Published in final edited form as:

Cell Rep. 2018 August 07; 24(6): 1653–1666.e7. doi:10.1016/j.celrep.2018.07.025.

Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development

Rajshekhar R. Giraddi[1], **Chi-Yeh Chung**[1], **Richard E. Heinz**[2], **Ozlen Balcioglu**[2], **Mark Novotny**[3], **Christy L. Trejo**[1], **Christopher Dravis**[1], **Berhane M. Hagos**[2], **Elnaz Mirzaei Mehrabad**[2], **Luo Wei Rodewald**[1], **Jae Y. Hwang**[2], **Cheng Fan**[4], **Roger Lasken**[3], **Katherine E. Varley**[2], **Charles M. Perou**[4], **Geoffrey M. Wahl**[1,*], and **Benjamin T. Spike**[2,5,*]

[1]Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[2]Huntsman Cancer Institute, Department of Oncological Sciences, University of Utah, Salt Lake City, UT 84112, USA

[3]J. Craig Venter Institute, La Jolla, CA 92037, USA

[4]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel HI, NC 27599, USA

[5]Lead Contact

SUMMARY

The mammary gland consists of cells with gene expression patterns reflecting their cellular origins, function, and spatiotemporal context. However, knowledge of developmental kinetics and mechanisms of lineage specification is lacking. We address this significant knowledge gap by generating a single-cell transcriptome atlas encompassing embryonic, postnatal, and adult mouse mammary development. From these data, we map the chronology of transcriptionally and epigenetically distinct cell states and distinguish fetal mammary stem cells (fMaSCs) from their precursors and progeny. fMaSCs show balanced co-expression of factors associated with discrete adult lineages and a metabolic gene signature that subsides during maturation but reemerges in some human breast cancers and metastases. These data provide a useful resource for illuminating mammary cell heterogeneity, the kinetics of differentiation, and developmental correlates of tumorigenesis.

[*]Correspondence: wahl@salk.edu (G.M.W.), benjamin.spike@hci.Utah.edu (B.T.S.).

AUTHOR CONTRIBUTIONS

Concept and Supervision, G.M.W. and B.T.S.; Writing, R.R.G., G.M.W., and B.T.S.; Resources, R.L., K.E.V., and C.M.P.; Experimental Design, R.R.G., C.-Y.C., R.E.H., O.B., C.L.T., B.M.H., C.M.P., G.M.W., and B.T.S.; Experimentation, R.R.G., C.-Y.C., R.E.H., O.B., M.N., C.L.T., C.D., B.M.H., L.W.R., and B.T.S.; Computational Design, R.R.G., C.-Y.C., C.D., K.E.V., C.M.P., G.M.W., and B.T.S.; Computational Analysis, C.-Y.C., E.M.M., J.Y.H., C.F., and B.T.S.; Editorial Revision, R.R.G., O.B., E.M.M., G.M.W., and B.T.S. All authors interpreted the data.

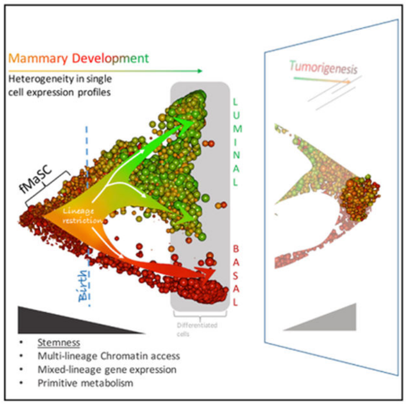
SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, five tables, and one data file and can be found with this article online at https://doi.Org/10.1016/j.celrep. 2018.07.025.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Graphical Abstract



In Brief

Single-cell RNA sequencing of developing mouse mammary epithelia reveals the timing of lineage specification. Giraddi et al. find that fetal mammary stem cells co-express factors that define distinct lineages in their progeny and bear functionally relevant metabolic program signatures that change with differentiation and are resurrected in human breast cancers and metastases.

INTRODUCTION

A deep understanding of complex tissues requires knowledge of the integrated molecular circuitry of each of the tissue’s constituent cells. Prior work used surface markers to fractionate the luminal, basal, and alveolar cells of the mouse mammary gland, and their lineage-restricted progenitors and stem cells (Shackleton et al., 2006; Shehata et al 2012; Sleeman et al., 2006; Stingl et al., 2006; Villadsen et al., 2007). Delineating how the ratios and molecular profiles of these cell types change over development can give valuable insights into the organization of the tissue and the regulators of differentiation and homeostasis. It should also provide insight into subversion of this organization by maladies such as cancer and identify cell states that are susceptible to tumorigenesis and therapeutic targets to prevent or revert tumorigenic phenotypes. We and others have previously reported relationships between the expression profiles of mouse mammary stem/progenitor cell populations and human breast cancers (Lim et al., 2009; Pfefferle et al., 2015; Prat et al., 2010; Spike et al., 2012). In particular, mouse fetal mammary stem cell (fMVaSC)-containing isolates show significant relatedness to aggressive human breast cancers (Pfefferle et al., 2015; Spike et al. 2012). However, it has been challenging to distill critical molecular regulators and cell type-specific biomarkers from bulk profiles since the cell type of interest often constitutes a small fraction of the cell population. For example, transplantation assays show adult mouse mammary stem cells comprise ~2% of sorted cell populations (Shackleton et al., 2006; Spike et al., 2012; Stingl et al, 2006; Wang et al., 2015). While the stem cell fraction is much higher during fetal mammary organogenesis, even the most enriched populations exhibit heterogeneity (Dravis et al., 2015; Spike et al., 2012; Spike et al., 2014).

Single-cell RNA sequencing (scRNA-seq) reveals the cellular and transcriptional heterogeneity of complex tissues (Kumar et al., 2017). For example, expression profiles have recently been obtained for single adult mouse mammary cells (Bach et al., 2017; Pal et al., 2017). However, these studies reveal neither the transcriptional programs that generate mature cell types from primitive embryonic antecedents nor the timing with which developmental transitions occur.

Mouse mammary organogenesis occurs with stereotyped structures at reproducible times (Veltmaat et al., 2003), and with dramatic changes in stem cell function (Spike et al., 2012; Makarem et al., 2013a). fMaSCs are the earliest cells shown by *in vitro* mammosphere formation, *in vivo* lineage tracing, and transplantation to fulfill all criteria for bipotent mammary stem cells (Makarem et al., 2013a; Spike et al., 2012; Van Keymeulen et al., 2011). They become measurable on embryonic day 16 (E16), increase dramatically to E18 (Spike et al., 2012), and then decline immediately after birth to produce the architecturally simple mature mammary epithelium (Giraddi et al., 2015; Makarem et al., 2013b; Prater et al., 2014; Spike et al., 2012). Luminal and basal compartments appear to be sustained by uni-potent cells in adults (Van Keymeulen et al., 2011; Giraddi et al., 2015; Wang et al., 2017; Wuidart et al., 2016), although rare bipotential adult mammary cells may also exist (Rios et al., 2014; Wang et al., 2015).

Here, we elucidate biological programs that distinguish fMaSCs from differentiating cells. We generate a scRNA-seq dataset encompassing fetal, postnatal, and adult mouse mammary epithelia, paying special attention to the perinatal interval, over which the prevalent, multipotent fMaSC phenotype declines and differentiation ensues (Makarem et al., 2013a; Spike et al., 2012). The data establish the chronology of emerging cell types in the mammary epithelium, and underlying changes in chromatin accessibility. We also identify fMaSC gene signatures related to chromatin architecture and metabolism that have significance for fMaSC function and relevance for human breast cancers. Mapping the relationships between single-cell tran-scriptomes as cells differentiate reveals two patterns of gene expression that revise classic models of transcriptionally discrete cell types. First, individual fMaSCs co-express genes associated with differentiating mammary lineages. This supports a model where opposing lineage differentiation factors aid in specifying the uncommitted stem cell state (Loh and Lim, 2011). Second, the fMaSC-containing population constitutes a single distribution of heterogeneous transcriptional states without discrete subclusters corresponding to their measured stem cell fraction. This suggests that stem cell capacity is distributed across heterogeneous cell profiles. The observation has implications for the generation of stem cell activity during tissue repair and cancer progression, and could explain historical difficulties in purifying stem cells using limited markers.

RESULTS

Single-Cell Transcriptomes from Developing Mouse Mammary Glands Distinguish Cell States

We used two approaches to generate scRNA-seq data from embryonic, postnatal, and adult mammary cells sorted for surface Epcam, an epithelial specific marker present throughout development on stem, progenitor, and differentiated mouse mammary epithelial cells (Figures 1A, S1A, and S1B) (Shehata et al., 2012; Spike et al., 2012). We used the Chromium Drop-Seq platform (10x Genomics) to obtain transcriptomes from E16 (n = 690), E18 (n = 1,047), postnatal day 4 (P4) (n = 849), and adult mammary cells (n = 3,838). This includes 786 Epcam$^+$ adult cells sorted for coexpression of Cd49f, ~2% of which are inferred to be functional stem cells based on mammary reconstitution assays (Prater et al., 2014; Shackleton et al., 2006; Spike et al., 2012; Stingl et al., 2006). We constructed a normalized gene expression matrix spanning these developmental stages composed of 6,060 cells with 500–2,000 expressed genes per cell, and a total of 22,184 genes expressed in five or more cells (Andrews, 2010; Dillies et al., 2013; Dobin et al., 2013; Hu and Smyth, 2009; Katayama et al., 2013; Li and Dewey, 2011; Li et al., 2009; Lin et al., 2016; Marinov et al., 2014) (Table S1). We also used the C1 system (Fluidigm) to focus in greater depth on the transition from the stem cell-rich, late embryonic stage to the immediate postnatal stage. With C1, we sequenced 262 cells, which included differentiated adult comparators with >1.5 million reads, 4,000–9,000 genes per cell, and 13,355 genes expressed in five or more cells (Figures 1A and S1C–S1I; Table S2).

We first visualized the Chromium-derived data using t-distributed stochastic neighbor embedding (tSNE), a commonly used dimensionality reduction approach that plots cells with similar profiles as nearby points (Figures 1B, 1C, 1E, and 1F) (van der Maaten and Hinton, 2008). This analysis identified one small group of cells from multiple developmental stages, and seven other groups in which all cells derive from a single stage (Figure 1B). Based on their lack of Epcam RNA, and high levels of Vi-mentin RNA and other non-epithelial markers, the small mixed cell group likely represents contaminating stroma (Figures 1C and S2A). Apart from these cells, adult cells comprised three major clusters, P4 cells comprised two clusters, and E16 and E18 cells comprised one cluster each (Figure 1B). An analysis of known lineage markers indicates the adult groups correspond to basal cells (Itgb1$^+$, Krt14$^+$), mature luminal cells (Ly6a$^+$, Krt8$^+$), and alveolar precursor cells (Cd14$^+$, Csn3$^+$, Krt8$^+$) (Figure 1C). However, we note that lineage-associated markers are often imperfect in that they show sporadic expression in other adult groups (Figure 1C). The data also show that individual cells from earlier developmental stages often express multiple lineage markers (Figure 1C). For example, while a minority of P4 cells exhibit a basal profile, the majority manifest luminal and alveolar features (e.g., Krt8, Csn3, and Cd14 expression) in addition to Krt14 (Figure 1). This contrasts with the interpretation that prepubertal cells are largely basal (Pal et al., 2017).

We next applied non-negative matrix factorization (NMF) as an alternative approach to cluster samples into transcriptional cell types based on the 1,000 most variably expressed genes in the data matrix (Figures 1D and S1L) (Brunet et al., 2004; Lee and Seung, 1999; Saeed et al., 2003; Shao and Höfer, 2017). This analysis included sorted adult basal cells (Epcam$^+$, Cd49f$^+$). We observed a small cluster of mixed-stage cells corresponding to the mixed tSNE cluster and three major adult cell clusters: (1) a basal cluster including the vast majority of presorted adult basal cells, (2) the presumptive luminal alveolar group, and (3) the mature luminal group. Some P4 cells from the smaller of the two P4 tSNE clusters were grouped with the adult basal compartment in this analysis (Figures 1D and 1E). Importantly, embryonic epithelial cells again comprise a single cluster in this analysis.

