal progenitor that is directly derived from a basal mammary stem cells[6,7]. These initial analyses showed that our scATAC-seq dataset represents a resource to explore the chromatin accessibility landscape in individual mouse MECs.

Defining the Distinct Gene Expression Signatures within Mammary Cell Types and States Using Single-Cell Transcriptomics

To further explore the distinct gene expression signatures underlying the cell states revealed by scATAC-seq, we performed scRNA-seq on fluorescence-activated cell sorting (FACS)-isolated MECs from age- and background-matched, 10-week-old, female FVB/NJ mice, yielding a dataset of 26,859 single-cell transcriptome libraries (Figure 3.2A; Figures 3.3A and 3.3B). Using clustering through Seurat, we detected three main clusters of MECs and their distinct marker genes (Figure 3.2B; Figure 3.3C;) that correspond to basal (Krt14+), L-Sec (Kit/Elf5+), and L-HR (Prlr+), in line with previous single-cell transcriptomics analyses[82,93]. All clusters were evenly composed of cells from all three individual experiments. We detected a small cluster of contaminating stromal cells, minor clusters of proliferating (P) cells (Mki67+), and small clusters expressing both luminal and basal keratins that displayed high levels of genes per cell, suggesting that these represent doublets (D). We detected two distinct cell states within the L-Sec cluster (Figure 3.2B; Figure 3.3D), which emerged as one homogeneous cluster in previous scRNA-seq studies[82,93]. Differential gene expression analysis revealed that one of these clusters was marked by genes associated with milk production, such as Lipa, Csn2, and Lalba, and thus labeled the lactation

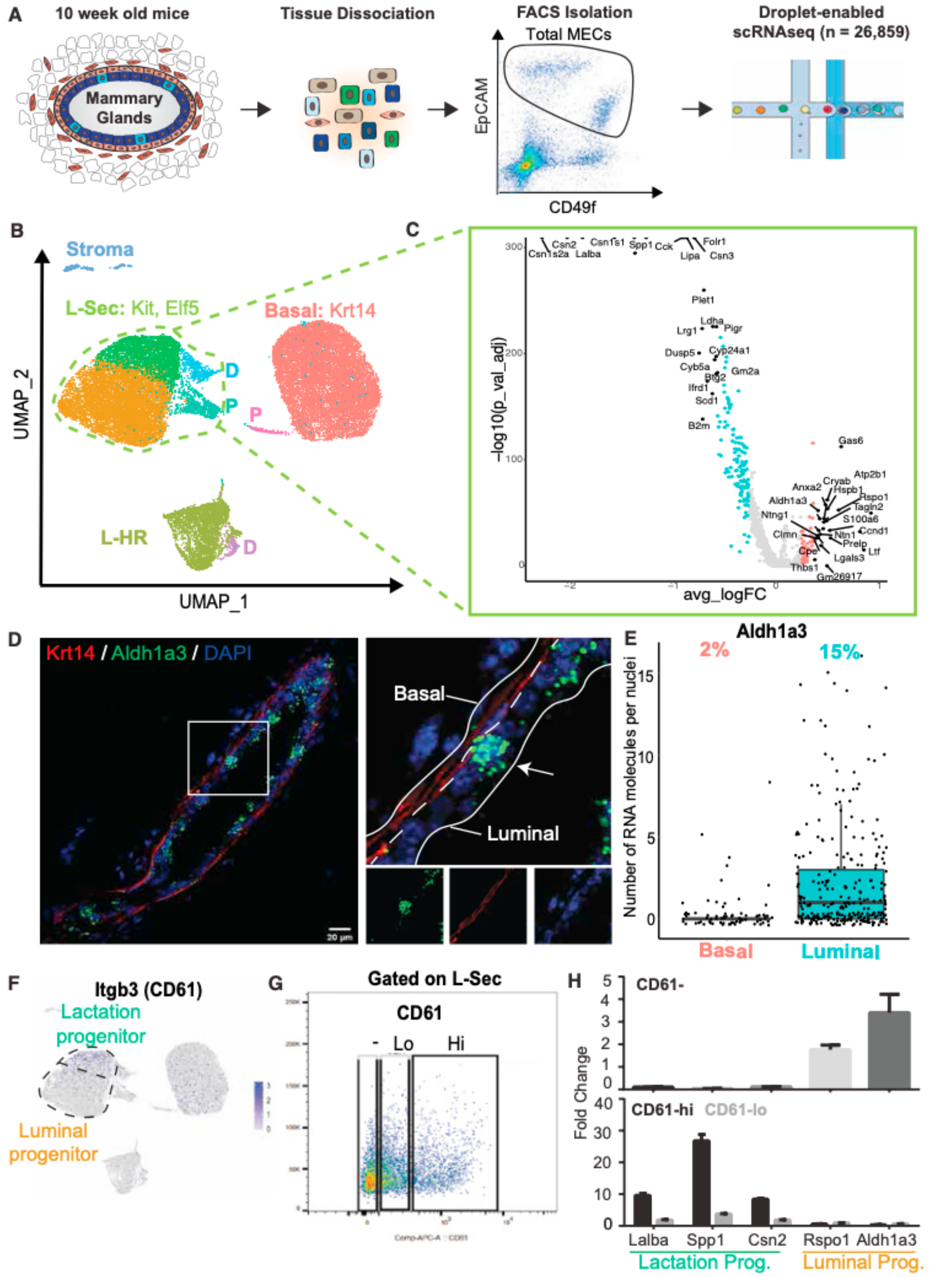


Figure 3.2. Single-Cell Transcriptomics of MECs Reveal the Lactation-Precursor Cell State (A) Schematic of the experimental workflow for scRNA-seq analysis of isolated mouse MECs. (B) UMAP visualization of scRNA-seq libraries anchored by sample, with colors corresponding to unbiased clustering and annotated by cell type and state. Basal cells are in red, L-HR cells are in light green, and L-Sec cells are outlined in dark green. Putative doublets are marked by D, and proliferative cells are marked by P. Within the L-Sec cell type, two distinct clusters emerged that were labeled mature or progenitor based on gene expression signatures. (C) Volcano plot showing genes that are differentially expressed between L-Sec luminal progenitor and lactation progenitor cells. (D and E) Fluorescence images from *in situ* RNAscope analysis for *Aldh1a3* in combination with immunostaining for basal-specific KRT14 are shown. Luminal and basal compartments are outlined in the blown-up image. Quantification of transcript counts per basal and luminal cells is shown; data were combined from three independent regions of mouse mammary gland sections. (F–H) Validation of two distinct cell states using flow cytometry. (F) Feature plot showing gene expression of *Itgb3* encoding CD61. (G) Flow cytometry analysis of primary mouse MECs gated on L-Sec cells only showing levels of CD61 ranging from negative () to low (lo) and high (hi). (H) Gene expression of marker genes from scRNA-seq analysis defining luminal progenitors and lactation progenitors measured in CD61, CD61-lo, and CD61-hi cells using qPCR. The error bar indicates inter-assay variability as SEM from n = 3 experiments.

