Published in final edited form as:
Cell Rep. 2020 October 20; 33(3): 108273. doi:10.1016/j.celrep.2020.108273.

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

Nicholas Pervolarakis[1,2], **Quy H. Nguyen**[1], **Justice Williams**[1], **Yanwen Gong**[1,2], **Guadalupe Gutierrez**[1], **Peng Sun**[1], **Darisha Jhutty**[3], **Grace X.Y. Zheng**[3], **Corey M. Nemec**[3], **Xing Dai**[1], **Kazuhide Watanabe**[4,*], **Kai Kessenbrock**[1,2,5,*]

[1]Department of Biological Chemistry, University of California, Irvine, Irvine, CA 92697, USA

[2]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA 92697, USA

[3]10X Genomics, 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566, USA

[4]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[5]Lead Contact

SUMMARY

The mammary epithelial cell (MEC) system is a bilayered ductal epithelium of luminal and basal cells, maintained by a lineage of stem and progenitor populations. Here, we used integrated single-cell transcriptomics and chromatin accessibility analysis to reconstruct the cell types of the mouse MEC system and their underlying gene regulatory features in an unbiased manner. We define differentiation states within the secretory type of luminal cells, which forms a continuous spectrum of general luminal progenitor and lactation-committed progenitor cells. By integrating single-cell transcriptomics and chromatin accessibility landscapes, we identify *cis*- and *trans*-regulatory elements that are differentially activated in the specific epithelial cell types and our newly defined luminal differentiation states. Our work provides a resource to reveal *cis*/*trans*-regulatory elements associated with MEC identity and differentiation that will serve as a reference to determine how the chromatin accessibility landscape changes during breast cancer.

In Brief

Pervolarakis et al. use integrated analysis of single-cell transcriptomics and chromatin accessibility data to reveal the spectrum of heterogeneity of mouse mammary epithelial cell types and states,

[*]Correspondence: kazuhide.watanabe@riken.jp (K.W.), kai.kessenbrock@uci.edu (K.K.).

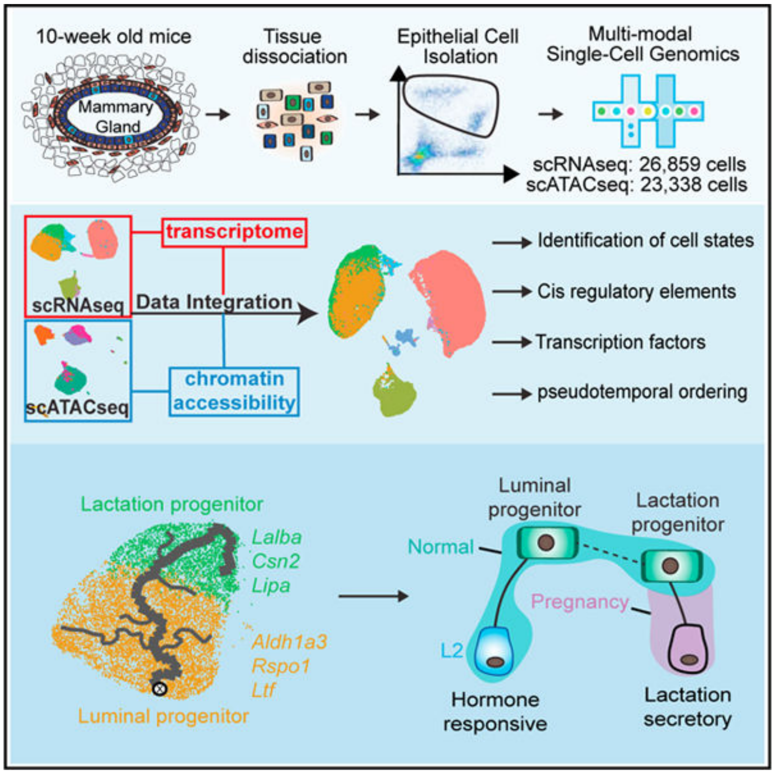
AUTHOR CONTRIBUTIONS
K.K. and K.W. designed and supervised the research; Q.H.N., K.W., G.G., P.S., and D.J. performed the research; X.D., J.R., G.X.Y.Z., and C.M.N. contributed reagents and analytic tools; N.P., Q.H.N., and K.W. performed bioinformatic analyses; N.P., K.W., and K.K. wrote the manuscript;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

SUPPLEMENTAL INFORMATION
Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.108273.

DECLARATION OF INTERESTS
D.J., G.X.Y.Z., and C.M.N. are employed by 10X Genomics.

along with regulatory elements that contribute to these differences. Analysis of secretory luminal cells defines their differentiation trajectory and underlying regulatory factors.

Graphical Abstract



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 (Perou et al., 2000). 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 (Visvader, 2009), 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 (Shackleton et al., 2006; Stingl et al., 2006), as well as subpopulations of progenitors and mature, hormone-responsive cells defined within the luminal compartment (Shehata et al., 2012).

Technological advances enable us to explore cellular heterogeneity without bias using single-cell RNA sequencing (scRNA-seq) (Pollen et al., 2014). This approach was used to describe the cell types and states within the human (Nguyen et al., 2018) and mouse mammary epithelium (Bach et al., 2017; Giraddi et al., 2018; Pal et al., 2017) 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 (Bach et al., 2017), 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/trans*-regulatory elements associated with cellular identity in bulk assays (Buenrostro et al., 2015a) and at the level of single-cell ATAC-seq (scATAC-seq) (Buenrostro et al., 2015b). This approach provided insights into the differentiation trajectories of the hematopoietic system (Buenrostro et al., 2018; Satpathy et al., 2019) and has elucidated transcriptional regulators of developmental lineages of the fetal mammary gland both using bulk ATAC-seq (Dravis et al., 2018) and scATAC-seq (Chung et al., 2019).