Functional assays indicate that up to 2% of adult basal cells, and 10%−50% of E18 cells possess mammary stem cell attributes, depending on the markers used for isolation (Dravis et al., 2015; Shackleton et al., 2006; Spike et al., 2012; Stingl et al 2006; Wang et al., 2015). We examined these populations alone by tSNE to attempt to identify minority stem cell clusters (Figure 1F). However, other than stromal or luminal alveolar contaminants, neither group showed obvious subclusters (Figure 1F). The contaminating cell types likely reflect the technical limitations of cell sorting, the biological inaccuracy of marker-based purification, or both. Similar to Pal et al. (2017), we did identify occasional basal cells expressing atypical markers such as Elf5, Muc1, and Kit, but these genes did not identify the same cells or a coherent subgroup, and mixed phenotype adult profiles did not occur at a rate exceeding the expected doublet frequency in these assays, ~1% for 10× derived profiles (see STAR Methods; 10× Chromium V2 Guide). These data suggest that either the functional stem cell state can be generated by different transcriptional programs or that multiple types of cells can act as facultative stem cells under the assay conditions employed for functional testing.

Discrete Mammary Epithelial Lineages Arise Postnatally with Loss of Balanced Lineage Factor Co-expression

As only the relationships between the most similar cells are well represented in tSNE, it can be difficult to interpret cellular relationship more globally across a diverse dataset. As an alternate approach to investigate the relationships between single-cell expression profiles, differentiation states, and developmental context, we plotted the single-cell data according to diffusion components (DCs), a noise-tolerant, non-linear dimensionality reduction method that reveals a global topology for the data based on local similarities between points (Coifman et al., 2005; Haghverdi et al., 2015) (Figures 2A and 2E). The resulting graph produced an intuitive developmental picture in which primitive (E16 and E18) and adult cells occupied opposite ends of the DC2 axis, while P4 cells localized to intermediate positions (Figure 2A). The regions occupied by E16 and E18 cells largely overlap, while P4 cells form a continuum between the more primitive cells and those of the adult. Importantly, while adult cells occupy discrete distributions at various extremes along the DC1 axis, E16 and E18 cells do not bifurcate and are instead positioned midway between the two DC1 extremes. P4 cells distribute in a unique pattern along DC1, with a small group extending toward a tightly grouped set of adult cells (Figure 2A, bottom right quadrant) and a second larger group of cells extending as a single group toward the two major adult groups (upper right quadrant). We provide a web tool for the interactive visualization of gene expression across this dataset (http://uofuhealth.utah.edu/huntsman/labs/spike/d3.php).

Graphing cells according to the ratio of transcripts for the well-known luminal and basal markers, Krt8 and Krt14, respectively, is concordant with *in situ* staining for the products of these lineage-associated genes over the same developmental window (Figure 2B). That is, we observed the following: (1) cells with mixed keratin expression predominating early (E18.5); (2) emergence of a minority Krt14-expressing P4 population that correlates with a well-defined, elongating myoepithelial cell layer lining the P4 epithelium; (3) a majority of P4 cells harboring a mixed lineage Krt8^{+}Krt14^{+} phenotype but positioned proximal to adult luminal cell types in the diffusion map; and (4) the resolution of this group into cells expressing Krt8 but lacking Krt14 in the adult gland, as would be expected for mature luminal cells lining the ducts (Figure 2B). In addition, each NMF-derived cluster represented a single region of the diffusion map in the first two DCs with the exception of NMF group 1. NMF1 corresponds to the group of cells from mixed developmental stages with stromal expression patterns and was largely distinguishable from epithelial cells along a higher DC (DCS) (Figure 2C). NMF-VII and NMF-II localize to the different extremes of DC1 and represent basal cells and Esr1$^+$ luminal cells expressing Ly6a (the gene for Seal), respectively (Figures 1C, 2C, and S2C). NMF-IV is positioned intermediately and contains cells expressing Wfdc18, Csn3, Csn2, Kit, and Itga2, indicative of an alveolar luminal phenotype (Figures 2C and S2B) (Pal et al., 2017).

These distributions imply that embryonic cells are distinguished from adult cells by expression patterns that change gradually over developmental time along DC2, and by an intermediate phenotype with respect to lineage differentiation (DC1). We chose a set of markers of these differences for cross-validation *in situ*. Consistent with the predicted pattern of expression from scRNA-seq, multiplex *in situ* hybridization distinguished adult cells expressing Krt14, both Wfdc18 and Krt8, or Krt8 alone (Figure 2D). By contrast, most cells in the fetal mammary epithelium co-express all three markers (Figure 2D). Similarly, targets identified in scRNA-seq data as being commonly expressed in fetal mammary cells, but not adult cells, e.g., Sostdc1, were identified *in situ* in fetal but not adult tissue (Figure 2D).

These in-depth analyses using tSNE, NMF, and diffusion mapping were concordant in that none was able to identify cellular subpopulations corresponding to the percentage of stem cells estimated by functional assays. Furthermore, they point to a predominance of mixed lineage phenotypes, corroborated *in situ*, in early development that is lost as development progresses.

Epithelial Lineage Precursors Exhibit Balanced Transcription Factor Activity

The above expression patterns suggest that, as development progresses, E18 cells resolve into two distinct cell types by P4. The first represents a basal population, while the second tends toward luminal specification in spite of persistent co-expression of some basal markers, such as Krt14. This population then resolves into the two distinct adult luminal populations. This model is consistent with the multi-potential state of fMaSCs, as well as recent observations from lineage-tracing experiments suggesting that independent Esr1$^+$ and Esrl$^-$ populations of luminal cells are established by adulthood (Van Keymeulen et al., 2011; Wang et al., 2017).

We constructed pseudotemporal trajectories for this simple lineage bifurcation model using principal curves through proximal populations in the first two components of the diffusion map, although the data do not rule out other differentiation trajectories (Figure 2E) (Hastie, 1989). We then combined pseudotemporal ordering of cells with SCENIC analysis to analyze the chronology of transcription program activation over mammary development (Figures 2F and S2C) (Aibar et al., 2017).

Among the differentially activated regulons identified by this approach were correlated sets that were highly expressed in one of the three adult cell types as well as regulons that typify embryonic and/or P4 cells and are downregulated as development progresses (Figure 2F). Each major adult cell type was also characterized by certain repressed regulons, such as Ehf for basal, Creb3l2 for alveolar, and Atf4, Cebpd, Cebpb, and YY1 for ER^+ luminal cells (Figure 2F). While E16 cells were largely negative for adult-associated regulon activity, there was a marked and balanced activation of regulons corresponding to adult cell types at E18 and this balanced pattern persisted into P4 cells (Figure 2F, asterisk). Primitive cells positioned after the branch points of the pseudotemporal trajectories grossly retained a high degree of similarity to their pre-branchpoint counterparts. However, differential regulation of lineage-associated factors could be observed. For instance, elevated Trp63 activity is evident in cells trending toward the basal branch, while Esr1 and Spdef activity is upregulated in P4/NMF-IV cells that trend toward the $Esr1^+$ luminal group (Figure 2F). The luminal determinant FoxA1 was upregulated in most P4 luminal cells but was subsequently downregulated in the $Esrl^-$ luminal cells of the adult (Figure 2F).

A Continuum of Gene Expression Profiles Defines Early Stem Cell State Transitions in the Mammary Gland

To gain a more highly detailed view of changing gene expression patterns over the critical period in early development where multipotent fMaSC generate more committed cells in the postnatal mammary epithelium (Figures 2 and S3) (Makarem et al., 2013a; Spike et al., 2012), we obtained deeper transcriptomes of single cells across the E18-to-P4 transition using the C1 microfluidics platform (Fluidigm), with inclusion of adult cells for reference (Figure 1A; Table S2). For this dataset, we clustered samples by NMF into putative cell types based on their expression patterns of the 1,500 most variably expressed genes (Figure S1J). We found that six clusters provide a suitable division of the data (Figures 3A and S1K). NMF clusters 1–4 almost exclusively comprise cells from a single developmental stage, with cluster 1 predominantly comprising P1 cells, while clusters 2–4 contain adult cells (Figure 3A; Table S2). In contrast, the compositions of clusters 5 and 6 were mixed among E18, P1, and P4 cells, suggesting the presence of cell states that vary in their prevalence across organismal development rather than being strictly defined by the time at which the samples were isolated.

Similar to our Chromium-based analysis, adult cells and E18 cells occupied opposite ends of the resulting graph along one DC axis (DC1), with P1 and P4 cells occupying intermediate positions (Figure 3B). The regions occupied by E18, P1, and P4 cells overlapped significantly in a continuum of transcriptional changes related to advancing organismal age. Adult subsets again occupied the extremes of a second DC axis (DC2), corresponding to luminal and basal lineages based on their expression of lineage-associated keratins (Figures 3B–3D), and could be further subdivided to reveal a population of presumptive alveolar cells despite their lower numbers in this analysis (Figures 3A, 3C, 3E, S4D, and S4E) (Visvader and Stingl, 2014). However, as in Chromium data, the lineage bifurcation noted for adult cells was largely absent from E18, P1, and P4 cells, and again we did not observe discrete subpopulations correlating in number to the predicted stem cell frequencies (Figures 3B–3E). However, we could identify other minority cell types as NMF1 and −4 segregated from the major epithelial clusters along DC5 and DC3, respectively (Figure S4A). In contrast, most of the variance in the remaining four NMF groups was directed along the first two DCs (Figures 3C and 3E).

We next found signature genes distinguishing each major predicted cell type (i.e., NMF1 to −6) using a pairwise, non-para-metric, rank-product (RP) approach (Breitling et al., 2004). We took the genes that were over-represented in each NMF group versus every other NMF group as the group-specific gene signature (Figures 3F and 3G). This unsupervised differential expression analysis led us to denote the small cluster NMF1 as "Matrix"-expressing cells based on their expression of mesenchymal/matrix-related transcripts (e.g., Coll, 3, 5, 6, 12, 15, 18; Fn1; Vim), and to denote NMF4 as "lmmune"-related cells based on their expression of class II transcripts (e.g., B2m; Fcerlg; H2-Aa, Ab1, Eb1, M2, Q6, Q7, T22) (Figure 3G). NMF1 and −4 in this analysis may represent subsets of the non-epithe-lial groups identified from Chromium data (Figures 1 and 2). While the origins and functions of these minority populations remain to be determined, we note that MHC-expressing mammary cells have been described (Elliott et al., 1988: Forero et al.. 2016). The analysis also confirmed the adult basal and luminal phenotypes of NMF2 and −3, respectively, as it re-identified a known basal and luminal markers for each group (Skibinski et al., 2014) (Figure 3G; Table S2).

Identification of fMaSC Expression Programs from scRNA-Seq

The data indicate that NMF5 corresponds to the fMaSC transcriptional state, as (1) DC analysis localizes NMF5 to a distal position relative to adult cells, (2) most NMF5 cells derive from a developmental time point with very high mammary stem cell activity, and (3) the NMF5 signature (857 genes, hereafter "fMaSC signature") includes Eya2, Itag6, Nrg1, Sostdcl, Sox10, Myb, Lsr, Sfrpl, Bcl11a, Pthlh, Sema3B, Slitrk2, and others that we previously identified as highly expressed in bulk fMaSC signatures, including some we have shown to be functionally relevant (Dravis et al., 2015; Spike et al., 2012) (Figures 3G and S4B). The single-cell-derived fMaSC signature also includes many genes not identified in our prior bulk population studies (Spike et al., 2012) (Figures 3G and S4B; Table S3). Gene ontology (GO) enrichment analysis revealed that the fMaSC signature comprises genes involved in a variety of cellular processes of potential importance for stem and progenitor cells, including cellular metabolism, chromatin conformation, cell cycle, and tissue development (Maere et al., 2005; Shannon et al., 2003) (Figure 3H).