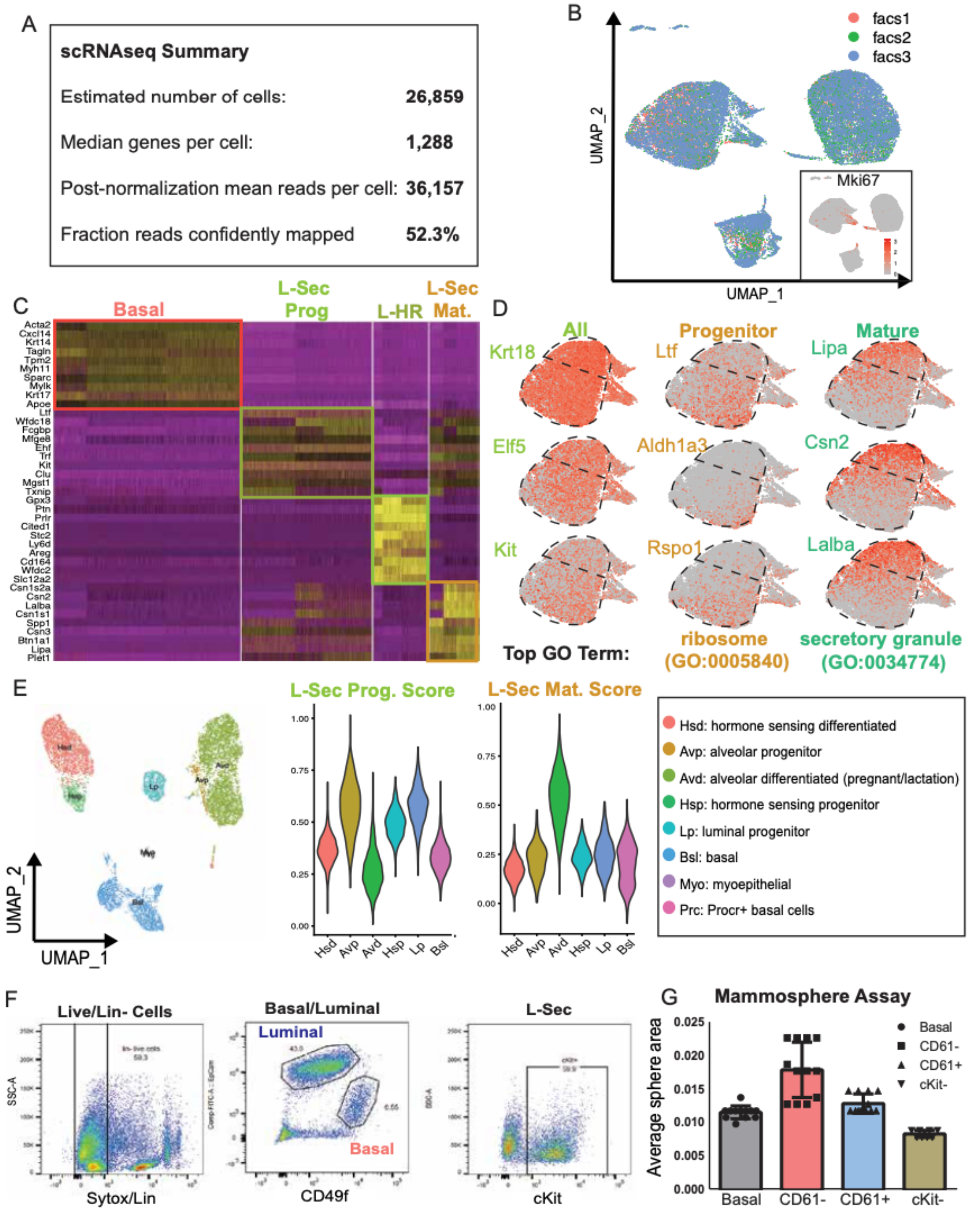


Figure 3.3: scRNAseq quality control and cell type identification (A) Sequencing and alignment metrics for the three scRNAseq libraries. (B) UMAP of scRNAseq analysis, with cells colored by library of origin. Proliferative cells are marked by high *Mki67* gene expression, with dark red corresponding to high expression and light grey to low. (C) Marker gene heatmap of clusters corresponding to epithelial cells, with yellow corresponding to high expression and purple with low. (D) Focused analysis of L-Sec cluster and corresponding marker gene is shown, where expression is scaled such that dark red corresponds to high expression of the gene and light grey corresponds low expression of the gene in question; top GO-term including accession number is listed for Progenitor and Mature cell states. (E) Bach et al. analysis of NP and G stage mouse scRNAseq data, with UMAP reduction colored by their cell type labels, scoring of our scRNAseq derived L-sec progenitor and lactation progenitor gene signatures visualized via ViolinPlot. (F) Summary of flow cytometry strategy to specifically gate on L-Sec cells from primary mouse mammary epithelial cell preparations. (G) Quantification of mammosphere formation assay using CD61- and CD61+ L-Sec cells quantified by mammosphere size after 4 days of culture (n = 3). Error bar indicates SEM. Difference between CD61- and CD61+ was statistically significant (t-test: p<0.05).

progenitor, whereas the second cluster expressed high levels of genes associated with general luminal progenitor cell capacity, including Aldh1a3[102] and Rspo1, and therefore labeled the luminal progenitor (Figure 3.2C). Mature alveolar luminal cells arise during pregnancy and lactation[91]. Because our dataset was generated from nulliparous mice, we hypothesized that lactation progenitors represent a subset of lactation-precursor cells even before pregnancy. To corroborate this, we explored an scRNA-seq analysis of mouse MECs from nulliparous, pregnant, and lactating mice (Figure 3.3E)[93]. We performed gene scoring analysis using our luminal progenitor and lactation progenitor gene signatures, which revealed that alveolar and luminal progenitors correspond to our luminal progenitor cluster, whereas differentiated alveolar cells from pregnant mice are highly comparable to our lactation progenitor cell state. Because Aldh1a3 marks a subset of luminal-restricted progenitor cells[102]we next used Aldh1a3 as a marker for in situ validation of this cell state. Using a specific RNA-based probe (RNAscope) for Aldh1a3, in combination with anti-KRT14 antibody staining to label the basal cell compartment, we detected a subset of luminal epithelial cells (KRT14-negative) with pronounced expression of Aldh1a3 located in both ductal and lobular regions of the mammary gland (Figure 3.2D). Quantification of cells with more than 5 transcripts per cell revealed ~15% of Aldh1a3+ in the luminal compartment detected by RNAscope (Figure 3.2E), which was in line with our scRNA-seq results showing ~13% of Aldh1a3+ luminal cells. We also found that the cell surface marker CD61 (Itgb3), which is known to mark luminal progenitors[103], is increased in lactation progenitor cells (Figure 3.2F). Using flow cytometry, we isolated cKit+/CD61+ and cKit+/CD61 MECs for further validation by qPCR of lactation progenitor genes (Figure 3.2G). In line with our scRNA-seq data, we found that cKit+/CD61+ cells express higher levels of the lactation associated genes Lalba, Spp1, and Csn2, whereas cKit+/ CD61 cells showed expression of Rspo1 and Aldh1a3, as detected in luminal progenitor cells (Figure 3.2H). Altogether, these findings confirmed the existence of two distinct states within the L-Sec cell type as predicted by scATAC-seq and allowed us to integrate these results with the previously proposed functional designations as luminal progenitor and lactation progenitor L-Sec cells.

Pseudotemporal Analysis Reveals Continuous Trajectory from Luminal Progenitor to Lactation Progenitor Cells

We next used pseudotemporal ordering using Monocle 3 pseudotemporal analyses[79] to reconstruct the lineage dynamics between luminal progenitors and lactation progenitors. In particular, we wanted to answer whether these are distinct cell states resulting in a loosely connected trajectory or whether they form a continuum with progenitor and lactation-precursor cell states at both ends of the spectrum. First, focusing on our scRNA-seq data, we generated a graph trajectory through the L-Sec cluster in uniform manifold approximation and projection (UMAP) space, which revealed one main trajectory connecting luminal progenitor and lactation progenitor cells with minor paths branching off to each side (Figure 3.4A). To learn more about the different regions in pseudotime, we divided the cells into 10 bins based on their position along the trajectory for further interrogation (Figure 3.4A). Using gene signatures from the Bach et al. (2017) dataset for luminal progenitor scores (LP scores) from nulliparous mice and alveolar differentiated scores (Avd scores) from pregnant mice, we found that L-Sec cells

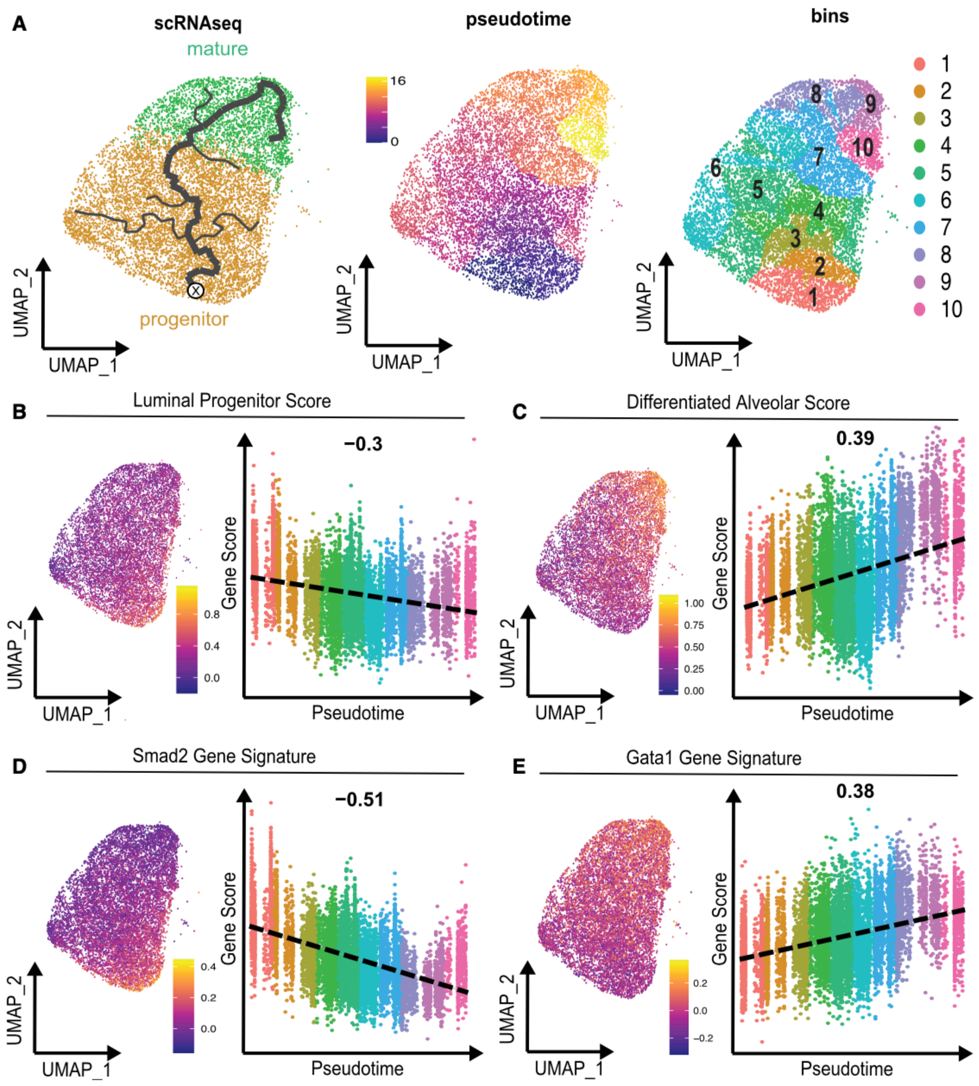


Figure 3.4. Pseudotemporal Analysis Shows a Continuous Differentiation Trajectory within L-Sec Cells (A) UMAP reduction of the scRNA-seq subset on L-Sec cells only colored by Seurat cluster with Monocle 3 pseudotemporal trajectory overlay, with the boldface path representing the major transitionary graph from luminal to lactation progenitors. UMAP reduction is shown (left plot), with cells colored by pseudotime with dark blue corresponding to early and light yellow corresponding to late (middle plot). The right plot shows UMAP reduction colored by pseudotime bin, with 1 as the earliest and 10 as the latest. (B) UMAP reduction colored by the LP score derived from Bach et al. (2017) (left), and a feature scatter showing individual cell gene scores colored by the pseudotime bin score, with the dotted line indicating the associated Pearson correlation (right). (C) UMAP reduction colored by the Avd score from cells in pregnant mice (left) derived from Bach et al. (2017), and a feature scatter showing individual cell gene scores colored by the pseudotime bin score, with the dotted line indicating the associated Pearson correlation (right). (D) UMAP reduction visualizing the Smad2 downstream target gene expression score, in which cells colored dark blue have low scoring and cells colored light yellow have high scoring, and a feature scatter colored by the pseudotime bin of score versus pseudotime and the associated Pearson correlation (right). (E) UMAP reduction visualizing Gata1 downstream target gene expression score, in which cells colored dark blue have low scoring and cells colored light yellow have high scoring, and a feature scatter colored by pseudotime bin of score versus pseudotime and associated Pearson correlation (right).