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 (Bach et al., 2017; Giraddi et al., 2018; Nguyen et al., 2018; Pal et al., 2017). 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 1A). After data processing using the Cell Ranger pipeline (10X Genomics), we performed unbiased clustering on all peaks using Seurat (Macosko et al., 2015), which revealed 4 main clusters (0–3) of MECs and minor populations of contaminating stromal cells (Figure 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 (Stuart et al., 2019). 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 1C; Figure S1C). We also generated pseudobulk profiles to visualize differentially accessible genomic regions. *Wnt10a* was found to be specifically accessible in basal cells (Figure 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 1E).

Interestingly, we observed two distinct clusters within the L-Sec cell type (Figure 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 [Sun et al., 2015]), and cluster 3 (marked by *Rspo1*, encoding a regulator of Wnt signaling, R-Spondin 1, that can mediate mammary stem cell renewal [Cai et al., 2014]). Further gene activity analysis revealed numerous specific marker loci that are specifically accessible in the respective clusters (Figure S1C; Table S1). Cluster 3 also showed moderate accessibility of the basal marker gene *Krt14* (Figure 1C), suggesting that this cell state within L-Sec shows similarity to basal cells, which could indicate a bipotent progenitor cell state that can differentiate into both basal and luminal lineages or a transitory luminal progenitor that is directly derived from a basal mammary stem cells (Shackleton et al., 2006; Stingl et al., 2006). 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 2A; Figures S2A and S2B). Using clustering through Seurat, we detected three main clusters of MECs and their distinct marker genes (Figure 2B; Figure S2C; Table S2) that correspond to basal (*Krt14+*), L-Sec (*Kit/Elf5+*), and L-HR (*Prlr*+), in line with previous single-cell transcriptomics analyses (Bach et al., 2017; Pal et al., 2017). All clusters were evenly composed of cells from all three individual experiments (Figure S2B). 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) (Figure S2B). We detected two distinct cell states within the L-Sec cluster (Figure 2B; Figure S2D), which emerged as one homogeneous cluster in previous scRNA-seq studies (Bach et al., 2017; Pal et al., 2017). 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 progenitor, whereas the second cluster expressed high levels of genes associated with general luminal progenitor cell capacity, including *Aldh1a3* (Eirew et al., 2012) and *Rspo1*, and therefore labeled the luminal progenitor (Figure 2C). Mature alveolar luminal cells arise during pregnancy and lactation (Visvader, 2009). 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 S2E) (Bach et al., 2017). 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 (Eirew et al., 2012), 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 2D). Quantification of cells with more than 5 transcripts per cell revealed ~15% of *Aldh1a3+* in the luminal compartment detected by RNAscope (Figure 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 (Asselin-Labat et al., 2007), is increased in lactation progenitor cells (Figure 2F). Using flow cytometry, we isolated cKit+/CD61+ and cKit+/CD61− MECs for further validation by qPCR of lactation progenitor genes (Figure 2G; Figure S2F). 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 *Rspol* and *Aldh1a3*, as detected in luminal progenitor cells (Figure 2H).

To functionally test for luminal progenitor capacity, we subjected cKit+/CD61− and cKit+/CD61+ cells to the mammosphere formation assay as recently described (Kessenbrock et al., 2013). This showed that cKit+/CD61− cells formed larger mammospheres compared with cKit+/CD61+ cells, which is in line with the notion of higher progenitor capacity (Figure S2G). However, cKit+/CD61+ cells still possess considerable growth potential similar to basal cells. 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 (Trapnell et al., 2014) 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 3A). 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 3A). 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 (Table S3), we found that L-Sec cells 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 3B and 3C), 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 (Pliner et al., 2018) 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 S3A). To identify modules of genomic peak regions in the scATAC-seq that are coaccessible and vary through pseudotime (Table S3), we employed the CisTopic pipeline to calculate topics of coaccessibility (Bravo Gonzalez-Blas et al., 2019). 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 S3B; Table S3). To link transcriptional and chromatin accessibility dynamics, we next analyzed these specific topics using HOMER (Heinz et al., 2010) to test for significant representation of the transcription factor (TF) binding motifs contained within (Table S3). 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 (Kuleshov et al., 2016) 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 3D and 3E; Figure S3). Smad family TF motifs are key mediators of transforming growth factor β (TGF-β) signaling (Sundqvist et al., 2012), indicating that this pathway is active in luminal progenitor cells. However, GATA signaling is generally associated with luminal differentiation (Kouros-Mehr et al., 2008). 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 (Stuart et al., 2019). 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 S4A). 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 S4B). We observed overall high correlation between ATAC-seq and RNA-seq data (Figure S4C). 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 4B).