As the data imply that fMaSCs are defined both by the unique features captured in the fMaSC signature and by co-expression of lineage-associated factors, we also wanted to identify gene sets reflecting this basal-luminal apposition. To this end, we examined the results of rank product tests between NMF2, −3, and −5 cells (adult luminal, adult basal, and fMASC groups, respectively), to identify a set of genes termed "balancer" signatures (Figure 3I; Table S3). We generated a "basal balancer signature" (i.e., genes expressed in fMASC and basal cells more highly than luminal cells) of 937 genes including known basal markers such as Itga6 and Krt14. Similarly, we generated a "luminal balancer signature" (i.e., genes expressed by both luminal cells and fMaSCs more highly than basal cells) of 477 genes, including, for example, Cd24a and Krt8 (Figure 3I; Table S3). The co-expression of these balancer signatures in individual fMaSC and their lineage-specific expression later in development is illustrated in Figure 3I. Although the specific mechanisms by which gene products belonging to balancer signatures effect the multipotent and uncommitted fMaSC state remain to be elucidated, we computationally identified numerous luminal balancer genes reported to directly interact with basal balancer genes in the MINT interaction database, suggesting that some of the mechanisms are likely to be direct (Figure S5A). The proteins encoded by balancer signature genes extend beyond transcription factors and represent varied aspects of signaling, metabolism, and microenvironmental response (Figure S5A; Table S3). These data, together with our regulon analysis, *in situ* detection of lineage-associated transcripts, the pattern of lineage-associated keratin protein expression, and our previous single-cell RT-PCR analysis support a model in which the fMaSC state is established by balanced expression of lineage factors and specifiers (Figures 1C, 2C, 2D, 2F, and 3I) (Villadsen et al., 2007; Rodilla et al., 2015; Spike et al., 2012; Sun et al., 2010).

Loss of Lineage Factor Balance Shortly after Birth

NMF6 cells principally derive from postnatal mammary epithelia and have an expression signature (234 genes) composed of transcription factors such as Jun, Rela/b, Nfkb2, and Sox4, chromatin modifiers including Arid1a,Top2A, Jmjdlc, and Jmjd1c, cell adhesion and cell-cell junction proteins including claudins 1 and 6, Icam, and many others (Figure 3F; Table S2). Although they comprise a single group in the initial analysis, NMF6 cells are more dispersed along the DC2 axis than the NMF5/fMaSC group. They also exhibit less balanced expression of the lineager balancer signatures described above, suggesting NMF6 cells might be divisible with regard to their relative luminal and basal characteristics (Figures 3C and 3I). Thus, while a few early postnatal cells have already committed to a basal fate and were grouped with NMF3, the majority of postnatal cells comprise NMF6 and appear to represent a mixed mammary precursor/progenitor (hereafter, MMPr) (Figures 1D, 1E, 3A–3C, 3E, and 3I; Table S2).

In light of this, and as fMaSC/NMF5 and MMPr/NMF6 are relatively large groups (76 and 66 cells, respectively), we repeated the divisive NMF procedure on NMF5 and NMF6 independently to determine whether they show early lineage bifurcation events. We could subdivide NMF5 and NMF6 into two subclusters each with high cluster stability (NMF5a and -b, CC = 0.96; NMF6a and -b, CC = 0.88) (Figure 3A). We then examined their relative positions in the diffusion map and their differential gene expression (Figures 3C and 3E). While these subclusters occupied largely overlapping regions, NMF5a and -b were centered at different points along the embryonic-adult axis (DC1) and NMF6a and 6b were centered at different points on the basal-luminal axis (DC2) (Figures 3C and 3E).

To uncover the basis of these distributions, we delineated rank-product gene expression differences for NMF5a versus −5b and NMF6a versus −6b, and their potential cell fate and type distinctions (Table S2). The MMPr groups are characterized by genes indicative of lineage specification including Cdh1 and Krt8 in NM6a (luminal) and Acta2 and Krt14 in NMF6b (basal) (Table S2). Consistent with this, there is also a positive correlation between the expression of luminal and basal balancer signatures and the genes subdividing these MMPr groups (Figures S4C and S4F). We therefore designate these MMPr subclusters and their signatures as MMPr-basal and MMPr-luminal according to their correlated balancer signatures (Figures 3G and 3I). In contrast, the data suggest that NMF5 is composed of a transcriptionally heterogeneous population not readily separated into luminal and basal subtypes. Rather, when subclusters from NMF5 (NMF5a and -b) were derived and contrasted, their differentially expressed genes correlated with graded expression of the fMaSC signature derived from the undivided NMF5 cluster (Figure 3J). We therefore sought to investigate select features represented in the fMaSC signature that distinguish fMaSCs from their adult counterparts.

Changes in Chromatin Regulation and Accessibility Accompany fMaSC Differentiation

The GO enrichment for chromatin regulators was of particular interest as it could help explain the multipotent expression pattern identified in fMaSCs (Figure 3H). To determine whether the fMaSC state is accompanied by altered chromatin states relative to adult epithelia, and to determine whether lineage-associated loci were heterochromatic, we examined global chromatin accessibility by assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Andrews, 2010; Bolstad, 2018; Buenrostro et al., 2015; Langmead et al., 2009; Li et al., 2009; McLean et al., 2010; Quinlan and Hall, 2010; Ramirez et al., 2014; Zang et al., 2009; Zhang et al., 2008). E18 cells have uniquely accessible regions (UARs) corresponding by proximity to 1,640 genes, and uniquely repressed regions (URRs) corresponding to 401 genes when compared to flow-sorted adult basal (Cd49f^{+}, Epcam^{l} $^{ow/med}$), luminal progenitor (Epcam^{+}, Cd61^{+}), and mature luminal (Epcam^{+}, Sca1^{+}) cells (Figures 4A–4D and S5B; Table S3). Genes at these loci correspond to developmental and metabolic processes and show significant overlap with fMaSC signature genes (Figures 4C and S5C). Overall, E18 cells show greater accessibility at fMaSC signature genes as expected (Figure S5D). Most fMaSC signature genes, however, were not uniquely accessible in E18 cells, suggesting other regulatory mechanisms of transcriptional control in response to changing signals in the developmental environment (Figure S5C). Interestingly, P4 cells showed intermediate accessibility across these loci (Figures 4A, 4B, and S5B).

We also identified UARs and URRs corresponding to each major adult lineage sorted by fluorescence-activated cell sorting (FACS) (Figures 4D and S5B) (Dravis et al., 2018). A large number of loci distinguished each lineage from the others in terms of both UARs and URRs (Figure S5A). fMaSCs exhibit modest and equivalent accessibility for both basal and luminal progenitor-associated loci, although they lack significant accessibility at loci associated with Epcam^{+}Sca1^{+} luminal cells. Even more strikingly, fMaSCs show accessibility at adult lineage URRs, suggesting that in spite of their more universal expression of Ezh2 and associated chromatin methylation (Figures S5E and S5F), fMaSCs have not yet silenced chromatin regions that define discrete lineages in the adult (Figures 4A–4C).

Closer inspection of lineage-associated genes illustrates the multilineage accessibility of fMaSCs compared to adult cells (Figure 4D). For example, while Sca1^{+} luminal cells show little accessibility in Krt14 and Csn3 proximal regions, Csn3 is accessible in Cd61^{+} luminal cells. Conversely, basal adult cells (Cd49f^{+}, Epcam$^{low/med}$) have lower accessibility at Krt8 and Csn3 than either luminal population, although basal cells generally have more open chromatin across multi-lineage markers than luminal cells (reported in detail in Dravis et al., 2018) (Figure 4D). Chromatin accessibility of P4 cells indicate that the FACS-sorted basal cells show greater accessibility at basal associated loci (e.g., Krt14) and luminal cells show a greater accessibility at luminal associated loci (e.g., Krt8, Csn3). However, each sorted population also shows appreciable accessibility for genes of the opposing lineages, indicating the P4 cells remain at an intermediate stage of lineage commitment (Figures 4B and 4D). These chromatin patterns mirror the gradual lineage restriction implied by diffusion maps of their associated single-cell transcriptomes (Figures 2F and 3G).

fMaSC Metabolic Profiles Are Lost during Differentiation

Metabolic factors are overwhelmingly represented among the fMaSC-signature gene ontologies (Figure 3H). Many of the fMaSC-signature genes encode enzymes involved in glycolytic metabolism, the Krebs cycle, and enzymes and transporters involved in anaplerotic mechanisms such as fatty acid oxidation. Therefore, we modeled how flux through these pathways might change as a function of development-associated changes in gene expression (Figure 5A). We noted elevated expression of several glycolytic enzymes in fMaSCs and MMPr cells relative to adult cells, but also a potential for shunting the glycolytic end-product, pyruvate, away from the mitochondria and oxidative phosphorylation, as in the stem cell- and tumor-associated Warburg effect (Figure 5A) (ShyhChang et al., 2013). fMaSCs have increased RNA for lactate dehydrogenase (Ldh), and ~90% of fMaSCs express Pkm2, a splice isoform of pyruvate kinase (Pkm). Pkm2 incorporates exon 10 rather than exon 9, reported previously to increase lactate formation (Li and Dewey, 2011; Mazurek et al., 2005; Robinson et al., 2011) (Figures 5A–5C). Surprisingly, all fMaSCs co-express the Pkm1 isoform. The majority of adult cells express either Pkm1 or Pkm2 or had Pkm levels below isoform detection limits (Figure 5C). We also noted marked elevation in embryonic cells of transcripts for several Krebs cycle enzymes and factors capable of providing acetyl-CoA to the mitochondria via free fatty acids. These observations are consistent with the evolving understanding that “Warburg-shifted” cells use alternative mechanisms to fuel Krebs cycle reactions (ShyhChang et al., 2013) (Figure 5A). E18 cells also have elevated Psatl and Gpt RNA, encoding two enzymes that can balance interconversion of pyruvate and glutamate with alanine and α-ketoglutarate (Figure 5A) (Coloff et al., 2016). While numerous metabolism-related transcripts associated with E18 mammary epithelium were also elevated in E16 cells, important differences were noted. For instance, Psat1 and Gpt transcripts were also elevated in E18 relative to E16, while E16 cells showed elevated levels for the glutamine transporter, Slc38a1 (Figure 5A).

The correlation between changing stem cell content in early mammary development and the above changes in expression of metabolic factors led us to hypothesize that differentiating cells would be more dependent on mitochondrial pyruvate transport than stem cells, an effect described previously in other systems (Flores et al., 2017; Schell et al., 2017). We therefore tested the prediction that inhibiting the mitochondrial pyruvate transporters (Mpc1, −2) should differentially affect stem and differentiating cells using *in vitro* clonal organoid formation assays in the presence and absence of a potent and specific Mpc inhibitor, UK5099 (Sigma-Aldrich) (Gray et al., 2014). UK5099 treatment significantly alters organoid composition in differentiation-promoting media where presumptive lineage-committed cells contribute significantly to organoid expansion (Spike et al., 2014) (Figures 5D, S6A, and S6B). Multicellular fMaSC-derived organoids grown in these conditions contain significantly fewer cells exclusively expressing Krt8 or Krt14 and an increased proportion of Krt8$^+$Krt14$^+$ co-expressing cells. This effect is associated with a partial reduction in organoid-forming efficiency and size that is not observed in a maintenance media where stem cell capacity is preserved during culture (Spike et al., 2014) (Figures 5E, 5F, and S6B).

The fMaSC Metabolic Program Has Parallels in Triple-Negative Breast Cancers and Metastases

We previously showed that bulk fMaSC signatures are related to expression profiles from human breast cancer patient samples (Pfefferle et al., 2013; Pfefferle et al.,. 2015; Spike et al., 2012). We therefore asked whether the refined single-cell-derived fMaSC signature and its metabolic components demarcate the same breast cancers (Cancer Genome Atlas, 2012; Ciriello et al., 2015; Forero et al., 2016; Forero-Torres et al., 2015; Parker et al., 2009; Varley et al., 2014). As we observed previously, fMaSCs share expression signature similarity with basal-like breast cancers, as well as occasional cancers in other aggressive subtypes such as luminal B and Her2 tumors (Figure 6A) (Spike et al., 2012). However, only Her2 tumors and basal-like tumors showed frequent elevation of fMaSC-like metabolic profiles, whereas the equally proliferative luminal B tumors did not (Figure 6B). We also generated a small eight-gene metabolic signature (Metab-8) to reflect metabolic processes proximal to the mitochondria in our fMaSC metabolism model (Figure 5A; Table S3). This signature was composed of the last three steps in lactate-directed glycolysis (Eno, Pkm, Ldh), three fatty acid pathway genes (Fabp5, Hadh, Acat), and genes controlling the balance between pyruvate, cytosolic α-ketoglutarate, alanine, and glutamate (Psat1, Gpt) (Figure 5A). The median expression of Metab-8 was highest in basal-like cancers, which include a subset of the most proliferative cancers and most of the clinically designated triple-negative breast cancers (TNBCs) (Figures 6C and 6D). Furthermore, among TNBCs, Metab-8 was highest in TNBC metastatic lesions (Figure 6E).