form a continuous gradient from luminal progenitors with high LP scores and low Avd scores to lactation progenitors with low LP scores and high Avd scores (Figures 3.4B and 3.4C), indicating that these cells exist on a continuum rather than in distinct states of progenitor and lactation-precursor L-Sec cells. Focusing on our scATAC-seq data, we used Cicero[104] to generate a subset L-Sec UMAP reduction for subsequent pseudotemporal analysis and binning as described earlier. This revealed a main trajectory connecting luminal and lactation progenitor cells similar to our scRNA-seq analysis (Figure 3.5A). To identify modules of genomic peak regions in the scATAC-seq that are coaccessible and vary through pseudotime, we employed the CisTopic pipeline to calculate topics of coaccessibility[105]. We found several topics to be dynamically correlated with pseudotime; for example, topic 5 showed accessibility early in pseudotime, whereas topic 1 represented features that were accessible late in pseudotime (Figure 3.5B). To link transcriptional and chromatin accessibility dynamics, we next analyzed these specific topics using HOMER[106] to test for significant representation of the transcription factor (TF) binding motifs contained within. We then used our scRNA-seq dataset to generate expression modules that are differentially expressed along the major trajectory in pseudotime and performed Gene Ontology (GO) term analysis for TF signaling outputs using Enrichr[75] to compare these with the TF motifs identified by HOMER on the scATAC-seq level. Interestingly, we found that Smad2 motif accessibility and Smad2 downstream gene expression were high early and gradually decreased in pseudotime, whereas TF motif accessibility and downstream gene expression associated with GATA1 started low early and then increased later in pseudotime (Figure 3.4D and 3E; Figure 3.5). Smad family TF motifs are key

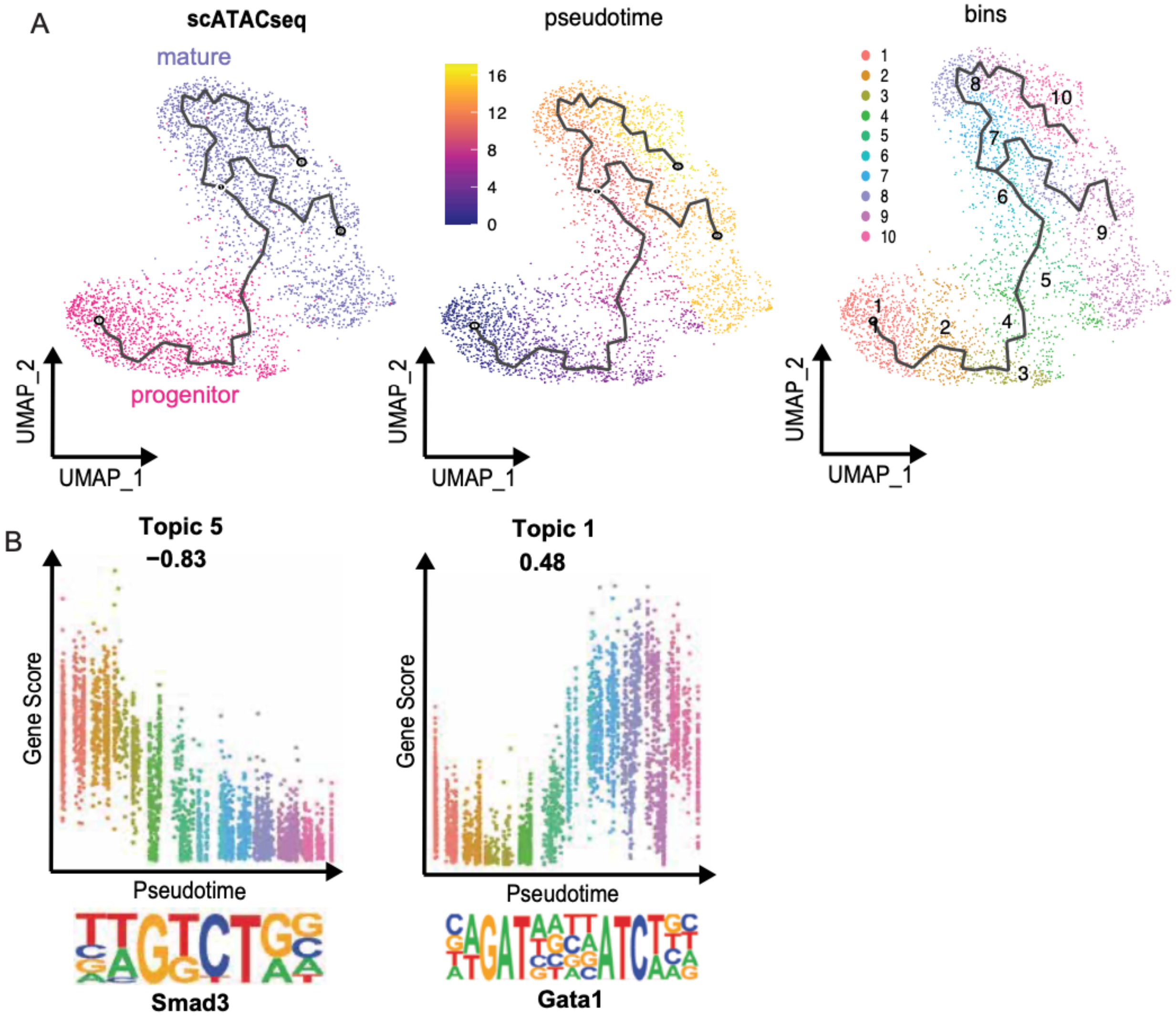


Figure 3.5: scATACseq-based pseudotemporal analysis in secretory luminal compartment (A) UMAP reductions and corresponding pseudotemporal trajectory overlay for secretory luminal cells from scATAC analysis, with cells colored by Seurat clusters. (A) UMAP reduction with cells colored by pseudotime, with dark blue corresponding to early and light yellow to late. (A) UMAP reduction with cells colored by pseudotime bin, with 1 as the earliest and 10 as the latest. (B) Feature scatter plots of Topics 5,7,and 1 probability vs Pseudotime, with cells colored by position in pseudotime and associated Pearson correlation. (B) Topic's 5 and 1 showed enrichment of the Smad2 and Gata1 binding motifs respectively highlighted below their feature scatters.

mediators of transforming growth factor b (TGF-b) signaling[107], indicating that this pathway is active in luminal progenitor cells. However, GATA signaling is generally associated with luminal differentiation[108]. Altogether, our findings support a continuous transition between L-Sec progenitor and lactation progenitor cells and highlight several chromatin accessibility changes and potential transcriptional regulators associated with this transition.

Integration of scRNA-Seq and scATAC-Seq Reveals Cell-Type-Specific Transcriptional Regulators and cis and trans-Regulatory Elements

We next sought to integrate our scRNA-seq and scATAC-seq datasets to gain deeper biological understanding about the link between chromatin accessibility and gene expression within MECs. To this end, we used an approach to anchor diverse datasets together for comprehensive integration of single-cell modalities[56]. This integrated object yielded consistent overlap between modalities within each of the main cell types and recapitulated the two clusters of luminal and lactation progenitor cells within the L-Sec cell type (Figures 4A and 4B; Figure 3.7A). Known hallmark genes for mammary cell types (e.g., Krt5, Krt8, Kit, and Foxa1) showed strong correspondence between chromatin accessibility and gene expression in this integrated analysis (Figure 3.7B). We observed overall high correlation between ATAC-seq and RNA-seq data (Figure 3.7C). In particular, we observed striking consistency for Rspo1 in progenitor cells and Lalba in mature L-Sec cells in terms of chromatin accessibility paired with gene expression (Figure 3.6B). We sought to use this integrated analysis to identify TFs that may be critical for regulating cell-type identity. We used the ChromVar analysis pipeline[109] to analyze accessibility of cell-type-specific TF motifs in our scATAC-seq dataset. Using Seurat's marker gene test on the resultant TF motif deviation matrix, we uncovered sets of cell-type specific TF motif enrichments (Figure 3.6A). We then performed cocorrelation analysis to pinpoint TF modules in the MEC system (Figure 3.7D), which revealed three major modules. Module 1 contained predominantly Jun and Fos-related TF motifs, indicating that this feature is related to a subset of cells showing stress response, most likely because of tissue dissociation and FACS isolation. Module 2 contained numerous TFs previously associated with basal epithelial biology, such as Tp63[110], but Gata3 and other Gata family TFs were also observed, which have been linked with regulation of luminal cell fate decisions[108]. Finally, module 3 contained mostly TFs associated with luminal epithelial biology, such as Foxa1[111] and Elf5[112], but also included a cluster of epithelial-to-mesenchymal transition (EMT)-related TFs, such as Tcf4, Snai2, and ID4[113]. Next, we devised cell-type-specific TFs displaying both motif accessibility and active downstream target gene expression as determined by Enrichr analysis. Reassuringly, the master regulator of basal cell biology, Tp63[110], emerged as one of the top TF motifs that was specifically accessible in basal cells and showed distinct gene expression as calculated using the gene score for a set of TP63 target genes (Figure 3.6C). Several SMAD TFs yielded top motif scores within basal cells; however, SMAD3 showed the highest target gene expression scores in basal cells, indicating that SMAD3 represents a key TF in the regulation of basal cell identity. SMAD family TFs are critical mediators of transforming growth factor b1 (TGFb1), which has wide implications in regulating mammary biology and cancer[114]. SMAD TFs also showed increased activity in L-Sec