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 (Schep et al., 2017) to analyze accessibility of cell-type-specific TF motifs in our scATAC-seq dataset (Table S4). Using Seurat's marker gene test on the resultant TF motif deviation matrix, we uncovered sets of cell-type- specific TF motif enrichments (Figure 4A). We then performed cocorrelation analysis to pinpoint TF modules in the MEC system (Figure S4D), 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 (Forster et al., 2014), but Gata3 and other Gata family TFs were also observed, which have been linked with regulation of luminal cell fate decisions (Kouros-Mehr et al., 2006). Finally, module 3 contained mostly TFs associated with luminal epithelial biology, such as Foxa1 (Liu et al., 2016) and Elf5 (Zhou et al., 2005), but also included a cluster of epithelial-to-mesenchymal transition (EMT)-related TFs, such as Tcf4, Snai2, and ID4 (Stemmler etal., 2019).

Next, we devised cell-type-specific TFs displaying both motif accessibility and active downstream target gene expression as determined by Enrichr analysis (Table S4). Reassuringly, the master regulator of basal cell biology, Tp63 (Forster et al., 2014), 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 4C). 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 β1 (TGF-β1), which has wide implications in regulating mammary biology and cancer (Moses and Barcellos-Hoff, 2011). SMAD TFs also showed increased activity in L-Sec progenitors (Figure 3E; Figure S3B), 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 (Bernardo et al., 2010).

To identify *cis*-regulatory elements that may contribute to cell-type distinction, we used the Cicero pipeline for coaccessibility analysis to determine cell-type-specific genomic connections (Pliner et al., 2018). The resulting connections were subset to those in which one peak of each pair corresponded to an enhancer region from Enhancer Atlas's mouse mammary putative enhancer list (Gao et al., 2016). 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 4D). Further interrogation of gene expression and chromatin accessibility revealed the specific signal for *Folr1* in L-Sec lactation progenitors (Figure 4E). *Folr1* has been identified as a putative regulator of milk protein synthesis in cow mammary glands (Menzieset al., 2009), which is in line with the notion that this cluster is a lactation-committed precursor (Figure 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.

STAR★METHODS

RESOURCE AVAILABILITY

Lead Contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kai Kessenbrock (kai.kessenbrock@uci.edu).

Materials Availability—This study did not generate new unique reagents.

Data and Code Availability—The accession number for the scATACseq and scRNaseq data reported in this study is GEO: GSE157890. All code will be made available upon request.

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 (Kessenbrock et al., 2013). 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 10× Genomics v2 chemistry was performed following the Chromium Single Cell 3’ Reagents Kits v2 User Guide: CG00052 Rev B. Library generation for single cell ATACseq were performed following the Chromium Single Cell ATAC Reagent Kits User Guide: CG000168 Rev B. Single cell RNaseq and ATACseq libraries were sequenced on the Illumina HiSeq4000 platform 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 (Applied Biosystems: 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 37°C cell culture incubator with 5% CO2. Each well then had 1mL of Epicult media added to it. Mammospheres were cultured in a humidified 37°C cell culture incubator with 5% CO2 for a total of 7 days.

Mammosphere image analysis—Brightfield 2× 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 represented 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 package version 3.0.2 (Macosko et al., 2015). 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 FindAllMarkers() function and the "wilcox" test option. GO term enrichment was performed using Enrichr (Kuleshov et al., 2016).

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 Enhancer Atlas mouse mammary list as the first peak of the connection (Gao et al., 2016). 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 (Schep et al., 2017). 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 Find Transfer Anchors () and the cluster labels were predicted using the

Transfer Data () 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 Transfer Data () 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 (Table S4), 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 AddModule-Score() 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 (Kuleshov et al., 2016) 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 comparison to the gene expression modules generated in the RNA analysis, we employed cis Topic 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 cis Topic. The bed files were used as input into the sfindMotifsGenome 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 (Macosko et al., 2015), 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

All data processing, statistical analyses, and figure generation were performed in R unless otherwise specified in the Method Details section above. Specific details regarding the tests applied to data as well as the corresponding n are delineated above as well both in the main text as well as the Method Details. Below, a brief additional summary is provided:

- All marker's for cell types/ states and clusters were identified using Seurat's implementation of a Wilcoxon Rank Sum test, with a p value of 0.05 as a lower bound cutoff, in the exception of the ChromVar TF markers employed in Figure 4, in which Seurat's implementation of a logistic regression was used.

- All output from Enrichr's Fisher's exact test was thresholded to only use output of a p value less than 0.05.
- Cistopic 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Devon Lawson for carefully reading the manuscript. This study was supported by funds from the NIH/NCI (1R01CA234496 and 4R00CA181490 to K.K.), the American Cancer Society (132551-RSG-18-194-01-DDC to K.K.), the Chan/Zuckerberg Initiative (CZF2019-002432), the NIH/NIGMS (R01GM123731 to X.D.), and a Grant-in-Aid for Scientific Research (KAKENHI) on Innovative Areas "Cellular Diversity" (JP18H05106 to K.W.).

REFERENCES

Asselin-Labat ML, Sutherland KD, Barker H, Thomas R, Shackleton M, Forrest NC, Hartley L, Robb L, Grosveld FG, van der Wees J, et al. (2007). Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. Nat. Cell Biol 9, 201–209. [PubMed: 17187062]

Bach K, Pensa S, Grzelak M, Hadfield J, Adams DJ, Marioni JC, and Khaled WT (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. Nat. Commun 8, 2128. [PubMed: 29225342]

Bernardo GM, Lozada KL, Miedler JD, Harburg G, Hewitt SC, Mosley JD, Godwin AK, Korach KS, Visvader JE, Kaestner KH, et al. (2010). FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. Development 737, 2045–2054.

Bravo Gonzalez-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, Davie K, Wouters J, and Aerts S. (2019). cis Topic: cis- regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods 76, 397–400.