DISCUSSION

This work provides a comprehensive scRNA-seq resource for the developing mammary gland. We mine the data to show that multipotent fMaSCs are typified by co-expression of lineage-associated factors and transcriptional programs reminiscent of opposing lineage specifiers described in embryonic stem cells (Loh and Lim, 2011). While the present work was under review, mixed-lineage expression patterns were reported for mammary cells as early as E14 (Wuidart et al., 2018). Although E14 cells lack stem cell activity by *in vitro* colony assays and transplantation of single cells, we have previously shown that transplant of intact rudiments at this stage does reconstitute mammary tissue, presumably through preservation of spatial cues and acquisition of repopulating capacity in the days subsequent to transplant (Spike et al., 2012). This suggests that additional factors are required to confer intrinsic mammary stem cell competence and subsequent lineage segregation.

At the population level, the fMaSC transcriptional profile is reflected in open chromatin across lineage-associated genes in spite of high PRC2 activity. It remains to be determined whether the PRC2 machinery is only operative in these cells at lineage-unassociated loci or, alternatively, whether there are regulatory mechanisms directly opposing PRC2-facilitated lineage restriction. Ultimately, as differentiation ensues a lineage bifurcation events are revealed where this balance is lost and opposing lineage genes become restricted. Thus, a continuum of E18 cell states resolve into two discrete populations by P4. One represents a distinct basal (myoepithelial) phenotype, and the other represents a luminal oriented precursor population that is itself a continuum between cells retaining basal features and more mature, luminally oriented expression profiles. The retention of mixed basal features in this majority population may have contributed to the apparently incorrect interpretation that prepubertal mammary epithelium is composed of basal oriented cells (Pal et al., 2017). As development progresses further, these mixed mammary precursors are replaced by discrete luminal cell types corresponding to Esr1^{+} and Esrl^{-} lineages. Although basal cells form a transcriptionally distinct cell type in the adult, it is interesting to note (as in Dravis et al., 2018) that they show some chromatin accessibility at luminal gene loci similar to fMaSCs. We speculate that this favors lineage plasticity and enables acquisition of multilineage potential upon transplantation, in wounding conditions or following *ex vivo* culture (Ge et al., 2017; Prater et al., 2014; Shackleton et al., 2006; Stingl et al., 2006).

Despite the similarities between fMaSC and adult basal populations in chromatin accessibility, fMaSCs are transcriptionally distinct from basal and other adult cells, and this extends beyond their active co-expression of lineage-specific genes. At the single-cell level, they show expression of factors that may relate to their developmental plasticity and proliferative potential, as well as to their connection to human breast cancers. In this regard, we examine metabolic transcript profiles in fMaSCs and show that they are shared by human breast cancers. Prompted by the changing pattern of these profiles over developmental time, we demonstrate a changing sensitivity to UK5099 as cells differentiate *in vitro*. We take these data to indicate that differentiating cells but not bipotent stem/progenitor cells critically depend upon mitochondrial pyruvate uptake. However, tools are not yet available to assess metabolomics directly at the resolution of single cells, and at this time we cannot rule out the possibility that the observed differential sensitivity relates to other effects of UK5099, although its affinity for secondary targets is several hundred-fold lower than for mitochondrial pyruvate carrier (MPC) (Gray et al., 2014).

Importantly, although proliferation is often associated with enhanced glycolytic rate, the fMaSC metabolic gene profile does not seem to be strictly linked to proliferation. For instance, there are many proliferative cells and tumor types that lack this specific metabolic profile (Figures 6, S6C, and S6D). In single-cell-derived organoids, Krt8^{+}Krt14^{+} cells increase in number and proportion under UK5099 treatment, while the generation of presumptive lineage committed cells is blocked. Still, it would not be surprising if the different metabolic programs lead to different rates of proliferation and smaller colonies in the absence of lineage committed progenitors. Most organoids produced from adult cells express only Krt8, but UK5099 treatment enabled production of organoids containing both basal and luminal cells indicative of effects on plasticity beyond the possible proliferative effects (Figure S6B). It will be important to determine whether this fMaSC-related profile in tumors helps pinpoint metabolic liabilities for more precise therapeutic targets than those aimed at all proliferating cells.

Our scRNA-seq analysis also led us to important observations on the nature of transcriptional heterogeneity in uncommitted cells. Although mammary stem cell estimates vary considerably depending upon the markers and assay used to assess their potential, the fetal cell population from which we derived our single-cell data exhibits 30%−50% stem cell activity measured by *in vitro* sphere formation, and 10%–30% activity measured by transplantation (Dravis et al., 2015; Spike et al., 2012; Trejo et al., 2017) (Figure S3). Despite this prevalence, our analyses did not reveal an equivalent transcriptionally distinct stem cell subpopulation. Similarly, although stem cell content is much lower in the adult basal population (~2%), our transcriptional profiling of >1,000 adult basal cells should have enabled us to identify a transcriptionally distinct stem cell population. However, we did not. Our detection of other rare transcriptional cell types suggests this was not due to the limitations of the sequencing and bioinformatics strategies employed. Work by Bach et al. (2017) also found the adult basal compartment to be relatively indivisible at the transcriptome level. We acknowledge that our analysis of stem cell content and the above scRNA-seq data would exclude any epithelial cells lacking Epcam expression.

We consider it likely that stem cell activity is probabilistically distributed throughout the heterogeneous cell population and is likely to be highly dependent on external cues for its maintenance (e.g., Spike et al., 2014). Performance in functional assays may therefore critically depend on the context under which a particular cell with stem cell potential is challenged (Spike, 2016; Wahl and Spike, 2017). This idea may help to refine a long-standing concept about the nature of compartmentalized functions and rigid cell types in complex tissues, and shed light on obdurate limitations of stem cell purification by surface markers. In our data, markers were often enriched in a given transcriptionally defined cell type but were never perfect, differing in their levels, being undetectable in some cells, and being expressed occasionally in unexpected cell types. Transcriptional heterogeneity may also help explain discrepancies in lineage tracing that have suggested lineage-restricted stem cells or multipotent stem cells depending on the chosen promoter for tracing, context, methodology, and interpretation.

Transcriptional heterogeneity associated with variable stem cell potential has also been identified in embryonic stem cells, intestine, lung, and skin, where perturbations of homeostasis can promote facultative stem cell activity and create functional states and niches (Blanpain and Fuchs, 2014; Hough et al., 2009). Expression profiles and their discreteness could be strongly influenced by environmental context. Indeed, it is tempting to speculate that a lack of a homeostatic niche environment is the critical feature linking early development, wounding, and cancer in the generation of cellular heterogeneity, mixed lineage phenotypes, and associated cellular plasticity among cells that may already harbor the intrinsic flexibility to respond (e.g., Ge et al., 2017, and the present study). Transcriptional heterogeneity and plasticity are likely to be even greater in tumor settings due to genomic instability, therapies, and inconstant microenvironments. It may be necessary to map the continuum of stem cell-like states to understand how they contribute to cell fitness in particular settings. This may also enable the development of effective therapeutic combinations directed against plastic cell states that critically contribute to intra-tumoral heterogeneity during cancer progression.

The present work provides a resource for uncovering functionally relevant mechanisms of stem cell regulatory biology with coopted roles in tumorigenesis. Based on our analysis of the data, it also has the potential to impact the way we conceptualize "cell type" in the mammary gland and other complex tissues, and to replace classical models based on rigid cell hierarchies with a more fluid, physiologically adaptable, and robust, if experimentally challenging, idea.

STAR ★ METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Alexa Fluor 647 anti-CD326 (Epcam)	Biolegend	Cat# 118212; RRID:AB_1134101
FITC anti-Cd49f (ITGA6)	Stem Cell Technologies	Cat# 60037FI.1; RRID:AB_2734790
Anti-Krt14	Biolegend	Cat# 905304; RRID:AB_2616896
Anti-Krt8(TROMA-1)	DSHB Univ. of Iowa	Cat# AB-531826; RRID:AB_531826
Anti-H3K27me3	Millipore	Cat# 07–449; RRID:AB_310624
FITC-anti rat lgG2	Thermo Fisher	Cat# PA1–84761; RRID:AB_933936
PE-anti rabbit IgG	Santa Cruz Biotechnology	Cat# SC-3739; RRID:AB_649004
Anti-BrdU	Bio-Rad	Cat# MCA2060GA; RRID:AB_10545551
Chemicals, Peptides, and Recombinant Proteins		
UK5099	Sigma Aldrich	Cat# PZ0160
bFGF	Stem Cell Technologies	Cat# 78003
EGF	Sigma Aldrich	Cat# E4127
DAPI	Thermo Scientific	Cat# 62248
Urea	Sigma	Cat# U5378
N.N.N’N’-tetrakis (2-hydroxypopryl) ethylenediamine	Sigma	Cat# 122262
Polyethylene glycol mono-p-isooctylphenyl ether/ Triton X-100	Sigma	Cat# 93443
2,2,2’-nitrilotriethanol	Sigma	Cat# 90279
Critical Commercial Assays		
Chromium prep	10x Genomics	Cat# 120237
Nextera DNA library kit	Illumina	Cat# FC-121–1030
TapeStation DNA high sensitivity kit (D1000)	Agilent	Cat# 5067–5585
NEBNext High Fidelity 2x PCR mix	NEB	Cat# M0541
RNAscope Multiplex Fluorescent v2	ACD Bio, Newark CA	Cat# 323110
SMARTer Ultra Low RNA kit	Clontech	Cat# 634936
SMART-Seq® v4 Ultra® Low Input RNA Kit for Sequencing	Clontech	Cat# 634888
Advantage 2 PCR Kit	Clontech	Cat# 634206
C1 Single-Cell auto Prep Reagent Kit for mRNA Seq	Fluidigm	Cat# 1006201
Bioanalyzer RNA Pico Kit	Agilent	Cat# 5067–1513
Quanti-it Pico-green dsDNA assay kit	Thermo Fisher	Cat# P11496
LIVE/DEAD kit	Invitrogen	Cat# MP03224
AM Pure XP beads	Agencourt	Cat# A63880
C1 Single-Cell mRNA Seq IFC, 10–17 μm	Fluidigm	Cat# 100–6041
ERCC RNA spike in mix	Thermo Fisher	Cat# 4456740
Illumina Nextera XT DNA sample preparation kit	Illumina	Cat# FC-131–1096
Illumina Nextera XT DNA sample preparation index kit	Illumina	Cat# FC-131–1002
Deposited Data		
FastQ sequencing files	NCBI sequence read archive and GEO https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111113 and https://trace.ncbi.nlm.nih.gov/Traces/sra/	SAMN07138894; GSE111113
Experimental Models: Organisms/Strains		
C57BL/6	Charles River	Strain Code: 027
CB17/lcr-Prkdcscid/lcrlcoCrl (SCID)	Charles River	Strain Code: 236
Software and Algorithms		
Bowtie	http://bowtie-bio.sourceforge.net/	Langmead et al., 2009