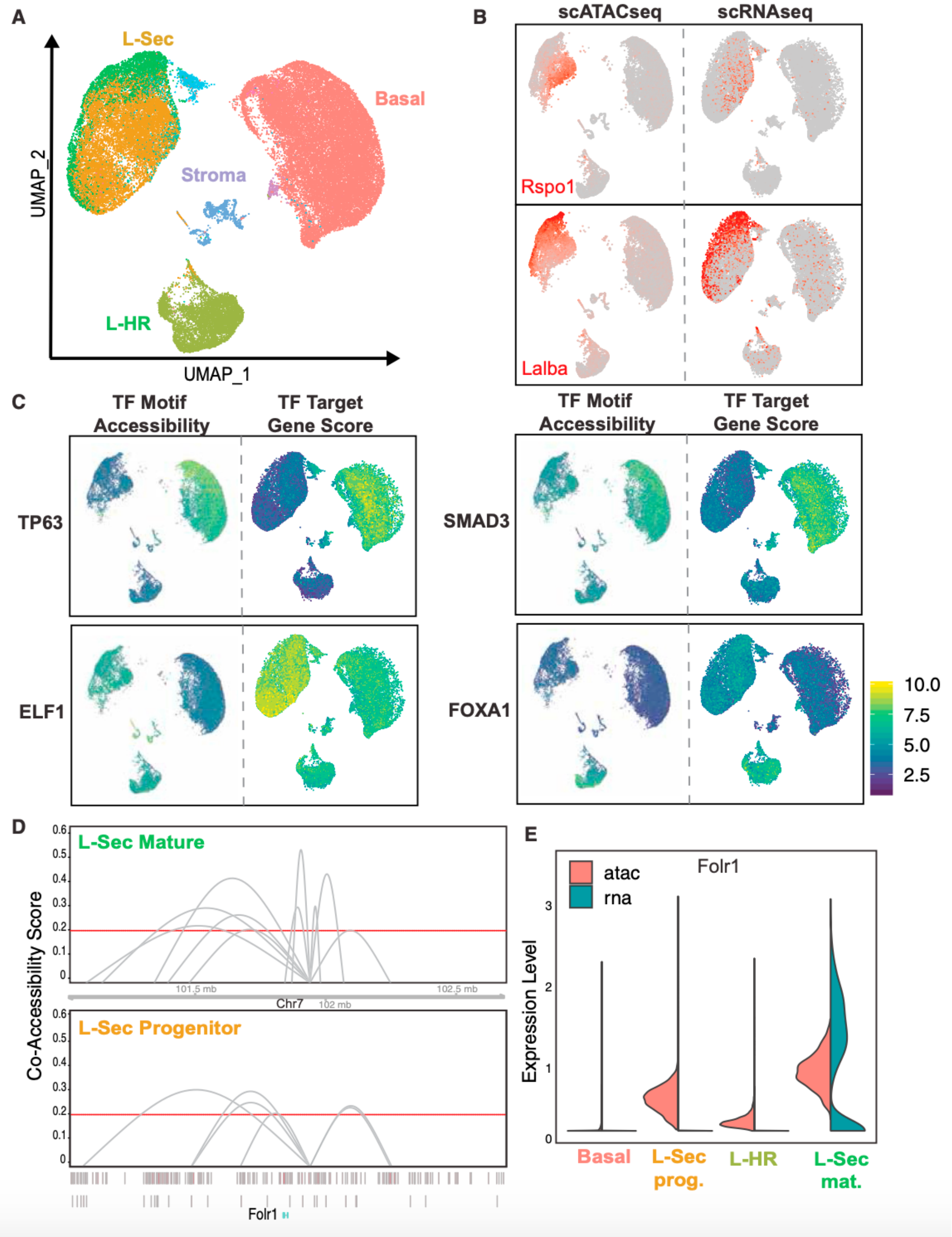


Figure 3.6. Integration of Single-Cell Chromatin Accessibility and Transcriptomics Datasets (A) Coembedding of scRNA-seq and scATAC-seq data into a single UMAP visualization, with cells colored by Seurat cluster or label-transferred cluster. (B) Coembedded UMAP faceted by technology type, with cells from scATAC-seq libraries on the left and cells from scRNA-seq libraries on the right. Cells are colored based on scaled expression, with gray corresponding to low expression and dark red corresponding to high expression. (C) Faceted UMAP visualization of coembedded analysis, with scATAC-seq cells on the left and scRNA-seq cells on the right. scATAC-seq data are colored by scaled deviations of TF motif accessibility, and scRNA-seq data are colored by gene scoring of downstream targets of TF signaling as annotated through GO terms. Yellow corresponds to high values, and dark blue to corresponds low values. (D) Cicero connection data at enhancer region chr7_101932449_101936345 generated by subset analysis by cluster. Connections from lactation progenitor cells are shown in the top panel, and connections from the L-Sec progenitor are shown in the bottom panel, with a minimum coaccessibility score of 0.2 visualized. (E) Violin plot ofFolr1expression in the coembedded analysis, split by technology type.

progenitors (Figure 3.4E; Figure 3.5B), which highlights the connection between basal and L-Sec progenitor cells. ELF1 showed the highest motif accessibility in both luminal clusters (L-Sec and L-HR); however, expression of ELF1 target genes is most predominantly detected in L-Sec cells. Finally, we explored FOXA1 as a known regulator of luminal differentiation, which showed strong correspondence between high TF motif accessibility and elevated target gene expression scores specifically in L-HR cells, corroborating the notion that FOXA1 is a master regulator of the L-HR cell type[115]. To identify cis-regulatory elements that may contribute to celltype distinction, we used the Cicero pipeline for coaccessibility analysis to determine cell-type-specific genomic connections[104]. The resulting connections were subset to those in which one peak of each pair corresponded to an enhancer region from EnhancerAtlas's mouse mammary putative enhancer list[116]. Directly comparing L-Sec cell states, we found enhancer-specific connections near the Folr1 locus that were specific to lactation progenitors, but not luminal progenitors (Figure 3.6D). Further interrogation of gene expression and chromatin accessibility revealed the specific signal for Folr1 in L-Sec lactation progenitors (Figure 3.6E). Folr1 has been identified as a putative regulator of milk protein synthesis in cow mammary glands (Menzies et al., 2009), which is in line with the notion that this cluster is a lactation-committed precursor (Figure 3.4 2C). Altogether, this suggests that this enhancer region on chromosome 7 represents a key regulatory element that becomes active during lactation-precursor differentiation. Altogether, our integrated single-cell transcriptomics and chromatin accessibility analysis of the MEC system revealed a cell-state hierarchy within the luminal epithelial compartment and defined transcriptional and epigenetic underpinnings regulating cellular identity in the mammary epithelium. In particular, we define distinct maturation states within L-Sec cells, which exist on a continuum ranging from general luminal progenitor cells (Rspo1 and Aldh1a3) to potentially lactation-committed progenitor cells (Lalba and Csn2). By directly integrating transcriptomics and chromatin accessibility datasets, we provide a framework to devise putative key TFs by combining motif accessibility with positive downstream target gene expression. We also identified enhancer regions that are systematically associated with gene accessibility and expression of effector genes associated with L-Sec differentiation (Folr1). Our findings lay the groundwork for future studies to functionally address the biological significance of these cis/trans-regulatory elements in mediating mammary stem and progenitor cell function and to determine how the chromatin accessibility landscape changes during breast cancer.

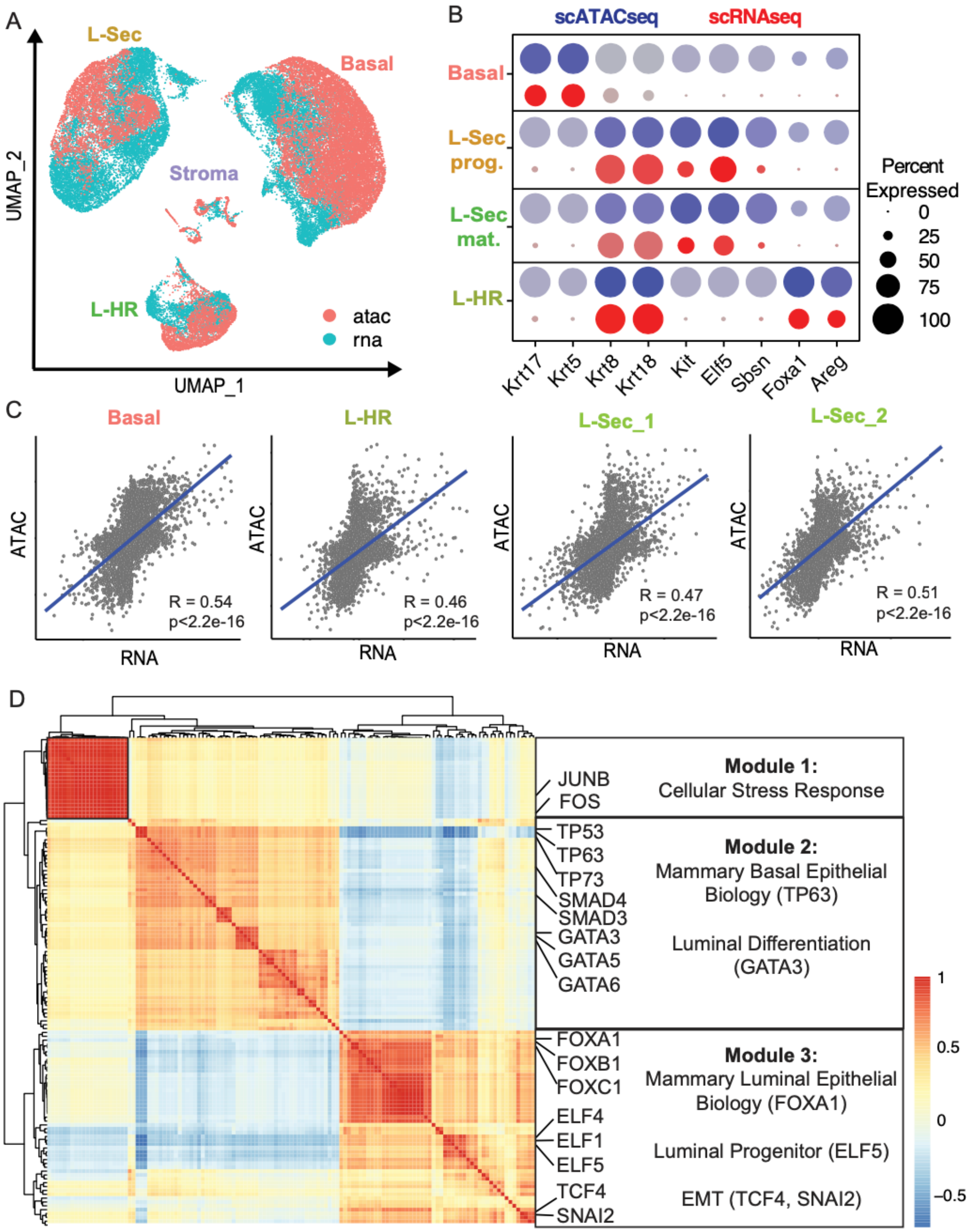


Figure 3.7: Data integration quality control and TF motif co-correlation analysis (A) Coembedded UMAP representation of both scATACseq cells and scRNAseq cells, with colors corresponding to the data type of origin. (B) Split Dot Plot of cell type markers, with imputed RNA expression intensity in the scATACseq cells scaled from grey to dark blue, and expression intensity in scRNAseq cells displayed in a scale from grey to dark red. The size of the dot corresponds to the percentage of cells within that cluster have positive expression of the gene. (C) Correlation of gene activity matrix from scATACseq cells and gene expression in cells from scRNAseq, split by cell type. (D) Heatmap displaying co-correlation of TF motif accessibility shown as z-scores (blue = low; red = high). Transcription factors were subset to those that were found through logistic regression to be significantly associated with a particular cluster post label transfer from scRNAseq data onto scATACseq data, and had an average log-fc greater than one. Key TFs are highlighted in relation to putative function on the box to the right.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice For sequencing, FVB/NJ mice are from Jackson Laboratory (Stock Number: 001800) were employed. In both scRNAseq and scATACseq experiment, 10 weeks old female mice were used for tissue collection. For RNAscope experiments, 10-week old C57BL/6 mice from Jackson Laboratory (Stock Number: 000664) were used. 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.