Buenrostro JD, Wu B, Chang HY, and Greenleaf WJ (2015a). ATAC- seq: A method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol 709, 21.29.1–21.29.9.

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, and Greenleaf WJ (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490. [PubMed: 26083756]

Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, and Greenleaf WJ (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell 773, 1535–1548.e16.

Cai C, Yu QC, Jiang W, Liu W, Song W, Yu H, Zhang L, Yang Y, and Zeng YA (2014). R-spondin1 is a novel hormone mediator for mammary stem cell self-renewal. Genes Dev. 28, 2205–2218. [PubMed: 25260709]

Chung C-Y, Ma Z, Dravis C, Preissl S, Poirion O, Luna G, Hou X, Giraddi RR, Ren B, and Wahl GM (2019). Single-cell chromatin accessibility analysis of mammary gland development reveals cell state transcriptional regulators and cellular lineage relationships. bioRxiv. 10.1101/624957.

Dravis C, Chung CY, Lytle NK, Herrera-Valdez J, Luna G, Trejo CL, Reya T, and Wahl GM (2018). Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. Cancer Cell 34, 466–482.e6.

Eirew P, Kannan N, Knapp DJHF, Vaillant F, Emerman JT, Lindeman GJ, Visvader JE, and Eaves CJ (2012). Aldehyde dehydrogenase activity is a biomarker of primitive normal human mammary luminal cells. Stem Cells 30, 344–348. [PubMed: 22131125]

Forster N, Saladi SV, van Bragt M, Sfondouris ME, Jones FE, Li Z, and Ellisen LW (2014). Basal cell signaling by p63 controls luminal progenitor function and lactation via NRG1. Dev. Cell 28, 147–160. [PubMed: 24412575]

Gao T, He B, Liu S, Zhu H, Tan K, and Qian J. (2016). Enhancer Atlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics 32, 3543–3551. [PubMed: 27515742]

Giraddi RR, Chung CY, Heinz RE, Balcioglu O, Novotny M, Trejo CL, Dravis C, Hagos BM, Mehrabad EM, Rodewald LW, et al. (2018). Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. Cell Rep. 24, 1653–1666.e7.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, and Glass CK (2010). Simple combinations of lineage- determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589. [PubMed: 20513432]

Kessenbrock K, Dijkgraaf GJP, Lawson DA, Littlepage LE, Shahi P, Pieper U, and Werb Z. (2013). A role for matrix metalloproteinases in regulating mammary stem cell function via the Wnt signaling pathway. Cell Stem Cell 13, 300–313. [PubMed: 23871604]

Kouros-Mehr H, Slorach EM, Sternlicht MD, and Werb Z. (2006). GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. Cell 127, 1041–1055. [PubMed: 17129787]

Kouros-Mehr H, Kim JW, Bechis SK, and Werb Z. (2008). GATA-3 and the regulation of the mammary luminal cell fate. Curr. Opin. Cell Biol 20, 164–170. [PubMed: 18358709]

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44 (W1), W90–W97. [PubMed: 27141961]

Liu Y, Zhao Y, Skerry B, Wang X, Colin-Cassin C, Radisky DC, Kaestner KH, and Li Z. (2016). Foxa1 is essential for mammary duct formation. Genesis 54, 277–285. [PubMed: 26919034]

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214. [PubMed: 26000488]

Menzies KK, Lefevre C, Sharp JA, Macmillan KL, Sheehy PA, and Nicholas KR (2009). A novel approach identified the FOLR1 gene, a putative regulator of milk protein synthesis. Mamm. Genome 20, 498–503. [PubMed: 19669235]

Moses H, and Barcellos-Hoff MH (2011). TGF-β biology in mammary development and breast cancer. Cold Spring Harb. Perspect. Biol 3, a003277.

Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, Phung AT, Willey E, Kumar R, Jabart E, et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat. Commun 9, 2028. [PubMed: 29795293]

Pal B, Chen Y, Vaillant F, Jamieson P, Gordon L, Rios AC, Wilcox S, Fu N, Liu KH, Jackling FC, et al. (2017). Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. Nat. Commun 8, 1627. [PubMed: 29158510]

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. (2000). Molecular portraits of human breast tumours. Nature 406, 747–752. [PubMed: 10963602]

Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. (2018). Cicero Predicts c/s-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell 71, 858–871.e8.

Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, et al. (2014). Low-coverage single cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol 32, 1053–1058. [PubMed: 25086649]

Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol 37, 925–936. [PubMed: 31375813]

Schep AN, Wu B, Buenrostro JD, and Greenleaf WJ (2017). chrom VAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods 74, 975–978.

Shackleton M, Vaillant F, Simpson KJ, Stingl J, Smyth GK, Asselin- Labat ML, Wu L, Lindeman GJ, and Visvader JE (2006). Generation of a functional mammary gland from a single stem cell. Nature 439, 84–88. [PubMed: 16397499]

Shehata M, Teschendorff A, Sharp G, Novcic N, Russell IA, Avril S, Prater M, Eirew P, Caldas C, Watson CJ, and Stingl J. (2012). Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. Breast Cancer Res. 74, R134.

Stemmler MP, Eccles RL, Brabletz S, and Brabletz T. (2019). Non-redundant functions of EMT transcription factors. Nat. Cell Biol 27, 102–112.