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MACS2	http://liulab.dfci.harvard.edu/MACS/	Zhang etal., 2008
Bedtools	http://bedtools.readthedocs.io/en/latest/	Quinlan and Hall, 2010
Samtools	http://www.htslib.org/doc/samtools.html	Li etal., 2009
Deeptools	https://deeptools.readthedocs.io/en/latest/	Ramirez etal., 2014
GREAT	http://great.stanford.edu/public/html/index.php	McLean et al., 2010
SICER	https://home.gwu.edu/~wpeng/Software.htm	Zang etal., 2009
preprocessCore (R)	https://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html	Bolstad, 2018
Plotly	Plotly Technologies; https://plot.ly	N/A
rtsne	https://cran.r-project.org/web/packages/tsne/	van der Maaten and Hinton, 2008
NMF (R)	https://cran.r-project.org/web/packages/NMF/	Lee and Seung, 1999
SCENIC	https://github.com/aertslab/SCENIC	Aibar etal., 2017
R 3.3.0 (for Mac OsX)	https://cran.r-project.org/bin/macosx/old/	R-3.3.0.pkg
Destiny (Diffusion Maps)	https://bioconductor.org/biocLite.R	Haghverdi etal., 2015
TM4-MeV4.8	mev.tm4.org	Saeed etal., 2003
Non-Negative Matrix Factorization	mev.tm4.org	Lee and Seung, 1999
Rank products	mev.tm4.org	Breitling etal., 2004
Cytoscape 3.3.0	www.cytoscape.org	Shannon etal., 2003
BiNGO	http://apps.cytoscape.org/apps/bingo	Maere etal., 2005
FastQC 0.11.2	http://www.bioinformatics.babraham.ac.uk/projects/fastqc	Andrews, 2010
RSEM 1.2.29	https://github.com/deweylab/RSEM/releases	Li and Dewey, 2011
STAR 2.4.2a	https://github.com/alexdobin/STAR	Dobin etal., 2013
IGV2.3.83	https://software.broadinstitute.org/software/igv/	Robinson etal., 2011
ELDA	http://bioinf.wehi.edu.au/software/elda/	Hu and Smyth, 2009
Other		
Epicult-B Basal Medium (mouse)	Stem Cell Technologies	Cat# 05611
Epicult-B Proliferation Supplement (mouse)	Stem Cell Technologies	Cat# 0562
B27 Supplement	GIBCO	Cat# 17504044
Hydrocortisone	Sigma Aldrich	Cat# H4001
Collagen ase/Hyaluronidase	Stem Cell Technologies	Cat# 07912
Dispase	Stem Cell Technologies	Cat# 25300–054
Trypsin	GIBCO by Life Technologies	Cat# 25300–054
Fetal Bovine Serum	Serum Source	Cat# FB22–500
Matrigel (complete)	Corning	Cat# 354234
Matrigel (growth factor reduced)	Corning	Cat# 356231
DMEM-F12 (no phenol red)	Thermo Fisher	Cat# 21041025

CONTACT FOR REAGENT AND RESOURCE SHARING

Requests for further information and resources may be directed to Lead Contact Benjamin T. Spike (benjamin.spike@hci.utah.edu)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice—C57BL/6 mice were obtained from Charles River as 10–16 week old adults or as timed pregnant mothers, or were bred in house to produce fetal, postnatal and adult tissues. Three-week old CB17 SCID mice were also obtained from Charles River as recipients for transplantation experiments and 8–10 week old CB17 SCID mice were obtained from

Charles River as recipients. All mice used in these studies were female. All animals were handled in accordance with IACUC/AAALAC and other institutional ethics guidelines.

METHOD DETAILS

Isolation of Single Cells—Mammary tissue was dissected from C57BL76 mice at the developmental stages indicated. The number 4 mammary gland from 10 to 16-week-old nulliparous C57BL/6 female mice was used for adult samples. Single cell suspensions were obtained as described previously (Spike et al., 2012). All dissociation reagents were purchased from Stem Cell Technologies (SCT), unless otherwise specified. Freshly dissected tissues were placed directly in EpiCult-B basal medium containing hydrocortisone. Following dissection, minced adult mammary glands were digested for 8–12 hours at 37°C in EpiCult-B basal medium containing Epicult-B proliferation supplement, collagenase and hyaluronidase and hydrocortisone. Mammary organ fragments resulting from overnight digestion were treated briefly with ammonium chloride solution (4min on ice), then 0.25% trypsin (5min RT) followed by dispase (4min 37°C), with washes and resuspension after each treatment using Hank’s Balanced Salt Solution (HBSS) supplemented with 2% FBS (Chemicon). Cell clumps and debris were removed by passing the suspension through a 40 μm nylon filter (BD Biosciences). For fetal and postnatal tissues, 150 mammary rudiments on average were dissected on each experimental day. Dissected fetal mammary glands were processed as adult cells except that 90 minutes incubation in collagenase/hyaluronidase was sufficient to obtain single cell suspensions and trypsin was found to be unnecessary and was therefore omitted to avoid unnecessary stress on the cells or alterations to their transcriptomes. Cells were immunostained with antibodies to Epcam, Cd49f, for 20 min on ice where indicated. Antibody-labeled cells were resuspended and incubated in HBSS with 2% FBS containing DAPI for live/dead discrimination. Cell sorting was carried out on a FACSDiva cell sorter to collect Epcam$^{\text{low-high}}$Cd49f$^{\text{medium-high}}$ cells for further analysis (Becton Dickinson).

Organoid Culture—Single cell suspension (as above) from mammary tissues at the developmental stages indicated were sorted for Epcam expression and plated in 4% Matrigel on a pre-congealed undiluted Matrigel bed in either Maintenance media or Differentiation media (Spike et al., 2014) supplemented where indicated with UK5099 at the doses indicated. Maintenance media was comprised of Dulbecco’s modified Eagle’s medium/F12 with 5% horse serum, 10 μg/ml insulin, 20 ng/ml epidermal growth factor, 100 ng/ml cholera toxin, 0.5 μg/ml hydrocortisone, and 10 μg/ml ciprofloxacin with 1 × B27 supplement (Invitrogen). Differentiation media was composed of Epicult-B mouse media containing B supplement (SCT), recombinant human epidermal growth factor, recombinant human basic fibroblast growth factor, and heparin, as previously described (Spike et al., 2012), with ciprofloxacin 10 μg/ml. Organoids were fixed after 7 days of growth for immunofluorescent staining.

Immunofluorescence—Samples were fixed in 4% formalin (NBF) at 4°C overnight, permeabilized using 0.1 % BSA, 0.2% Triton X-100, 0.05% Tween-20 in PBS, blocked in 10% goat serum and stained with antibodies against Keratin 8, Keratin 14 as previously described (Spike et al., 2012) prior to staining with secondary antibodies and imaging on a Zeiss LSM 880 with Airyscan FAST. Samples were counterstained with DAPI.

In Situ RNA-FISH—RNA-FISH was performed using RNAScope Multiplex Flourescent V2 Kit (Advanced Cell Diagnostics, catalog number 323110). The protocol was followed as per the manufacturer's recommendations with the target retrieval boiling time for 15 min and Protease IV at 40 °C incubation for 30 min. Slides were mounted with Slowfade Mountant+DAPI (Life Technologies, S36964) and sealed. All images were captured within one week of slide preparation. A bacterial gene probe was used as a negative control as per kit instructions. Images were captured on Carl Zeiss 880 Airyscan Super-Resolution microscope using 40X/1.2NA W objective with 1.8X digital zoom. All pictures were digitally edited to enhance the color and contrast levels using Zen (Carl Zeiss) and ImageJ (Fiji) software.

Transplantation—Single cell suspensions of primary mammary cells were obtained as above and varying numbers of cells were transplanted into de-epithelialized number 4 fat pads of recipient three-week old CB17 SCID mice as previously described (Spike et al., 2012). Glands were subsequently dissected, mounted on glass slides and fixed at 8 weeks post injection and were stained with Carmine Alum to score outgrowth. Frequency of the repopulating unit was estimated using Extreme Limiting Dilution Analysis (ELDA).

Single Cell RNA-Sequencing

Production of cDNA Libraries

Microfluidic Assay (Fluidigm C1).: FACS-sorted, Epcam-positive mammary cells were resuspended in 20–30ul of ice cold HBSS/FBS. A 2ul volume of the cell suspension was manually counted between a coverslip and slide and the remaining cells were diluted to to 250 cells/ul for loading onto Fluidigm C-1 platform microfluidic chips with with 10–17um capture well sizes. Loading and staining of cells with the Invitrogen LIVE/DEAD kit (i.e., Calcein-AM and EtBr) to distinguish viable from dead cells was carried out according to the manufacturer's instructions. Following loading and staining, all capture wells were imaged in green and red fluorescent channels and brightfield including z axis examination on a Zeiss 710 or 780 confocal microscope. Each well was then scored as containing a single live cell (Calcein+), single dead cell (Calcein-), multiple cells or no cells. Mean live, single cell capture efficiency was 73% across E18-adult (a representative image is given in Figure S2). After imaging, the chip was processed according to the manufacturer's instructions to produce cDNA libraries from each well. Initial experiments employed ERCC RNA spike in controls (dil. 1:40,000), but it was determined that they provided no benefit over normalization according to cellular transcript abundance and spiked in controls were omitted from select runs to maximize sequencing of cellular transcriptomes (see analysis below, Figure S2, and Dillies et al., 2013, and Lin et al., 2016). cDNA yield was determined by Quant-iT Picogreen fluorescence on a FlexStation II (Molecular Devices). Libraries with sufficient yield were diluted to 100pg/ul in water. Controls were also generated from pellets of ~1000 cells in parallel in PCR tubes as recommended in the Fluidigm C1 protocol. The capture efficiency of E16 cells was low for unknown reasons and individual E16 cells were therefore processed manually by first sorting single cells in 2ul starting volumes in 96 well plates and following tube control ratios for Smart-seq2 (Clontech) and Nextera XT reagents.

Drop-Seq Platform (1 Ox Genomics Chromium).: FACS-sorted, Epcam-positive mammary cells were resuspended in 500ul of HBSS. Cells were centrifuged at 4°C for 5 minutes at 1500rpm and resuspended in 32.5ul of HBSS and processed immediately and loaded on the microfulidic chip together with barcoded beads and other reagents as described in the 10X Chromium Single Cell Reagent Kit V2 protocol (Cat# 120237,10X Genomics Inc). Subsequent cell lysis, first strand cDNA synthesis and amplification were carried out according to the instructions with cDNA amplification set for 12 cycles. cDNA quality was measured using TapeStation (Agilent Biosystems) after bead-based purification.

Production of Sequencing Libraries

ATAC-Seq Library Preparation.: The ATAC-seq transposition assay was performed as previously described with minor modifications (Buenrostro et al., 2015). Two biological replicates from each cell population were assayed to ensure reproducibility. For the adult populations, 2×10^4 nuclei were subjected to 2 μL TDE1 digestion in 20 μL reaction mix, while for the P4/fMaSC populations, 1×10^4 nuclei were subjected to 1 μL TDE1 digestion in 10 μL reaction mix (lllumina Nextera FC-121–1030). The equal Tn5:cell ratio is crucial to ensure similar signal-to-noise ratio during downstream analysis. The cell-Tn5 mix was incubated at 37°C for 30 minutes. qPCR was performed to determine the cycle number for 25% library saturation. Typically, 10–14 total cycles were performed. The library was purified with AMPure XP beads (Beckman A63881), and then analyzed by Agilent TapeStation to ensure proper digestion.

Microfluidic Library Preparation.: For C1 libraries, 125 pg of each diluted library was used as input for Nextera tagmentation and barcoding using volumes recommended by Fluidigm (https://www.fluidigm.com/productsupport/c1-support-hub). Subsequently, single cell Nextera libraries were individually purified using Ampure XP beads (Agencourt) and magnetic force at a library:bead-suspension ratio of 10:9. Bead-pellets dried just to the point of visible cracking were resuspended in Tris/0.1mM EDTA/0.05% Tween-20 to elute libraries. Nextera libraries were quantified by Quant-iT Picogreen fluorescence as above and fragment sizes were determined on a 2100 bioanalyzer (Agilent Genomics) (Figure S1D). Libraries were combined at equal ratios into pools of 50–100 uniquely bar-coded samples which were again precipitated using ampure beads, evaluated for fragment size using the bioanalyzer and quantified by Quant-iT Picogreen.

Drop-Seq Library Preparation.: Sequencing libraries were prepared as per the manufacturer’s protocol (Cat# 120237, 10XGenomics Inc) with Index-PCR set for 14 cycles. Qualities of the sequencing libraries were measured using TapeStation (Agilent Biosystems) after bead-based purification.

Sequencing

ATAC-Seq Library Sequencing.: The ATAC-seq libraries were sequenced with 50 or 125 bp single- (P4) or paired-end (fMaSC and adult) lllumina HiSeq 2500.

Microfluidic Library Sequencing.: For C1 derived libraries, pooled library were loaded at 12–20pM on an Illumina HiSeq 2500 High Throughput Sequencing System and sequenced using a paired-end, 100bp+ read protocols. We obtained 160–260 million reads per sequencing lane totaling 1.4–14 million reads per sample. Sequencing quality was assessed using FastQC software (Ref. 7) and all samples exhibited acceptable sequence quality including base wise quality scores > 30 over the majority of the read length.