METHOD DETAILS

Cell Isolation and single-cell RNA and ATAC sequencing library generation

Mammary glands number 4 were collected and pooled from a total of four 10-week old, female FVB/NJ mice. Glands were minced into pieces ~1mm in diameter and processed as previously described[117]. In brief, minced glands were incubated with a 2mg/ml collagenase type IV solution at 37C while shaking for 1 hour. Digested organoids were collected by differential centrifugation. Collected organoids were further dissociated with trypsin into single cells. Cells were stained for flow cytometry using fluorescently labeled antibodies for CD49f, EpCAM, CD31, CD45, Ter119, and SytoxBlue. For scRNAseq, live epithelial cells were collected for sequencing. For scATACseq, basal and luminal cells were collected separately. Library generation for 10x Genomics v2 chemistry was performed following the Chromium Single Cell 30Reagents Kits v2 UserGuide: CG00052 Rev B. Library generation for single cell ATACseq were performed following the Chromium Single Cell ATAC Re-agent Kits User Guide: CG000168 Rev B. Single cell RNAseq and ATACseq libraries were sequenced on the Illumina HiSeq4000 plat-form targeting approximately 50,000 reads per cells.

Validation by qPCR

Mammary Glands number 2, 3, 4, and 5 were collected and combined from a total of four 13-week old, female FVB/NJ mice. Mammary glands were processed with the same procedure as those isolated for scRNAseq. Cells were stained for flow cytometry using fluorescently labeled antibodies for CD49f, EpCAM, cKit, CD61, CD31, CD45, Ter119, and SytoxBlue. Gates were set to sort out and collect CD61-, CD61lo, and CD61+ cells from the cKit+ luminal epithelial population. Directly after sorting, RNA was collected using a Quick-RNA Microprep RNA isolation kit (Zymo Research: R1054). The extracted RNA was immediately processed into cDNA using an iScript cDNA Synthesis Kit (Biorad: 1708891). qPCR reactions were performed using PowerUp SYBR Green Master Mix (AppliedBiosystems: A25742) and Ct values were normalized to Gapdh Ct values.

Mammosphere assay

Mammary Glands number 2, 3, 4, and 5 were collected and combined from a total of four female FVB/NJ mice in triplicate (10- 13 weeks in age). Mammary glands were processed with the same procedure as those isolated for scRNAseq. Cells were stained with the same panel as for the qPCR validation, and gates were set to sort and collect basal cells, CD61- and CD61+ cells from the cKit+ luminal population, as well as cKit- luminal cells. Each cell type was then individually resuspended in complete Epicult- B Mouse Medium (Stemcell: 05610) medium and mixed 1:1 with Matrigel (Corning: 354230). Cells were plated in the center of individual wells on a 24-well cell culture plate (Genesee: 25-107) at a density of 10,000 cells/well, in a final cell-containing Matrigel solution volume of 40uL/well. The cell-containing Matrigel was solidified for 15 mins in a humidified 37C cell culture incubator with 5% CO2. Each well then had 1mL of Epicult media added to it. Mammospheres were cultured in a humidified 37C cell culture incubator with 5% CO2 for a total of 7 days.

Mammosphere image analysis

Brightfield 2x magnification z stack images of the mammosphere cultures were taken using a Keyence Microscope on day 4 and day 7 of culture. The full focus z stack images were then analyzed using ImageJ v1.52p software. Each image was converted to binary with the ''Make Binary'' function, then underwent ''Fill Holes'' and ''Watershed'' processing. To count the number and average area of spheres in each image we then used the ''Analyze Particles'' feature. A lower sphere area threshold was set to 0.005 inch2 (smallest size that still appears to be a real sphere in the image) for every image, and the upper sphere area threshold was determined by measuring the area of the largest sphere in each image. Circularity was set to 0.50-1.00. Each condition in every experiment is rep- resented by four images (quadruplicate), each from an individual well.

Sequence alignment and data processing

Alignment of scRNAseq analyses was completed utilizing 10x Genomics Cell Ranger pipeline (version 2.1.0). Alignment of scATAC seq analyses was completed utilizing 10x Genomics Cell Ranger ATAC pipeline (version 1.1.0). Each library was aligned to an indexed mm10 genome using Cell Ranger Count and Cell Ranger ATAC Count. ''Cell Ranger Aggr'' function was used to normalize the number of confidently mapped reads per cells across the libraries from different libraries for scRNAseq and scATACseq separately.

Cell-type clustering analysis and marker identification using Seurat

The aggregated peak-by-cell data matrix was read into R (R version 3.6.0) and processed using the Seurat single cell analysis pack- age version 3.0.2[12]. Along with the peak matrix, the Cicero-generated gene activity matrix (see below) and ChromVar deviations score matrix (see below) were added as assays to the Seurat object. A quality control cutoff of a minimum of 2500 fragments per cell was applied to trim the dataset of low-quality cells. Next, variable features of the peak matrix were set to peak regions of > 100 across the matrix. These variable features were used to perform Latent Semantic Indexing (LSI), and the first 50 components were calculated. These components were then used to generate a Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction. Post UMAP, a Shared-Nearest-Neighbor graph was generated from the first 14 LSI components chosen via the elbow plot method and was used to cluster the cells via Seurat's Louvain algorithm.

Marker genes for peak-based clustering were generated using Seurat's default FindAllMarkers() function on the gene activity matrix. Pseudobulk profiles by cluster highlighting fragment stack ups at particular genomic regions were generated using Signac (version 0.1.0).

Post label transfer, cell type-specific transcription factor motifs were calculated using the logistic regression method option implemented in Seurat's FindAllMarkers() function. Those TF motifs that had an average log fold change greater than one were used to generate the correlation heatmap to find co-correlated modules of transcription factor motif enrichment.

Single-cell RNAseq analysis

Each of the scRNAseq data libraries were independently read into R version 3.6.0 and processed using the Seurat pipeline version 3.0.2. Genes had to be expressed in at least three cells to be considered for analysis. Cells were trimmed to those that had at least 200 minimum unique genes expressed, no more than 6000 unique genes, and less than 30% of counts aligning to the mitochondrial genome. Libraries were anchored and integrated using the top 2000 variable features per library calculated via the "vst" method in Seurat. Canonical correlation analysis (CCA) on these 2000 features between the libraries was calculated, and the first 20 dimensions used as input for anchoring. Post anchoring, PCA was performed and the first 10 PC's were used for UMAP dimensionality reduction and subsequent clustering using the default Louvain implementation. Marker genes per cluster were calculated using Seurat's Find AllMarkers() function and the "wilcox" test option. GO term enrichment was performed using Enrichr[75].

Gene activity matrix generation

The aggregated peak-by-cell data matrix was read into R version 3.6.0, binarized, and processed with the Cicero analysis package version 1.2.0 and the monocle 3 alpha version 2.99.3 to generate a gene activity matrix for all cells sequenced in the study. The generation of the matrix took into account not only fragments that aligned to regions proximal to the promoter site of each protein coding gene in the genome took into account peak co-accessibility scores also generated through Cicero for all cells to factor in distal genomic relationships to the promoter site of each gene.

Single-cell ATACseq analysis using cisTopic

After cell filtration, the binarized matrix was inputted in an R package, cisTopic(v.0.2.2), to cluster the ATAC-seq data and analyze the chromatin accessibility difference among cell groups. It generated probabilities of a region–topic distribution and topic–cell distribution which were calculated using a latent Dirichlet allocation model with a collapsed Gibbs sampler. Regions were identified as associated with certain topics by automatically selecting a probability threshold based on a fit of the region scores to a gamma distribution.

Cis-regulatory regions by cluster

Post label transfer, scATACseq cell libraries were subset by their predicted ID label, whereupon the Cicero pipeline was utilized on each subset. Co-accessibility networks were generated, with pairs of peak regions and their corresponding score in a data frame. This data frame was subset to only those pairs that overlapped with regions in the EnhancerAtlas mouse mammary list as the first peak of the connection[116]. This trimmed connection matrix was then thresholded for each cell type to those that had a co-accessibility score greater than 0.2. Next, the second non-enhancer peak in the pair was annotated to its closest protein coding gene. Conserved expression markers between technology (RNA and ATAC in the RNA-imputed matrix) were found by cell type and the respective co-accessible gene regions that were both highly connected to an enhancer region, and represented a marker for a cell type were selected.

Transcription factor (TF) motif analysis using ChromVar

Motif enrichment analysis was performed using an R package ChromVAR version 1.4.1[109]. Open chromatin peaks and read counts at open chromatin were defined by the Cell Ranger pipeline as described above. After correction of GC bias, TF deviation score was calculated using a total of 579 TF motif position weight matrices provided with the 10X Genomics Cell Ranger package. For TF clustering analysis, only cells corresponding to epithelial clusters post label transfer (0,1,2,3) were selected. TF enrichment scores were averaged by cluster and hierarchically clustered using hclust( ) and pheatmap( ) in R.

Combined scATACseq and scRNAseq analysis

To generate a coembedding of cells from both scATACseq and scRNAseq libraries, cells from the scRNAseq analysis were used as a reference dataset to predict cluster labels in the scATACseq dataset and transfer them. This prediction used the variable features of the scRNAseq analysis on the RNA assay, and the gene activity matrix of the scATACseq analysis as the query data. Transfer anchors were learned using FindTransferAnchors( ) and the cluster labels were predicted using the TransferData( ) function together with the peak-based scATACseq LSI reduction as the weight.reduction function option input. Next, an imputed gene activity matrix was generated by using the TransferData( ) function again, with the previously learned transfer anchors and a matrix consisting of only the variable features of the scRNAseq analysis and its corresponding cells as the reference. This imputed expression matrix was then used to merge the two Seurat objects, allowing for co-visualization of cells labeled by the scRNAseq cluster labels or their predicted cluster labels for the scRNAseq based or scATACseq respectively.