Stingl J, Eirew P, Ricketson I, Shackleton M, Vaillant F, Choi D, Li HI, and Eaves CJ (2006). Purification and unique properties of mammary epithelial stem cells. Nature 439, 993–997. [PubMed: 16395311]

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, and Satija R. (2019). Comprehensive Integration of Single-Cell Data. Cell 777, 1888–1902.e21.

Sun Y,Xu Y, Xu J, Lu D, and Wang J. (2015). Role of TM4SF1 in regulating breast cancer cell migration and apoptosis through PI3K/AKT/mTOR pathway. Int. J. Clin. Exp. Pathol 8, 9081–9088. [PubMed: 26464650]

Sundqvist A, Ten Dijke P, and van Dam H. (2012). Key signaling nodes in mammary gland development and cancer: Smad signal integration in epithelial cell plasticity. Breast Cancer Res. 14, 204. [PubMed: 22315972]

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, and Rinn JL (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol 32, 381–386. [PubMed: 24658644]

Visvader JE (2009). Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. Genes Dev. 23, 2563–2577. [PubMed: 19933147]

Zhou J, Chehab R, Tkalcevic J, Naylor MJ, Harris J, Wilson TJ, Tsao S, Tellis I, Zavarsek S, Xu D, et al. (2005). Elf5 is essential for early embryo-genesis and mammary gland development during pregnancy and lactation. EMBO J. 24, 635–644. [PubMed: 15650748]

Highlights

- Generate combined scATAC-seq and scRNA-seq libraries of adult mouse MECs
- Uncover luminal cell heterogeneity and molecular underpinning driving cell type/state
- Identify the pre-committed lactation progenitor within the luminal progenitor compartment
- Regulatory the grammar of MEC types and putative enhancer identification

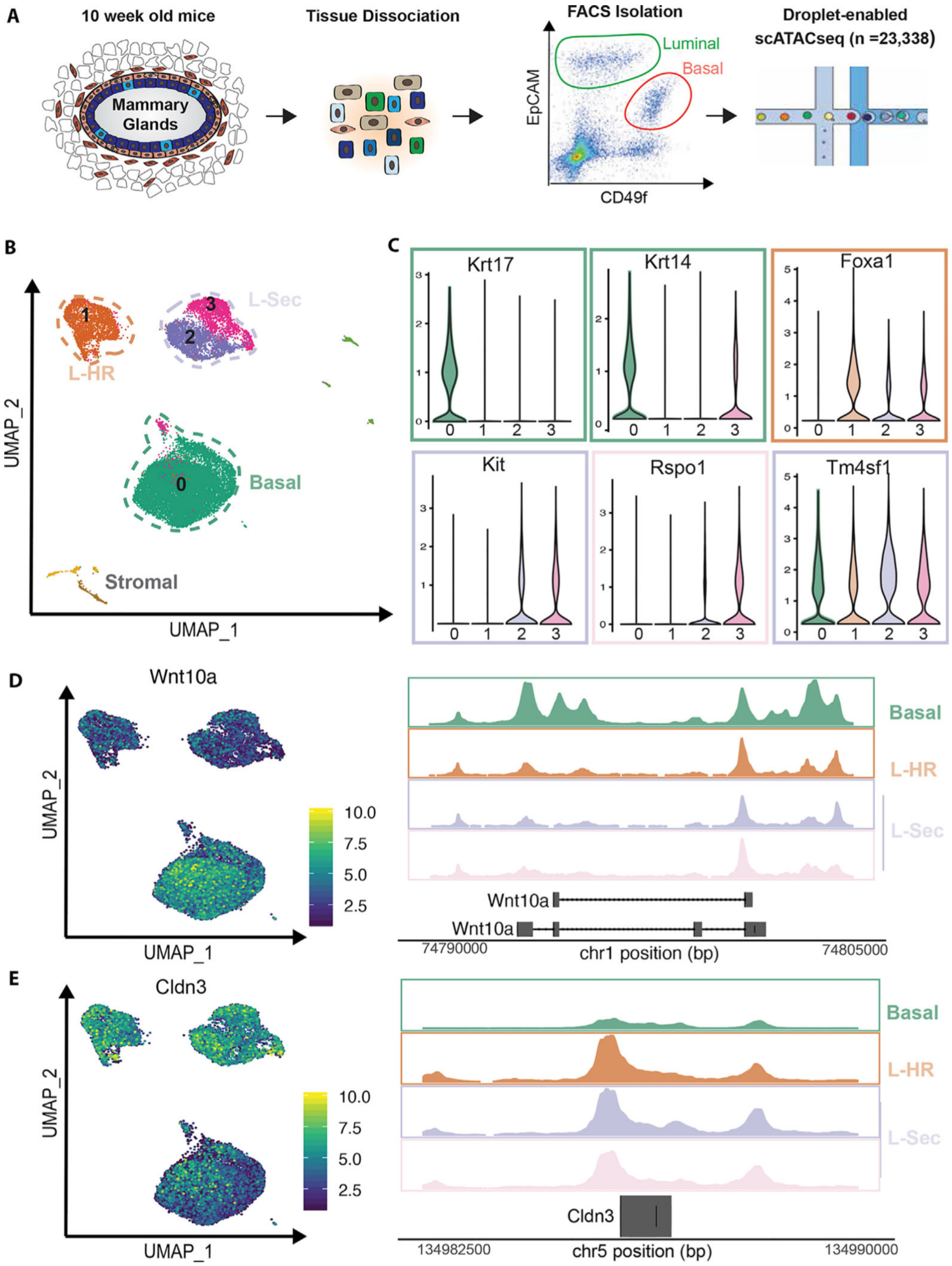


Figure 1. Single-Cell Chromatin Accessibility Profiling of MECs from Post-pubertal Mice Reveals Luminal Epithelial Cell States

(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.