Drop-Seq Library Sequencing.: Pooled libraries were sequenced on the Illumina HiSeq2500 Rapid Sequencing (Illumina Inc) System as per instructions provided in the Drop-Seq protocol (Cat# 120237, lOXGenomics Inc). Sequencing quality was assessed using FastQC software and all samples exhibited acceptable sequence quality including base wise quality scores > 30 over the majority of the read length.

Sequence Data Submission: C1 derived FastQ files and the filtered normalized expression matrix (262 cells × 13355 genes) are available at the NCBI sequence read archive https://trace.ncbi.nlm.nih.gov/Traces/sra/ under the BioSample accession SAMN07138894. Chromium Sequencing data is archived under GSE111113 (Bioproject PRJNA435951; SRA: RP133477).

Analysis

Reproducibility: Chromium data yielded coherent clusters of cells corresponding to established cell types in the adult gland. For example, basal cells from three independent isolates co-clustered by each of the approaches we employed suggesting biological differences supersede technical variation in the data and analysis. C1 derived single cell samples similarly identified known cell types and samples run on duplicate sequencing lanes showed an average r^2 value of > 0.99 between replicates (Figure S1E). Select E18 samples were processed independently from the cDNA stage onward and were run independently in a subsequent sequencing run. The resulting expression profiles bore an average r^2 value of 0.965 when compared to their initial profiles. Averaging (Geometric means of values +1 count each) of single cell samples showed high correlation to samples processed as pools (Figure S1G). Potentially as a function of our dual pre- and post pooling library cleanup procedure, barcoding mismatches were negligible as judged by spill over of ERCC reads to samples lacking ERCC at preparation, i.e., was determined to be <0.04% based on means for counts in non-ERCC samples co-processed with ERCC containing samples.

Mapping and Normalization: We mapped reads to a custom mouse transcriptome compatible with RSEM and comprised of gene models from MM10 sequences and sequences for the ERCC RNA spike-in set. Transcript counts were enumerated using RSEM. This pipeline is presented in the accompanying “Scripts and Command Line Procedures” document (Data S1) under the perl script “RNASIuice.pl.” As we noted variable total counts per cell, we applied inter-sample normalization. Since previous studies showed that Upper Quantile normalization compares favorably to other normalization approaches including RPKM, DESeq and TMM except for sensitivity to highly overexpressed genes (Dillies et al., 2013; Lin et al., 2016), we used a derivative approach, normalizing samples for read depth based on the 19th ventile of expressed transcripts (i.e., count > 0; excluding ERCC spike-ins if present). We first determined the sum of counts in the 19th ventile of expressed transcripts for each sample (σ_j), and then defined a standard value (α) near the lower end of σ values in the dataset to which samples were subsequently normalized (e.g., ~1.5 stdev below the mean σ). We then calculated a coefficient *(W)* for normalizing each sample (j) where,

$$W_j = \frac{\alpha}{\sigma_j}$$

We then multiply each gene's counts in each sample ($C_{i,j}$) by the sample specific coefficient and round the values to integers to obtain the normalized expression value of each gene in the dataset ($\mathrm{Exp}_{i,j}$):

$$\mathrm{Exp}_{i,\,j} = \mathrm{int}\left[W_j * C_{i,\,j}\right]$$

In addition to its relative computational simplicity and avoidance of perturbations from very highly expressed genes, this approach has the added advantage of avoiding complications from lowly expressed genes that have been shown to contain many null values in single cell RNA Seq data (Marinov et al., 2014), and also avoids unwarranted assumptions about total RNA content and complexity between samples that has been shown to inflate Type 1 errors (Katayama et al., 2013). Normalized values for all transcripts were subsequently rounded to integers. We noted that this approach performed as well or better than normalization by FPKM (which assumes equivalent total RNA content per cell) or ERCC (which assumes equivalent cell lysis and mRNA recovery efficiency per cell) when measuring minimization of expression difference between technical replicates (Figure S1F) and when examining cluster distinction between adult luminal and basal candidates. Following normalization, we evaluated transcriptomic complexity for each sample by determining the number of genes represented by 5 or more counts in each sample following normalization. As we noted that most samples from each developmental stage fell within a complexity distribution unique to that stage, we removed from further analysis outlier samples with complexities lower than 1.5 standard deviations below the mean complexity for all samples from the same developmental stage (Figure S1H). To produce a final filtered expression matrix for use in identifying groups of related cells, we also manually removed the majority of transcripts with alphanumeric gene names indicative of uncharacterized transcript accession numbers and Riken expression tags as well as mitochondrial genes with names beginning "mt-" from further analysis as well as all transcripts that were present (>0 counts) in fewer than five independent samples following normalization.

Sample Graphing and Clustering: Relationships between samples, cell stages and clusters were visualized using t-SNE and diffusion coordinates as implemented in the tsne and Destiny packages for R, respectively. The interactive online rendering of diffusion maps (see Spike Lab website link, http://uofuhealth.utah.edu/huntsman/labs/spike/d3.php) was generated in javascript using the Plotly libraries (Plotly Technologies). Alternative approaches (i.e., clustering by Spearman rank correlation) showed general agreement with our diffusion mapping (Figures 3 and S4C). Plots were generated with 'plot' or 'plot3d' in the rgl library in R, respectively specifying fill transparency or point radius based on scaled count values.

We clustered samples (i.e., single cell transcriptomes) into related cell types using the 1000 genes (Chromium) or 1500 genes (C1) with the highest local variance across the dataset. Local variance for each gene in the filtered matrix was defined as the ratio of the gene-specific squared coefficient of variation (CV^2) of normalized raw counts to the gene neighborhood CV^2 value. The neighborhood CV^2 value was calculated as the geometric mean of CV^2 values for the neighboring 100 genes in a list ordered by mean expression values from high to low (i.e., 50 genes in each direction) (Figure S1J). We then applied non-negative matrix factorization (NMF) with minimum value subtraction and random number seed generation as implemented in R (Chromium data) or the MeV suite (C1 data) on log2 transformed data for these 1000 or 1500 genes to cluster samples. We tested multiple ranks (i.e., numbers of clusters, 2–15) using a cost convergence cutoff of 1.0 for 10 runs of up to 1000 iterations per rank and divergence update rules and cost measurements (Lee and Seung, 1999). Cophenetic correlation coefficients, change in Residual sum of squares and visual inspection of cluster position in DC and t-SNE graphs was used to identify stable clusters that describe putative cell type changes in the data (Brunet et al., 2004).

SCENIC: We used the standard SCENIC pipeline (Aibar et al., 2017) with depth normalized log2 transformed values from Chromium sequencing as input (Table S1). The SCENIC application (version 0.1.7) uses an older mm9 genomic reference and did not account for 6,516 genes. Per SCENIC recommendations, we refiltered genes to exclude those absent (counts = 0) in > 99% of cells. SCENIC identified 10206 genes that were pruned to 324 regulons based on *cis* regulatory motif analysis and a threshold of at least 10 co-regulated genes per regulon. Cell activity scores for these regulons were layed onto the pseudotemporal cell ordering from diffusion maps and were clustered based on covariance.

ATAC-Seq Analysis: ATAC-seq analysis was performed as described in Dravis et al. (2018). In brief, after quality check with FastQC, sequencing reads were mapped to the mouse genome (mm9) with Bowtie (Langmead et al., 2009). Low quality and duplicated reads were removed and peak calling was done with MACS2. The signal correlations between the biological replicates were checked to ensure reproducibility (Pearson $r > 0.9$) before the replicates were merged with samtools (Li et al., 2009). To normalize the ATAC-seq signal between different samples, the genome-wide signal was binned into 100 bp and quantile normalized with preprocessCore in R. The average signal profile at all genes was then checked for each sample to ensure similar signal-to-noise level. Bedgraph files generated were converted into BigWig format and visualized on UCSC genome browser (https://genome.ucsc.edu/). Signal profiling, correlation analysis and clustering were performed using deepTools (Ramirez et al., 2014). Functional annotation of peaks and peak-gene association were done with GREAT using the default “basal plus extension” parameter (McLean et al 2010). To isolate UARs and URRs, pairwise differential peaks ($FC > 2$ and $FDR < 1 \times 10^{-30}$) between each cell type were first determined using SICER-df (Zang et al 2009), and enrichment score (ES) for each peak calculated as $ES = FC \times -\log(FDR)$. Cell type specific regions were then isolated by cross comparison of peaks using bedTools intersect (Quinlan and Hall, 2010). Afterward, total enrichment score (TES) was calculated by adding up cell type specific ES. For example, the TES of fMaSC = $ES_{fMasc\text{-}Ba}$ + $ES_{mabsc\text{-}Lp}$ + $ES_{fMaSC\text{-}ML}$ Thus, cell-type specificity of UARs and URRs can be ranked by their TES.

Identification of Differentially Expressed Genes: Gene expression differences between NMF designated cell types (i.e., NMF clusters @ rank = 7(Chromium) or 6(C1), and subdivisions of adult cells and clusters 5 and 6 @ rank = 2 and 3 (C1)) were identified using two-class unpaired rank products analysis as implemented in the TM4 suite MeV software (Saeed et al., 2003) with 100 permutations per comparison and the proportion of false significant genes controlled so as not to exceed 0.05. Gene expression values for the 13355 genes in the filtered expression data matrix were compared between each NMF designated group of cells and each other group of NMF designated cells to generate differentially expressed gene lists for each pairwise group-to-group comparison. Cell type specific gene signatures for the 6 NMF-derived clusters in C1 data were determined from the overlap of more highly expressed genes in their five pairwise rank products analyses.

GO Enrichment: Cell type specific gene signatures and other gene lists evaluated in the manuscript were assessed for their content related to specific biological functions using curated GO databases. GO enrichment was assessed and graphed using the BiNGO plugin for Cytsoscape 3.3.0 (Shannon et al., 2003; Maere et al.. 2005). We assessed the overlap of signature gene lists with biological process ontologies using a hypergeometric test statistic and Benjamini & Hochberg False Discovery Rate correction (FDR significance level = 0.05). Enrichments were corroborated with PANTHER (pantherdb.org) and DAVID (https://david.ncifcrf.gov).

Comparative Transcriptomics: Human tumor data was collected as previously described with classification based on PAM50, consensus by a pathology committee, or mode of sample acquisition (Parker et. al., 2009; Cancer Genome Atlas, 2012; Varley et al., 2014; Cirielloet al., 2015; Forero-Torres et al., 2015; Foreroetal., 2016). Differential signature expression was determined by ANOVA or a Kolmogorov-Smirnovtest, as indicated, on the median centered or normalized probe and transcript expression values with normalization as in Pfefferle et al. (2013), or the sum of scaled FPKM values.

Other Graphing: Additional graphics generation used R/Quartz (including plot, rgl and plot3d libraries), Excel, Powerpoint and Adobe Photoshop.

QUANTIFICATION AND STATISTICAL ANALYSIS

Details of the statistical tests used in this manuscript and the number of replicates (n) are presented in the figures and figure legends, are reiterated in the text, and described in detail in the methods sections above. They are also summarized below:

1. Non-negative matrix factorization, as implemented in the TM4-MeV Suite, employed 10 runs of up to 1000 runs each with a cost convergence cutoff set to 1.0 with a check frequency of 40. Update rules and cost measurements were based on divergence. Cophenetic correlation coefficients were based on

Euclidean distance. Chromium data was processed using every fifth sample beginning with sample 1 or 2 or 3 or 4 and resulting statistical metrics were averaged for each rank across these runs to select a suitable NMF rank (i.e., cell cluster number) for further modeling.

2. Rank products, as implemented TM4-MeV suite, used a two class unpaired test statistic with 100 permutations and false discovery control set on the proportion of false significant genes not exceeding 0.05.

3. The correlation in expression of gene signatures across samples were calculated in Excel using the Pearson correlation.

4. GO enrichment, as implemented in the BinGO app for Cytoscape, used “Biological Process” definitions and a hypergeometric test and Benjamini-Hochberg false discovery rate correction at 0.05.

5. Differentially represented ATAC-Seq peaks were called using SICER-df with FC >2 and FDR < 1×10^{-30} and associated enrichment scores were calculated as ES = FC × -log(FDR).

6. Quantification of organoid cultures was conducted manually on 5 replicate wells per treatment. Replicates and significant differences were determined using a two-tailed Student’s t test.