For combined TF motif accessibility and target gene expression analysis, we first identified cell type-specific TF motifs in our ChromVar analysis (see above), and then performed Enrichr analysis using cell type marker genes from scRNAseq to identify ‘‘ENCODE and ChEA Consensus TFs from ChIP-X’’ for each cell type. Transcription factor targets came from the Enrichr analysis, where marker genes by cluster were analyzed and those genes that had pathway hits in the "ENCODE and ChEA Consensus TFs from ChIP-X" annotation for particular transcription factors were used to score all cells using Seurat's AddModuleScore( ) function.

Pseudotemporal Analyses

For the scRNAseq analysis, using R version 3.6.3, cells pertaining to clusters 1 and 3 (L-sec progenitor and lactation progenitor) were subset into their own raw counts data matrix. This matrix was then processed using Monocle 3 version 0.2.1 functionality. Using the subset UMAP cell positions from the scRNAseq analysis, we next employed monocle to learn a graph trajectory through this space. The beginning of pseudotime was chosen as the branch node that started within our L-sec progenitor cluster. Cells were then binned into 10 groups based on their positions along pseudotime. To further explore the branch of the trajectory that traveled from the start of pseudotime toward the lactation progenitor population, we manually selected cells along the branch for subset analyses. We performed differential expression as a function of pseudotime and clustered the genes from the output into expression modules that varied along the trajectory. These modules were then each separately entered into Enrichr[75] to interrogate TF downstream signaling genes, which were then compared to scATACseq analysis.

For the scATACseq pseudotemporal analysis, we first used Cicero (version 1.3.4.8, built of the aforementioned monocle 3 version), to project clusters 2 and 3 in a subset analysis. Contaminating cells were removed, and the resulting peak region matrix was binarized and processed using cicero. A UMAP dimensionality reduction was calculated and used as the basis for learning the graph trajectory. The beginning of pseudotime was selected as the branch point harboring progenitor cells as previously annotated. An identical binning approach was then applied to the cells in the analysis as described above for scRNAseq data. Seeking an analogous com- parison to the gene expression modules generated in the RNA analysis, we employed cisTopic to generate topics of peak regions that we could then visualize using our binned pseudotime designations to observe which topics had an increased probability at different positions along the graph. The regions associated with each topic were output as bed files containing the genomic coordinates by getBedFiles function in cisTopic. The bed files were used as input into the findMotifsGenome command of HOMER(v4.7). The size parameter was set to 200 and the repeat-masked sequence was used. Mm10 was used as the reference genome. HOMER screened its library of known motifs against the input regions and background for enrichment. These motifs were then cross-referenced to the scRNA pseudotime gene module Enrichr output for those TF's that had both a hit on our HOMER analysis and in their downstream signaling outputs among modules and topics that exhibited similar patterns (probability or gene module scores) through pseudotime.

Comparison with scRNAseq dataset from pregnancy

Single cell gene expression matrix from Bach et al. (2017) was downloaded and loaded into R version 3.6.3. Using meta data supplied by the authors, cells corresponding to their published analysis of both the Nulliparous (NP) and Gestational (G) stages of mouse samples were separated into a single matrix and analyzed using Seurat version 3.1.4. No additional trimming was performed to maintain consistency with the published analyses.

Following standard Seurat workflow[12], a UMAP was generated using the top 2,000 variable genes selected via the default ‘‘vst’’ method. Cluster / cell type labels were preserved from the manuscript for visualization and downstream analysis. Using Seurat, marker genes were generated for the labeled clusters, whereupon the top 100 markers by log fold change for the LP and Avd cell types were used for scoring in the scRNAseq pseudotime analysis (see above). Using the top 100 marker genes by log fold change derived from our scRNAseq data, cells in the Bach et al. (2017) analysis were additionally scored using the Seurat function AddModuleScore() using genes corresponding to the L-sec progenitor and lactation progenitor cell designations and visualized.

In situ RNA analysis using RNAscope

Mammary glands were harvested from a 10-week old C57BL/6 mouse and frozen in O.C.T Compound (4583, Sakura). 10-micron sections were fixed with fresh 4% PFA made from 40% PFA (15715-S, Electron Microscopy Sciences) diluted in PBS (21-031- CV, Corning) for 1 hour at RT. The RNAscope assay for the Aldh1a3 probe (501201, ACDBio) was performed according to the manufacturer’s protocol for fresh frozen sections. The images were acquired with a Zeiss LSM 700 confocal microscope. Fiji was used to calculate the number of Aldh1a3 foci (RNA molecules) per nuclei manually. Nuclei enveloped in Krt14 protein are called basal for this analysis. Nuclei adjacent to, but not enveloped, are called luminal. To quantify the percentage of Aldh1a3-positive cells, we applied a cut-off of n > 5 molecules per nuclei and calculated the percentage of all cells in basal or luminal compartment.

Reprinted with permission from:

Pervolarakis, Nicholas, et al. "Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Regulators of Mammary Epithelial Cell Identity." Cell Reports 33.3 (2020): 108273.

Author Contributions:

K.K. and K.W. designed and supervised the research; Q.H.N., K.W., G.G., P.S., and D.J. performed the research; X.D., J.R., G.X.Y.Z., and C.M.N. contributed reagents and analytic tools; N.P., Q.H.N., and K.W. performed bioinformatic analyses; N.P., K.W., and K.K. wrote the manuscript; and all authors discussed the results and provided comments and feedback.

Chapter 4: Conclusion and Future Directions

The mammary gland remains a critical area of study at the intersect developmental, cell, and cancer biology. With continued advances in standard of care for breast cancer patients (https://www.breastcancer.org/research-news/20100930), greater and greater value should be placed on tools for early detection and informing risk for the patient population. That value is generated by the thorough characterization of the normal breast epithelial hierarchy to establish a foundation. Moving forward, we as scientists can clearly establish the molecular transitions that accompany cancer development. Armed with this information we will save lives and reduce healthcare costs across the board. From a developmental perspective, the majority of the mammary gland takes place postnatally and so mouse model systems are ideal to interrogate and test hypotheses. Heterogeneity has always been relevant to the underlying questions being asked in mammary gland, but it is only now with the advent of single cell technologies that we can truly begin to explore an unbiased view of the biology at hand.

This is not without its challenges, and through the process of the thesis described above, we examine the boons and potential pitfalls of the nature of library preparation and the resultant data's analysis. Due to the fast moving nature of this field, many of the specific computational approaches applied to ameliorate issues such as batch effects and unwanted technical variation are likely to have new solutions in the works or already published. That being said, the foundational considerations of these issues will never go away when designing an experimental plan and the more that researchers can set themselves up for success by reducing potential for issues will only improve a study. Benjamin Franklin famously said “An ounce of prevention is worth a pound of cure” in reference to fire susceptibility in Philadelphia, and this is no less true for in science. It is better to cover your bases beforehand than have to put out “fires” computationally post dataset generation.

Chapter 2 of this thesis employed scRNA-seq as a tool to characterize the normal, adult mammary epithelial compartment from mammoplasty reduction tissue sources. When working with human tissue, this source is one of the better ways to get “normal” tissue without capture postmortem but not without its own biases (weight, age, parity status). We show that despite patient to patient variation driving individual datasets to separate in our analysis that their remains a distinct underlying structure to the epithelial hierarchy of Basal Cells, L1, and L2 cell types respectively. Basal and Myoepithelial cells marked by high *KRT14* expression, with the former having more specific expression of *APOD* and *TIMP1*, while the latter is characterized by higher *TAGLN* and *ACTA2* expression(Figure 2.4) The luminal compartment presented two main cell types that were both positive for *KRT8* and *KRT18* expression, with a secretory-like cell type referred to as L1 marked by *SLPI* and a more hormone responsive-like population marked by specific expression of *ANKRD30A* and *AGR2* (Figure 2.4). The L1 compartment also exhibited two cell states within, with L1.1 characterized by higher *LTF* positivity and L1.2 presenting higher *CLDN4* expression (Figure 2.4).

The mRNA gene expressions of these cell types were consistent across individuals sampled in the study, but things were taken a step further to validate the existence of these cell types through immunofluorescence analysis in tissue sections. In the basal compartment, it is shown that the heterogeneous observation of KRT14 expression along with rare population characterized by *ZEB1* and *TCF4* positivity are preserved from mRNA to protein (Figure 2.6). We show distinct separation spatially of the L1 and L2 cell types (Figure 2.8), through staining of SLPI, ANKRD30A, and three different hormone receptors (ER, PR, AR). Additionally, it was observed through expression and staining that in the luminal compartment there exists a subset of cells that are double positive for both KRT14 and KRT18 (Figure 2.*7*).

The next step of the analysis was to attempt to relate these different cell types and states to one another in a developmental context bioinformatically. Because this is primary human data, we cannot easily establish a true timeline of development in a healthy tissue and so must rely on the "snapshot" nature of our epithelial cell capture for sequencing and the dynamic process of homeostasis within the adult gland. For this, we employed the pseudotemporal reconstruction algorithm Monocle to generate a trajectory of the putative developmental relationship between the observed cell types. With *ZEB1*/*TCF4* positive basal cells set as the beginning of pseudotime, the graph then branched into 3 major trajectories (Figure 2.9,2.10,2.11). Basal cells comprised the right most path, with an enrichment for the Myoepithelial subtype the later into pseudotime the cells progressed. To the other side of the trajectory, the luminal cells segregated themselves into two separate branches largely enriched for L1 and L2 respectively. It is also of note that the L1.2 subtype presented the highest density at the branch point between the two, suggesting it may be a more progenitor-like cell state that gives rise to the distinct luminal cell types.

Chapter 3 of this thesis took the next steps past scRNA-seq analysis of the mammary gland epithelium and employed scATAC-seq to interrogate the notion of heterogeneity encoded at the epigenetic level of cells and what regulatory machinery contributes to cell identity in a mouse model system. Reassuringly, the basic architecture of the mammary gland between human and mouse with one major basal population represented, and two major luminal populations (L-HR and L-Sec) (Figure 3.1). The L-Sec cell type showed additional sub-structure by splitting into two distinct subclusters. Using the chromatin accessibility data, we show that through pseudoexpression generated using Cicero that basal cells are characterized by increased accessibility at the *Krt17* and *Krt14* genomic loci (Figure 3.1). L-HR exhibited specific accessibility associated with *Foxa1*, while L-Sec alternatively possessed unique accessibility associated with the Kit genomic loci.