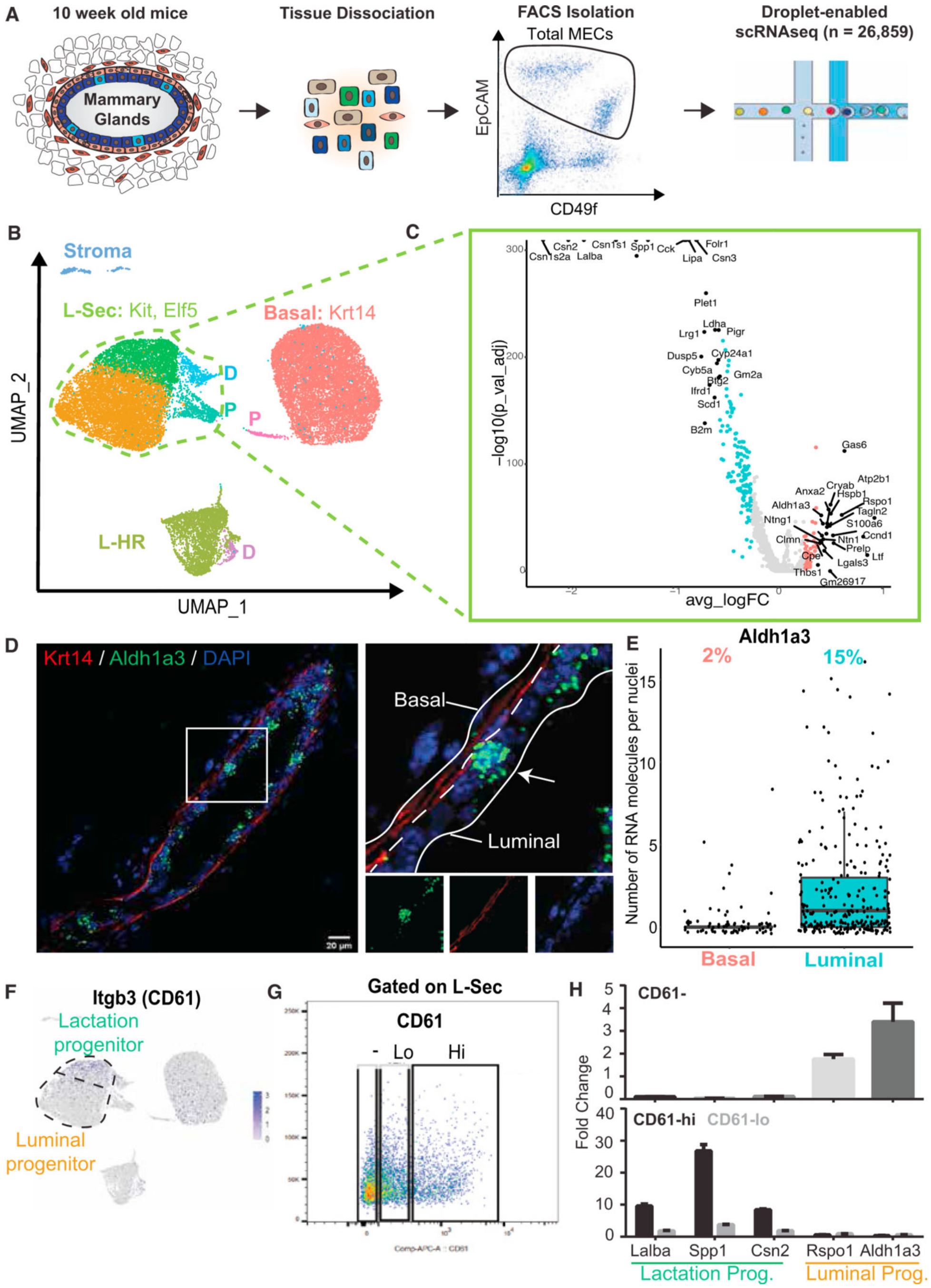


Figure 2. Single-Cell Transcriptomics of MECs Reveal the Lactation-Precursor Cell State

(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* RNA scope 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.