7. Extreme Limiting Dilution Analysis for transplants as implemented at the URL: http://bioinf.wehi.edu.au/software/elda/, were calculated to include 95th percentile confidence intervals and likelihood ratio tests to determine differential stem cell content.

8. Differences in signature expression among TCGA archived breast cancers that were organized by intrinsic subtype was determined by ANOVA with p values indicated in the corresponding figures.

9. Differences in signature expression in primary TNBC versus TNBC metastases was determined with a Kolmogorov-Smirnoff test with p values indicated in the corresponding figures.

DATA AND SOFTWARE AVAILABILITY

All software is commercially available, cited to previous publications or is included in the accompanying “ScriptsandCommandline” text document.

Single cell RNA-sequencing data files are available at the NCBI Sequence Read Archive (SRA) (https://wwvii.ncbi.nlm.nih.gov/sra/) under BioSample accession SAMN07138894 and the NCBI Gene Expression Omnibus (GEO) under accession GSE111113 (Bioproject: PRJNA435951；SRA:RP133477).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank C. Benner, J. Gertz, S. Heinz, Z. He, J. Rutter, M. Shokhirov, and P. Venepally for critical advice; A. Collazo, C. Conley, A. Domissy, M.-C. Ku, G. Luna, and S. Naik for technical assistance; and D. Ayer and B. Cairns for critical reading of the manuscript. This work was supported by the Breast Cancer Research Foundation, Susan G. Komen Foundation (SAC110036), Helmsley Charitable Trust (2012-PG-MED002), Chapman Foundation, and NIH/NCI (R35 CA197687) (to G.M.W.), funds from the Huntsman Cancer Institute (to B.T.S.), and Cancer Center Support Grants P30 CA014195 (Salk) and P30 CA42014 (HCI).

REFERENCES

AIbar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086. [PubMed: 28991892]

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data (Bab rah am Bioinformatics).

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]

Blanpain C, and Fuchs E (2014). Stem cell plasticity. Plasticity of epithelial stem cells In tissue regeneration. Science 344, 1242281. [PubMed: 24926024]

Bolstad B (2018). preprocessCore: a collection of pre-processing functions. R package version 1.42.0.

Breitling R, Armengaud P, Amtmann A, and Herzyk P (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 573, 83–92. [PubMed: 15327980]

Brunet JP, Tamayo P, Golub TR, and Mesirov JP (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA 101, 4164–4169. [PubMed: 15016911]

Buenrostro J, Wu B, Chang H, and Greenleaf W (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol 109, 21.29.1–21.29.9.

Cancer Genome Atlas, N., and Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. [PubMed: 23000897]

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Past ore A, Zhang H, McLellan M, Yau C, Kandoth C, et al.; TCGA Research Network (2015). Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163, 506–519. [PubMed: 26451490]

Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, and Zucker SW (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. USA 102, 7426–7431. [PubMed: 15899970]

Coloff JL, Murphy JP, Braun CR, Harris IS, Shelton LM, Kami K, Gygi SP, Selfors LM, and Brugge JS (2016). Differential glutamate metabolism in proliferating and quiescent mammary epithelial cells. Cell Metab. 23, 867–880. [PubMed: 27133130]

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al.; French StatOmique Consortium (2013). A comprehensive evaluation of normalization methods for lllumina high-throughput RNA sequencing data analysis. Brief. Bioinform 14, 671–683. [PubMed: 22988256]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Dravis C, Spike BT, Harrell JC, Johns C, Trejo CL, Southard-Smith EM, Perou CM, and Wahl GM (2015). Sox10 regulates stem/progenitor and mesenchymal cell states in mammary epithelial cells. Cell Rep. 12, 2035–2048. [PubMed: 26365194]

Dravis C, Chung C-Y, 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, Published online 4 7, 2018. 10.2139/ssrn. 3155661.

Elliott BE, Maxwell L, Arnold M, Wei WZ, and Miller FR (1988). Expression of epithelial-like markers and class I major histocompatibility antigens by a murine carcinoma growing in the mammary gland and in metastases: orthotopic site effects. Cancer Res. 48, 7237–7245. [PubMed: 3191495]

Flores A, Schell J, Krall AS, Jelinek D, Miranda M, Grigorian M, Braas D, White AC, Zhou JL, Graham NA, et al. (2017). Lactate dehydrogenase activity drives hair follicle stem cell activation. Nat. Cell Biol 19, 1017–1026. [PubMed: 28812580]

Forero A, Li Y, Chen D, Grizzle WE, Updike KL, Merz ND, Downs-Kelly E, Burwell TC, Vaklavas C, Buchsbaum DJ, et al. (2016). Expression of the MHC class II pathway in triple-negative breast cancer tumor cells is associated with a good prognosis and infiltrating lymphocytes. Cancer Immunol. Res 4, 390–399. [PubMed: 26980599]

Forero-Torres A, Varley KE, Abramson VG, Li Y, Vaklavas C, Lin NU, Liu MC, Rugo HS, Nanda R, Storniolo AM, et al.; Translational Breast Cancer Research Consortium (TBCRC) (2015). TBCRC 019: a phase II trial of nanoparticle albumin-bound paclitaxel with or without the anti-death receptor 5 monoclonal antibody tigatuzumab in patients with triple-negative breast cancer. Clin. Cancer Res 21, 2722–2729. [PubMed: 25779953]

Ge Y, Gomez NC, Adam RC, Nikolova M, Yang H, Verma A, Lu CPJ, Polak L, Yuan S, Elemento O, and Fuchs E (2017). Stem cell lineage infidelity drives wound repair and cancer. Cell 169, 636–650.e14. [PubMed: 28434617]

Giraddi RR, Shehata M, Gallardo M, Blasco MA, Simons BD, and Stingl J (2015). Stem and progenitor cell division kinetics during postnatal mouse mammary gland development. Nat. Commun 6, 8487. [PubMed: 26511661]

Gray LR, Tompkins SC, and Taylor EB (2014). Regulation of pyruvate metabolism and human disease. Cell. Mol. Life Sci 71, 2577–2604. [PubMed: 24363178]

Haghverdi L, Buettner F, and Theis FJ (2015). Diffusion maps for highdimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989–2998. [PubMed: 26002886]

Hastie TSW (1989). Principal curves. J. Am. Stat. Assoc 84, 502–516.

Hough SR, Laslett AL, Grimmond SB, Kolle G, and Pera MF (2009). A continuum of cell states spans pluripotency and lineage commitment in human embryonic stem cells. PLoS ONE 4, e7708. [PubMed: 19890402]

Hu Y, and Smyth GK (2009). ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. J. Immunol. Methods 347, 70–78. [PubMed: 19567251]

Katayama S, Töhönen V, Linnarsson S, and Kere J (2013). SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics 29, 2943–2945. [PubMed: 23995393]

Kumar P, Tan Y, and Cahan P (2017). Understanding development and stem cells using single cell-based analyses of gene expression. Development 144, 17–32. [PubMed: 28049689]

Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. [PubMed: 19261174]

Lee DD, and Seung HS (1999). Learning the parts of objects by non-nega-tive matrix factorization. Nature 401, 788–791. [PubMed: 10548103]

Li B, and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12, 323. [PubMed: 21816040]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 20782079.

Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, et al.; kConFab (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat. Med 15, 907–913. [PubMed: 19648928]

Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YL, Sultana H, Oliver B, and Harbison ST (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila mel-anogaster. BMC Genomics 17, 28. [PubMed: 26732976]

Loh KM, and Lim B (2011). A precarious balance: pluripotency factors as lineage specifiers. Cell Stem Cell 8, 363–369. [PubMed: 21474100]

Maere S, Heymans K, and Kuiper M (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21, 3448–3449. [PubMed: 15972284]

Makarem M, Kannan N, Nguyen LV, Knapp DJ, Balani S, Prater MD, Stingl J, Raouf A, Nemirovsky O, Eirew P, and Eaves CJ (2013a). Developmental changes in the in vitro activated regenerative activity of primitive mammary epithelial cells. PLoS Biol. 11, e1001630. [PubMed: 23966837]

Makarem M, Spike BT, Dravis C, Kannan N, Wahl GM, and Eaves CJ (2013b). Stem cells and the developing mammary gland. J. Mammary Gland Biol. Neoplasia 18, 209–219. [PubMed: 23624881]

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, and Wold BJ (2014). From single-cell to cell-pool transcriptomes: sto-chasticity in gene expression and RNA splicing. Genome Res. 24, 496–510. [PubMed: 24299736]

Mazurek S, Boschek CB, Hugo F, and Eigenbrodt E (2005). Pyruvate kinase type M2 and its role in tumor growth and spreading. Semin. Cancer Biol 15, 300–308. [PubMed: 15908230]

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, and Bejerano G (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol 28, 495–501. [PubMed: 20436461]

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]

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol 27, 1160–1167. [PubMed: 19204204]

Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, et al. (2013). Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol. 14, R125. [PubMed: 24220145]

Pfefferle AD, Spike BT, Wahl GM, and Perou CM (2015). Luminal progenitor and fetal mammary stem cell expression features predict breast tumor response to neoadjuvant chemotherapy. Breast Cancer Res. Treat 149, 425–437. [PubMed: 25575446]

Prat A, Parker JS, Kargin ova O, Fan C, Livasy C, Herschkowitz JI, He X, and Perou CM (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 12, R68. [PubMed: 20813035]

Prater MD, Petit V, Alasdair Russell I, Giraddi RR, Shehata M, Menon S, Schulte R, Kalajzic I, Rath N, Olson MF, et al. (2014). Mammary stem cells have myoepithelial cell properties. Nat. Cell Biol. 16, 942–950, 1–7. [PubMed: 25173976]

Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. [PubMed: 20110278]

Ramirez F, Dundar F, Diehl S, Gruning BA, and Manke T (2014). deep-Tools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 42, W187–191. [PubMed: 24799436]

Rios AC, Fu NY, Lindeman GJ, and Visvader JE (2014). In situ identification of bipotent stem cells in the mammary gland. Nature 506, 322–327. [PubMed: 24463516]

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, and Mesirov JP (2011). Integrative genomics viewer. Nat. Bio-technol 29, 24–26.

Rodilla V, Dasti A, Huyghe M, Lafkas D, Laurent C, Reyal F, and Fre S (2015). Luminal progenitors restrict their lineage potential during mammary gland development. PLoS Biol. 13, e1002069. [PubMed: 25688859]

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, et al. (2003). TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–378. [PubMed: 12613259]

Schell JC, Wisidagama DR, Bensard C, Zhao H, Wei P, Tanner J, Flores A, Mohlman J, Sorensen LK, Earl CS, et al. (2017). Control of intestinal stem cell function and proliferation by mitochondrial pyruvate metabolism. Nat. Cell Biol 19, 1027–1036. [PubMed: 28812582]

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]

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504. [PubMed: 14597658]

Shao C, and Höfer T (2017). Robust classification of single-cell transcrip-tome data by non negative matrix factorization. Bioinformatics 33, 235–242. [PubMed: 27663498]

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. 14, R134. [PubMed: 23088371]

Shyh-Chang N, Daley GQ, and Cantley LC (2013). Stem cell metabolism in tissue development and aging. Development 140, 2535–2547. [PubMed: 23715547]

Skibinski A, Breindel JL, Prat A, Galván P, Smith E, Rolfs A, Gupta PB, LaBaer J, and Kuperwasser C (2014). The Hippo transducer TAZ interacts with the SWI/SNF complex to regulate breast epithelial lineage commitment. Cell Rep. 6, 1059–1072. [PubMed: 24613358]

Sleeman KE, Kendrick H, Ashworth A, Isacke CM, and Smalley MJ (2006). CD24 staining of mouse mammary gland cells defines luminal epithelial, myoepithelial/basal and non-epithelial cells. Breast Cancer Res. 8, R7. [PubMed: 16417656]

Spike BT (2016). Breast cancer stem cells and the move toward high resolution stem cell systems. In Cancer Stem Cells, Liu H and Lathia J, eds. (Elsevier), pp. 121–148.