Using scRNA-seq, we show that a similar architecture of the gland. Here canonical cell type markers are consistent with basal cells expressing high levels of Krt14, Acta2, and Tagln (Figure 3.2). In the luminal compartment, L-HR cells were characterized by *Krt18*, *Prlr*, and *Areg* while the both L-Sec populations were marked by *Krt18 (Figure 3.2,3.3)*. Taking a deeper dive into the substructure of the L-sec compartment, we show through differential expression, RNAscope, and qPCR after FACS that there is consistent heterogeneous expression of *Aldh1a3* associated with the L-Sec Progenitor-like population as compared to the L-Sec Mature (Figure 3.2,3.3).

With the language of progenitor and mature-like being used to describe these cell states within the L-Sec compartment, it was a natural next step to employ pseudotemporal trajectory construction algorithms to investigate how gene expression and chromatin accessibility profiles change as a cell travels the path from a more progenitor like state to that of the mature. The resultant graph for the scRNA-seq trajectory begins with cells of the L-Sec Prog cluster, and as cells travel through positive pseudotime end up in regions enriched for the L-sec Mat cluster (Figure 3.4). We show that there is a reduction in gene expression associated with a Luminal Progenitor type signature (taken from an analysis of Bach et al. 2017) and an increase in gene expression associated with their published Differentiated Alveolar through pseudotime (Figure 3.4). The scATAC-seq data provided a graph reconstruction similar to that of the scRNA-seq in their relationship to progenitor and mature states and pseudotime, as well as a host of genomic regions differentially accessible through pseudotime (Figure 3.5). These regions were then processed with HOMER to calculate motif enrichment of transcription factors. Specifically, we found that the SMAD3 motif was associated with cells early in pseudotime while the GATA1 motif exhibited accessibility late in the trajectory (Figure 3.5). Using GO annotations, we then took genes associated with SMAD3 and GATA1 downstream signaling respectively and show that in the scRNA-seq based trajectory a very similar pattern of expression changes that mirror that of the motif's pattern (Figure 3.4,3.5).

Up until this point, we have highlighted the existence of three major cell types in the mouse mammary epithelium and dug deeper into insights within the secretory luminal compartment but we have fallen short of driving home the relationship of what chromatin accessibility profiles correspond to what gene expression profiles across our two data modalities. To accomplish this, we employed the Seurat label transfer and integration workflow to first predict cell type labels with the scRNA-seq data as our reference and projecting onto the scATAC-seq data as the query. From there, we calculated anchors across the data and generated a co-embedding resulting in a final UMAP that contains cells from both the scRNA-seq and scATAC-seq with common cluster labels (Figure 3.6). With this space now generated, we were interested in asking what transcription factor motifs might be specifically associated with variation in accessibility and if we can find evidence for their downstream signaling activity through gene expression in a cell type specific fashion. To accomplish this, we first employed ChromVar to calculate TF motif enrichment scores for every cell in the scATAC analysis (Figure 3.6,3.7). Using this resultant cells-by-TF matrix as input, we show specific variation in accessibility of TP63 and SMAD3 in Basal cells, ELF1 in L-Sec cells, and FOXA1 in L-HR cells (Figure 3.6,3.7). We additionally find the aforementioned TF's specific downstream gene expression signatures in the same cell type's scRNA-seq data, highlighting the concordance between the two data modalities and the power of leveraging the two together.

Taken together, all of these results point to additional previously uncharacterized heterogeneity present in the mammary gland with an emphasis on the secretory luminal cell type and substates present within. In human, these cell states manifested in our analysis as being at important branch points in pseudotime as well as having similarities to different molecular subtypes on breast cancer. In mice, we highlight in the analogous cell type the existence of a pre-committed lactation progenitor in nulliparous mice and dynamics of a putative developmental relationship between the traditional luminal progenitor cell and these more mature like cells in the gland. This work additionally challenges previously held notions of markers for progenitor like cells in the mammary gland, through the identification of heterogeneous CD61+ expression in the secretory luminal compartment and who's positivity is associated with the more mature like cells within the compartment. What remains as a critical gap in understanding is how this Aldh1a3$^+$ L-Sec luminal progenitor state contributes to in the development of the mature state / lactation committed cells as well as L-HR cells, and if their progenitor capacity is indeed manifest in a tissue context.

To this end, we propose the generation of a WAP-cre Aldh1a3$^{fl/fl}$ mouse model that will specifically delete Aldh1a3 in the mouse mammary gland and disrupt the L-Sec progenitor function[118]. The notion of Aldh1a3$^+$ positivity marks cells as putative mammary progenitor cells is not new[102] and has additionally been discussed through the work presented in Chapter 3, but through the thorough exploration of this model we can gain insight into Aldh1a3's role in progenitor capacity. Firstly, we will characterize this model and the effect of Aldh1a3 deletion on gland development. To accomplish this, WAP-cre Aldh1a3$^{fl/fl}$ adult mouse mammary glands will be harvested at 10 weeks and sectioned for histopathology. Immunohistochemistry staining of Krt14/Krt8/Csn2/Prlr will provide insight into proportions of L-Sec Prog vs L-Sec Mat vs L-HR in the adult gland as well as any morphological differences that may have arose. We hypothesize that ablation of Aldh1a3 in L-Sec cells will disrupt progenitor capacity and cause an accumulation of this more naïve state in the gland, bottlenecking development and greatly reducing the representation of more mature / L-HR cells in the mouse. In H&E, we would expect this result to be shown as a disruption of the regular morphology of the epithelial cell layers, as well as a marked reduction in Csn2 and Prlr staining.

Carrying this idea forward, we will employ the FACS strategy previously described in Chapter 3 to perform mammosphere outgrowth assays to calculate the sphere forming potential of cKit+ luminal cells from the Aldh1a3$^{fl/fl}$ mouse as compared to a wild type B6 background, with the expected result of our floxxed gland cells having less sphere forming potential vs the WT. Additionally, we will employ our developed CD61+ based gating strategy and qPCR to assay the secretory heterogeneity and determine if there has been any additional disruption of the transition of cells from the more progenitor like state to that of the pre-committed lactation phenotype.

These experimental propositions additionally hinge on the assumption that in deleting Aldh1a3 in mammary epithelial cells, we will affect largely the progenitor cell capacity of this substate without otherwise dysregulating the L-Sec Prog identity and our ability to identify and differentiate this cell state from other luminal cell types in the perturbed mouse. It could be the case that through this deletion, we will disrupt cKit / CD61 expression and be unable to stratify our luminal cells via FACS. Additionally, the bottleneck hypothesis may not hold true, and instead we may accumulate other types of luminal cell types present only in the experimental model that will require deeper characterization. To this end, given sufficient preliminary data, we propose to interrogate this mouse model extensively through different stages of lactation and pregnancy through single cell sequencing.

Cohorts of wild type and Aldh1a3$^{fl/fl}$ nulliparous mice at 10 weeks, pregnant, and actively lactating using same cell scRNA-seq / scATAC-seq sampling from 10x Genomics will be subjected to sequencing with the aim of dissecting the role of Aldh1a3 in mouse development and contribution to proper lactation. Although previous studies have investigated the mouse mammary gland with similar time courses, they fell short of identifying the L-Sec heterogeneity that we have highlighted. The combined scATAC-seq data will also provide extremely valuable insight into drivers of cell state and identity at the epigenetic level, and can help answer why cell types that are perturbed manifest as they do in this experimental model.

Through this sequencing, we would expect in the wild type to closely mirror the results of Bach et al. 2017, where from nulliparous to pregnant mice we observe a transition from an enrichment of more progenitor like luminal secretory cells to a strong representation of *Csn2* positive luminal cells in the pregnant model. When disrupted through the Cre construct, we expect that there will be a massive depletion of these *Csn2*+ cells at the pregnancy timepoint concomitant with an accumulation of *Rspo1*+ cells. Incorporating the post-birth / lactation timepoint, with the gland affected as such we would expect the inability for these mice to properly lactate and such only find many basal / myoeptheial cells dominating the gland similar to what was observed in the Bach et al. data.

The scATAC-seq data will also provide for a deeper definition of cell type and state in these models, having shown the strength of this combinatorial approach earlier in this document. We are delving into new territory for how the gland regulates with a disrupted homeostasis and this same cell data will be critical to highlighting the developmental dynamics at play in the gland. The data will also allow for us to observe disruption at the putative enhancer regions described in Chapter 3, and the role of Fol1R accessibility associated with milk regulation in both the WT and experimental groups. We hypothesis that L-Sec specific TF dynamics will not be preserved, and if there is not wholesale reconstruction of co-connectivity of genomic regions in Luminal epithelial cells then at least observable differences in magnitude of connection strengths will make themselves apparent.

This approach is not with its potential pitfalls and caveats. Primary among these is that Aldh1a3 is not essential for proper progenitor function of L-Sec Prog cells in the mouse mammary gland. It has been described as a marker for these cells in human and mouse[102], but this does not mean that functionally it plays an important role. It could be the case that when deleted, other progenitor state drivers pick up the slack so to speak and the cells are able to perform business as usual in the gland without disruption. If that is the case, our first suite of experiments should show this well before sequencing and we will be forced to re-evaluate our usage of this model. It may be necessary to implement a DTP based ablation of these cells in the gland rather than just deleting *Aldh1a3* expression to better disrupt luminal progenitor capacity in the mammary epithelium. The Aldh1a3$^{fl/fl}$ mouse is also of the B6 background as opposed to FVB which was employed in the majority of the study presented in Chapter 3, and so model to model differences can also contribute to potential difficulties.

Beyond these model considerations, the transition from scRNA-seq and scATAC-seq performed separately also presents some potential issues. In Chapter 3 we characterize heterogeneity in the mouse mammary epithelium using scRNA-seq data taken from whole cells, whereas the scRNA-seq data generated in the simultaneous scRNA/scATAC protocol is actual single nuclear RNA data, capturing nuclear transcripts vs cytoplasmic. It could be the case that the heterogeneity described will not be maintained in these nuclear transcriptomic data and we will have a difficult time bioinformatically relating the cell types and states captured in the our newly generated data to the observations and conclusions from past studies that the model was designed to investigate. The wild type control will help with this to serve as a baseline for what heterogeneity can be expected at this resolution, but this may still fall short. In this case, we will lean heavily on the chromatin accessibility data that had examples independent of what gene expression data to serve as an anchor point for making comparison.