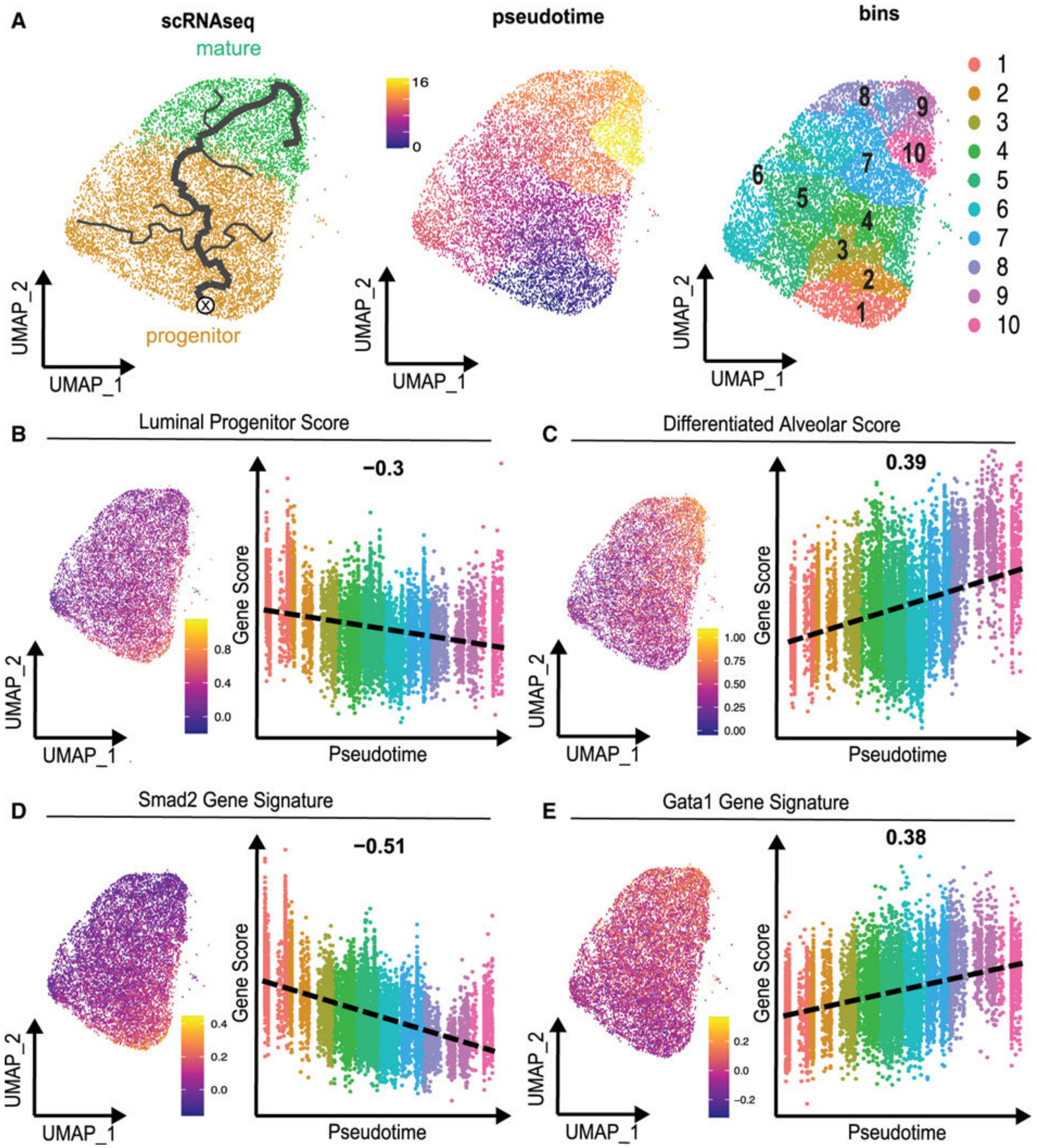


Figure 3. Pseudotemporal Analysis Shows a Continuous Differentiation Trajectory within L-Sec Cells

(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).