References

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

2. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

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

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

5. Dontu, G. & Ince, T. A. Of Mice and Women: A Comparative Tissue Biology Perspective of Breast Stem Cells and Differentiation. *Journal of Mammary Gland Biology and Neoplasia* **20**, 51–62 (2015).

6. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006).

7. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993–997 (2006).

8. Wang, D. *et al.* Identification of multipotent mammary stemcells by protein C receptor expression. *Nature* **517**, 81–84 (2015).

9. Fu, N. Y., Nolan, E., Lindeman, G. J. & Visvader, J. E. Stem Cells and the Differentiation Hierarchy in Mammary Gland Development. *https://doi.org/10.1152/physrev.00040.2018*

100, 489–523 (2020).

10. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).

11. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–90 (2015).

12. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

13. Pott, S. & Lieb, J. D. Single-cell ATAC-seq: Strength in numbers. *Genome Biol.* **16**, 1–4 (2015).

14. Muto, Y. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat. Commun. 2021 121* **12**, 1–17 (2021).

15. Jia, G. *et al.* Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. Commun. 2018 91* **9**, 1–17 (2018).

16. AM, R. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28**, 472-487.e7 (2021).

17. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 1–14 (2020).

18. Tung, P. Y. *et al.* Batch effects and the effective design of single-cell gene expression

studies. *Sci. Rep.* **7**, (2017).

19. Qiu, X., De Jesus, J., Pennell, M., Troiani, M. & Haun, J. B. Microfluidic device for mechanical dissociation of cancer cell aggregates into single cells. *Lab Chip* **15**, 339–350 (2015).

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

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

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

23. Will, B. & Steidl, U. Multi-parameter fluorescence-activated cell sorting and analysis of stem and progenitor cells in myeloid malignancies. *Best Practice and Research: Clinical Haematology* **23**, 391–401 (2010).

24. Van Den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods* **14**, 935–936 (2017).

25. Kornbluth, J. & Hoover, R. G. Anti-HLA Class I Antibodies Alter Gene Expression in Human Natural Killer Cells. in *Immunobiology of HLA* 150–152 (Springer Berlin Heidelberg, 1989). doi:10.1007/978-3-662-39946-0_39

26. Christaki, E. *et al.* A monoclonal antibody against RAGE alters gene expression and is

protective in experimental models of sepsis and pneumococcal pneumonia. *Shock* **35**, 492–498 (2011).

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

28. Romero-Santacreu, L., Moreno, J., Pérez-Ortín, J. E. & Alepuz, P. Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. *RNA* **15**, 1110–1120 (2009).

29. Schroeder, A. *et al.* The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, (2006).

30. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science (80-.).* **347**, (2015).

31. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).

32. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

33. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

34. Heath, J. R., Ribas, A. & Mischel, P. S. Single-cell analysis tools for drug discovery and development. *Nature Reviews Drug Discovery* **15**, 204–216 (2016).

35. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells.

Nat. Commun. **8**, (2017).

36. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

37. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344**, 1396–1401 (2014).

38. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352**, 189–196 (2016).

39. Yuan, G. C. *et al.* Challenges and emerging directions in single-cell analysis. *Genome Biology* **18**, (2017).

40. Lawson, D. A. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).

41. Grindberg, R. V. *et al.* RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19802–19807 (2013).

42. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).

43. Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science (80-.).* **353**, 925–928 (2016).

44. Krishnaswami, S. R. *et al.* Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).

45. Lacar, B. *et al.* Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, (2016).

46. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science (80-.).* **352**, 1586–1590 (2016).

47. Ambati, S. *et al.* Adipocyte nuclei captured from VAT and SAT. *BMC Obes.* **3**, (2016).

48. Barthelson, R. A., Lambert, G. M., Vanier, C., Lynch, R. M. & Galbraith, D. W. Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells. *BMC Genomics* **8**, (2007).

49. Trask, H. W. *et al.* Microarray analysis of cytoplasmic versus whole cell RNA reveals a considerable number of missed and false positive mRNAs. *RNA* **15**, 1917–1928 (2009).

50. Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* **8**, (2017).

51. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**, (2017).

52. Hicks, S. C., Teng, M. & Irizarry, R. A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv* (2015).

53. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

54. Stoeckius, M. *et al.* Cell "hashing" with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *bioRxiv* 237693 (2017). doi:10.1101/237693

55. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat.*

Biotechnol. **36**, 421–427 (2018).

56. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

57. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, (2015).

58. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

59. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).

60. duVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H. & Tsuda, K. CellTree: An R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics* **17**, (2016).

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

62. Visvader, J. E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev.* **28**, 1143–58 (2014).

63. Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.* **14**, 1–19 (2012).

64. Stingl, J., Eaves, C. J., Zandieh, I. & Emerman, J. T. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer*

Res. Treat. **67**, 93–109 (2001).

65. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009).

66. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).

67. Ting, D. T. *et al.* Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **8**, 1905–1918 (2014).

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

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

70. Morel, A. P. *et al.* A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat. Med.* **23**, 568–578 (2017).

71. Ye, X. *et al.* Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature* **525**, 256–260 (2015).

72. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491–1498 (2015).

73. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).

74. Gudjonsson, T., Adriance, M. C., Sternlicht, M. D., Petersen, O. W. & Bissell, M. J. Myoepithelial cells: their origin and function in breast morphogenesis and neoplasia.

Journal of mammary gland biology and neoplasia **10**, 261–272 (2005).

75. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

76. Filipczyk, A. *et al.* Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat. Cell Biol.* **17**, 1235–1246 (2015).

77. Robin, Y. M. *et al.* 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**, 502–510 (2013).

78. Spike, B. T. *et al.* A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. *Cell Stem Cell* **10**, 183–197 (2012).

79. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

80. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, (2016).

81. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).

82. Pal, B. *et al.* Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat. Commun.* **8**, (2017).

83. Rios, A. C., Fu, N. Y., Lindeman, G. J. & Visvader, J. E. In situ identification of bipotent stem cells in the mammary gland. *Nature* **506**, 322–327 (2014).

84. Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189–193 (2011).

85. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

86. Seil, I. *et al.* The differentiation antigen NY-BR-1 is a potential target for antibody-based therapies in breast cancer. *Int. J. Cancer* **120**, 2635–2642 (2007).

87. Choudhury, S. *et al.* Molecular profiling of human mammary gland links breast cancer risk to a p27+ cell population with progenitor characteristics. *Cell Stem Cell* **13**, 117–130 (2013).

88. Huh, S. J. *et al.* The proliferative activity of mammary epithelial cells in normal tissue predicts breast cancer risk in premenopausal women. *Cancer Res.* **76**, 1926–1934 (2016).

89. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, (2011).

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

91. Visvader, J. E. Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes and Development* **23**, 2563–2577 (2009).

92. Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, (2018).

93. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.* **8**, 1–11 (2017).

94. Giraddi, R. R. *et al.* Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep.* **24**, 1653-1666.e7 (2018).

95. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–9 (2015).

96. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

97. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

98. Dravis, C. *et al.* Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. *Cancer Cell* **34**, 466-482.e6 (2018).

99. Chung, C.-Y. *et al.* Single-cell chromatin accessibility analysis of mammary gland development reveals cell state transcriptional regulators and cellular lineage relationships. (2019). doi:10.1101/624957

100. Sun, Y., Xu, Y., Xu, J., Lu, D. & Wang, J. Role of TM4SF1 in regulating breast cancer cell migration and apoptosis through PI3K/AKT/mTOR pathway. *Int. J. Clin. Exp. Pathol.*

8, 9081–9088 (2015).

101. Cai, C. *et al.* R-spondin1 is a novel hormone mediator for mammary stem cell self-renewal. *Genes Dev.* (2014). doi:10.1101/gad.245142.114

102. Eirew, P. *et al.* Brief report: Aldehyde dehydrogenase activity is a biomarker of primitive normal human mammary luminal cells. *Stem Cells* **30**, 344–348 (2012).

103. Asselin-Labat, M. L. *et al.* Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nat. Cell Biol.* **9**, 201–209 (2007).

104. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).

105. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).

106. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

107. Sundqvist, A., ten Dijke, P. & van Dam, H. Key signaling nodes in mammary gland development and cancer: Smad signal integration in epithelial cell plasticity. *Breast Cancer Research* **14**, 1–13 (2012).

108. Kouros-Mehr, H., Kim, J. whan, Bechis, S. K. & Werb, Z. GATA-3 and the regulation of the mammary luminal cell fate. *Current Opinion in Cell Biology* **20**, 164–170 (2008).

109. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat.*

Methods (2017). doi:10.1038/nmeth.4401

110. Forster, N. *et al.* Basal cell signaling by p63 controls luminal progenitor function and lactation via NRG1. *Dev. Cell* **28**, 147–60 (2014).

111. Liu, Y. *et al.* Foxa1 is essential for mammary duct formation. *Genesis* **54**, 277–285 (2016).

112. Zhou, J. *et al.* Elf5 is essential for early embryogenesis and mammary gland development during pregnancy and lactation. *EMBO J.* **24**, 635–644 (2005).

113. Stemmler, M. P., Eccles, R. L., Brabletz, S. & Brabletz, T. Non-redundant functions of EMT transcription factors. *Nature Cell Biology* **21**, 102–112 (2019).

114. Moses, H. & Barcellos-Hoff, M. H. TGF-β Biology in mammary development and breast cancer. *Cold Spring Harb. Perspect. Biol.* **3**, 1–14 (2011).

115. Bernardo, G. M. *et al.* FOXA1 is an essential determinant of ERα expression and mammary ductal morphogenesis. *Development* **137**, 2045–2054 (2010).

116. Gao, T. *et al.* EnhancerAtlas: A resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw495

117. Kessenbrock, K. *et al.* A Role for matrix metalloproteinases in regulating mammary stem cell function via the Wnt signaling pathway. *Cell Stem Cell* **13**, 300–313 (2013).

118. Chassot, A. A. *et al.* Retinoic acid synthesis by ALDH1A proteins is dispensable for meiosis initiation in the mouse fetal ovary. *Sci. Adv.* **6**, (2020).