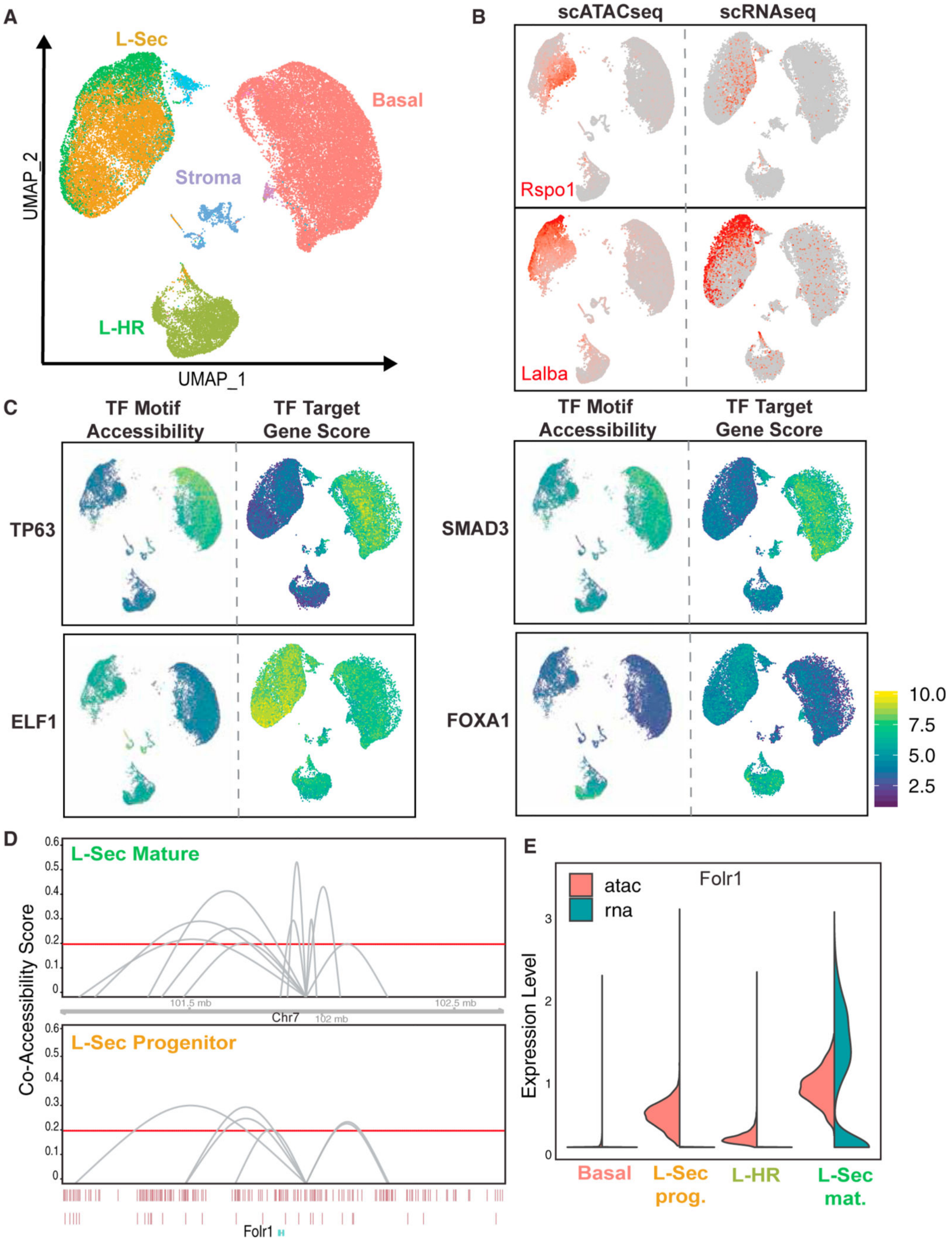


Figure 4. Integration of Single-Cell Chromatin Accessibility and Transcriptomics Datasets

(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 of *Folr1* expression in the coembedded analysis, split by technology type.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
PE-Cyanine7 anti-human/mouse CD49f (integrin alpha 6)	eBioscience	Cat#25-0495-82; RRID: AB_10804881
FITC anti-mouse CD326 (EpCAM)	Biolegend	Cat#118207; RRID: AB_1134106
eFluor450 anti-mouse CD45	eBioscience	Cat#48-0451-82; RRID: AB_1518806
eFluor450 anti-mouse CD31	eBioscience	Cat#48-0311-82; RRID: AB_10598807
eFluor450 anti-mouse Ter119	eBioscience	Cat#48-5921-82; RRID: AB_1518808
PE anti-mouse CD117 (cKit)	Biolegend	Cat#105807; RRID: AB_313216
APC anti-mouse/rat CD61	Biolegend	Cat#104315; RRID: AB_2561733
APC Anti-Mo CD326 (EpCAM)	eBioscience	Cat#17-5791-82; RRID: AB_2716944
Sytox Blue Dead Cell Stain	Invitrogen	Cat#S34857
PE-CD49f	Invitrogen	Cat#12-0495-83; RRID: AB_891476
Chemicals, Peptides, and Recombinant Proteins		
Collagenase Type IV	Sigma Aldrich	Cat#C5138-1g
Epicult-B Mouse Medium	Stemcell	Cat#05610
Matrigel GFR Membrane Matrix	Corning	Cat#354230
Tissue-Tek O.C.T. Compound	Sakura	Cat#4583
Paraformaldehyde Aqueous Solution	Electron Microscopy Sciences	Cat#15715-S
Dulbecco’s Phosphate-Buffered Saline (PBS)	Corning	Cat#21-031-CV
Critical Commercial Assays		
RNAscope Probe- Mm-Aldh1a3	ACDbio	Cat#501201
Chromium Single Cell 3’ Library &Gel Bead Kit v2	10x Genomics	Cat#PN-120237
Chromium Single Cell ATAC Library & Gel Bead Kit	10x Genomics	Cat#PN-1000111
Quick-RNA Microprep RNA isolation kit	Zymo Research	Cat#R1054
iScript cDNA Synthesis Kit	Biorad	Cat#1708891
PowerUp SYBR Green Master Mix	Applied Biosystems	Cat#A25742
Deposited Data		
scATACseq Data	This Paper	GSE157888
scRNaseq Data	This Paper	GSE157889
Mouse NP and Pregnancy scRNaseq data	Bach et al., 2017	GSE10627
Experimental Models: Organisms/Strains		
Mouse: FVB/NJ	Jackson Laboratory	Cat#001800
Mouse: C57BL/6J	Jackson Laboratory	Cat#000664
Software and Algorithms		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R (v3.6.0 or v3.6.2)	The R Project for Statistical Computing	https://www.r-project.org/
Cell Ranger v2.1.0	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
Cell Ranger ATAC v1.1.0	10x Genomics	https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/what-is-cell-ranger-atac
ImageJ v1.52p	ImageJ	https://imagej.nih.gov/ij/download.html
Seurat v3.0.2	Stuart et al., 2019	https://cran.r-project.org/web/packages/Seurat/index.html
Cicero (v1.2.0 or v1.3.4.8)	Pliner et al., 2018	http://bioconductor.org/packages/release/bioc/html/cicero.html
ChromVar v1.4.1	Schep et al., 2017	https://bioconductor.org/packages/release/bioc/html/chromVAR.html
Monocle 3 (v2.99.3 or v0.2.1)	Trapnell et al., 2014	https://cole-trapnell-lab.github.io/cicero-release/docs/#installing-cicero
cisTopic v0.2.2	Bravo Gonzalez-Blas et al., 2019	https://github.com/aertslab/cisTopic
Signac v0.1.0	Stuart et al., 2019	https://cloud.r-project.org/web/packages/Signac/index.html
HOMER v4.7	Heinz et al., 2010	http://homer.ucsd.edu/homer/introduction/install.html