Spike BT, Engle DD, Lin JC, Cheung SK, La J, and Wahl GM (2012). A mammary stem cell population identified and characterized in late embryo-genesis reveals similarities to human breast cancer. Cell Stem Cell 10, 183–197. [PubMed: 22305568]

Spike BT, Kelber JA, Booker E, Kalathur M, Rodewald R, Lipianskaya J, La J, He M, Wright T, Klemke R, et al. (2014). CRIPTO/GRP78 signaling maintains fetal and adult mammary stem cells ex vivo. Stem Cell Reports 2, 427–439. [PubMed: 24749068]

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]

Sun P, Yuan Y, Li A, Li B, and Dai X (2010). Cytokeratin expression during mouse embryonic and early postnatal mammary gland development. His-tochem. Cell Biol. 133, 213–221.

Trejo C, Luna G, Dravis C, Spike B, and Wahl G (2017). Lgr5 is a marker for fetal mammary stem cells, but is not essential for stem cell activity or tumorigenesis. NPJ Breast Cancer 3, 16. [PubMed: 28649656]

van der Maaten L, and Hinton G (2008). Visualizing data using t-SNE. J. Mach. Learn. Res 9, 2579–2605.

Van Keymeulen A, Rocha AS, Ousset M, Beck B, Bouvencourt G, Rock J, Sharma N, Dekoninck S, and Blanpain C (2011). Distinct stem cells contribute to mammary gland development and maintenance. Nature 479, 189–193. [PubMed: 21983963]

Varley KE, Gertz J, Roberts BS, Davis NS, Bowling KM, Kirby MK, Nesmith AS, Oliver PG, Grizzle WE, Forero A, et al. (2014). Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res. Treat 146, 287–297. [PubMed: 24929677]

Veltmaat JM, Mailleux AA, Thiery JP, and Bellusci S (2003). Mouse embryonic mammogenesis as a model for the molecular regulation of pattern formation. Differentiation 71, 1–17. [PubMed: 12558599]

Viliadsen R, Fridriksdottir AJ, Rønnov-Jessen L, Gudjonsson T, Rank F, LaBarge MA, Bissell MJ, and Petersen OW (2007). Evidence for a stem cell hierarchy in the adult human breast. J. Cell Biol. 777, 87–101.

Visvader JE, and Stingl J (2014). Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes Dev. 28, 1143–1158. [PubMed: 24888586]

Wahl GM, and Spike BT (2017). Cell state plasticity, stem cells, EMT, and the generation of intra-tumoral heterogeneity. NPJ Breast Cancer 3, 14. [PubMed: 28649654]

Wang D, Cai C, Dong X, Yu QC, Zhang XO, Yang L, and Zeng YA (2015). Identification of multipotent mammary stem cells by protein C receptor expression. Nature 517, 81–84. [PubMed: 25327250]

Wang C, Christin JR, Oktay MH, and Guo W (2017). Lineage-biased stem cells maintain estrogen-receptor-positive and -negative mouse mammary luminal lineages. Cell Rep. 18, 2825–2835. [PubMed: 28329676]

Wuidart A, Ousset M, Rulands S, Simons BD, Van Keymeulen A, and Blanpain C (2016). Quantitative lineage tracing strategies to resolve multipotency in tissue-specific stem cells. Genes Dev. 30, 1261–1277. [PubMed: 27284162]

Wuidart A, Sifrim A, Fioramonti M, Matsumura S, Brisebarre A, Brown D, Centonze A, Dannau A, Dubois C, Van Keymeulen A, et al. (2018). Early lineage segregation of multipotent embryonic mammary gland progenitors. Nat. Cell Biol 20, 666–676. [PubMed: 29784918]

Zang C, Schones DE, Zeng C, Cui K, Zhao K, and Peng W (2009). A clustering approach for identification of enriched domains from histone modification ChlP-seq data. Bioinformatics 25, 1952–1958. [PubMed: 19505939]

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. (2008). Model-based analysis of ChlP-seq (MACS). Genome Biol. 9, R137. [PubMed: 18798982]

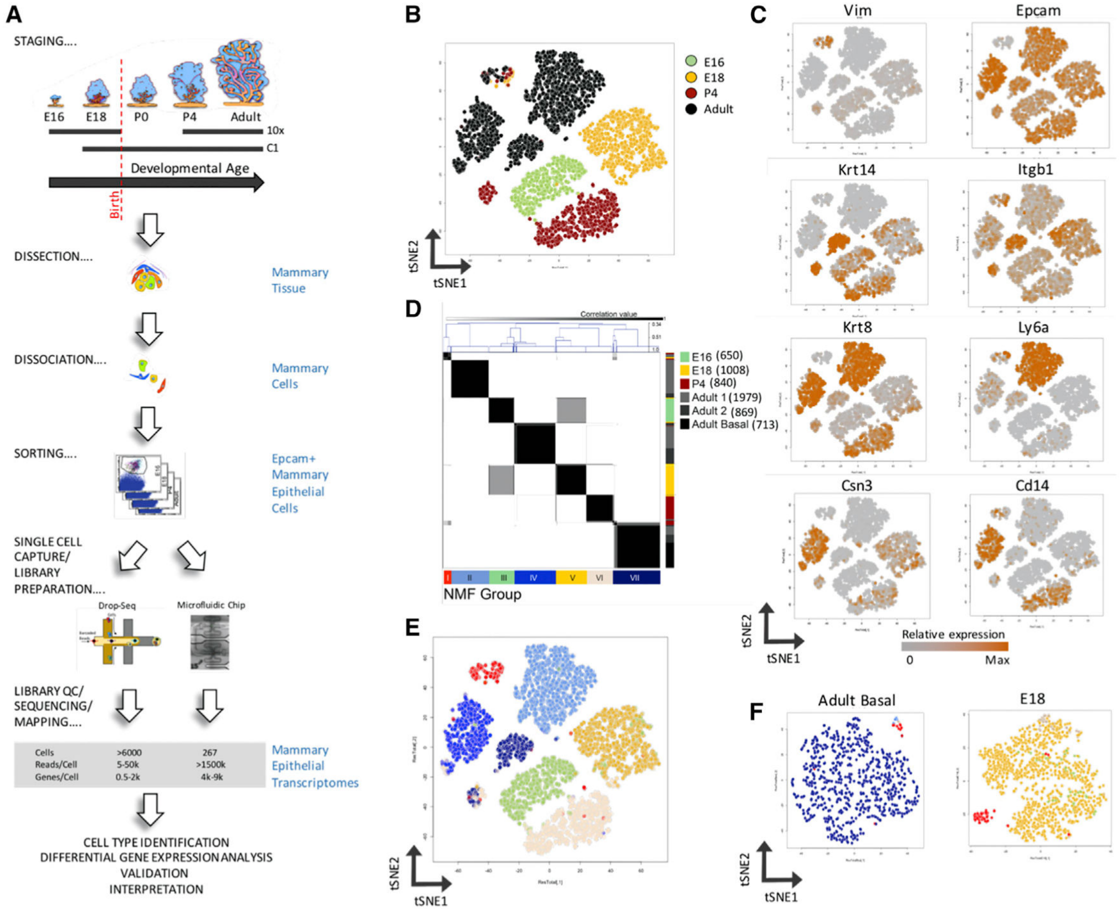


Figure 1. Derivation and Clustering of Mouse Mammary Epithelial Single-Cell Transcriptomes from Embryonic Development to Adulthood

(A) Isolation and sequencing of mammary cells from different developmental stages. Two strategies are shown, Chromium Drop-Seq (10× Genomics) and C1-microfluidic capture (Fluidigm), with differential output (gray box).

(B) tSNE plot of single-cell transcriptomes from indicated developmental stages.

(C) The plot in (B), overlaid with relative expression levels of mammary lineage markers.

(D) NMF clustering of single-cell expression profiles (n = 6,059), shown by the white-to-black correlation scale. Colored bars correspond to developmental context (right; y axis) or NMF group (x axis).

(E) Projection of NMF groups identified in (D) onto the tSNE plot from (B).

(F) tSNE plots for adult basal (Epcam^{+}, Cd49f^{+}, left panel) and E18 cells (right panel) isolated, processed, and plotted separately with colors corresponding to the NMF grouping in (D).

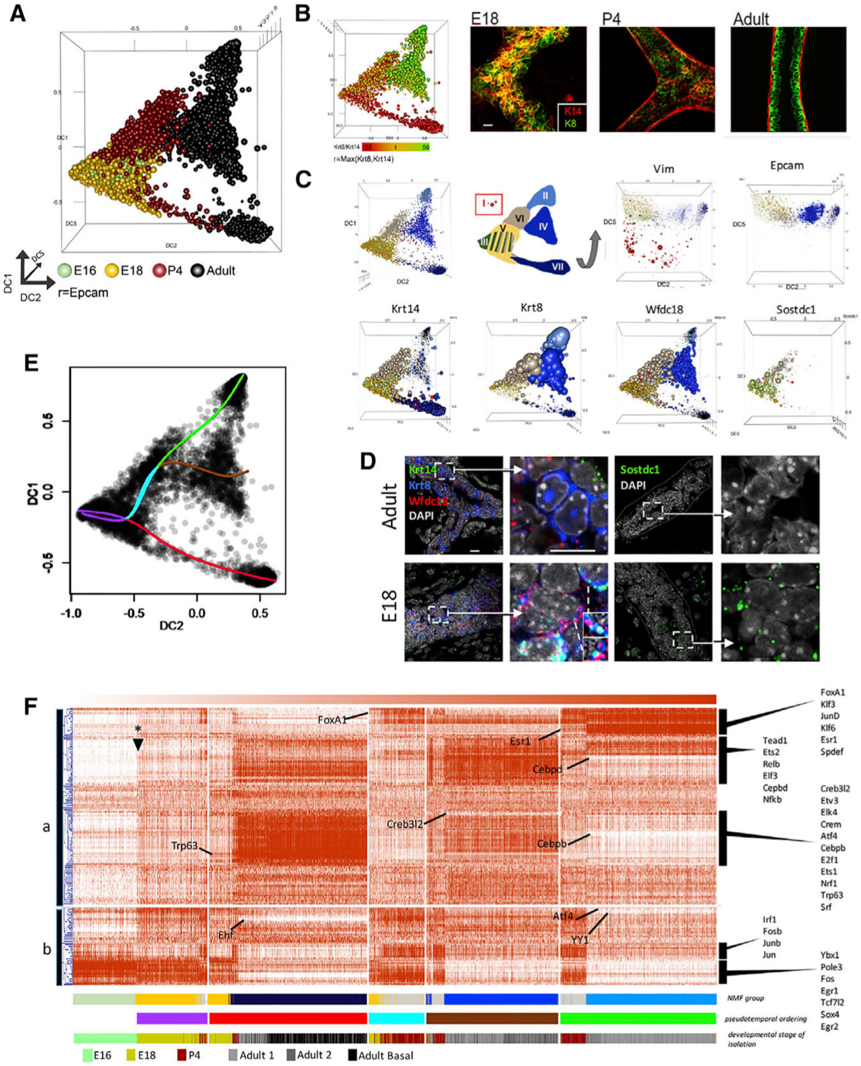


Figure 2. Diversification of Cell Types in the Mouse Mammary Epithelium

(A) Diffusion map of single-cell transcriptomes annotated by stage of collection.

(B) A color gradient is used to depict the ratio of Krt8 and Krt14 on the same diffusion map. Cells with much higher levels of Krt8 levels than Krt14 are green, and cells with the opposite configuration are red. Orange/yellow color indicates balanced coexpression. Sphere radius assigned by maximum value of Krt8 or Krt14. The right three panels represent immuno-staining of mouse mammary epithelium for Krt8 (green) and Krt14 (red) at the given stages of development.

(C) Placement of NMF groups in the diffusion map from (A) and their correspondence to markers of known cell types as indicated. Sphere radius represents normalized expression values for the given factors. Color, NMF group.

(D) Multiplex fluorescent *in situ* hybridization for select lineage markers shown in (C), in adult and embryonic mammary tissue. Insets in second lower panel, Extreme digital zoom showing close-proximity red, green, and blue signals.

(E) Pseudotemporal vectors through the diffusion map from (A) color-coded to represent lineage branch points as shown in (F).

(F)Heatmap of regulon scores from SCENIC analysis. Rows, Individual regulons. Columns, Cells organized according to pseudotemporal trajectories as indicated below the heatmap. a, Adult lineage oriented regulons; b, primitive regulons; *onset of balanced lineage regulon activity. Color coding is also given for NMF groups and developmental stage as in Figure 1D.

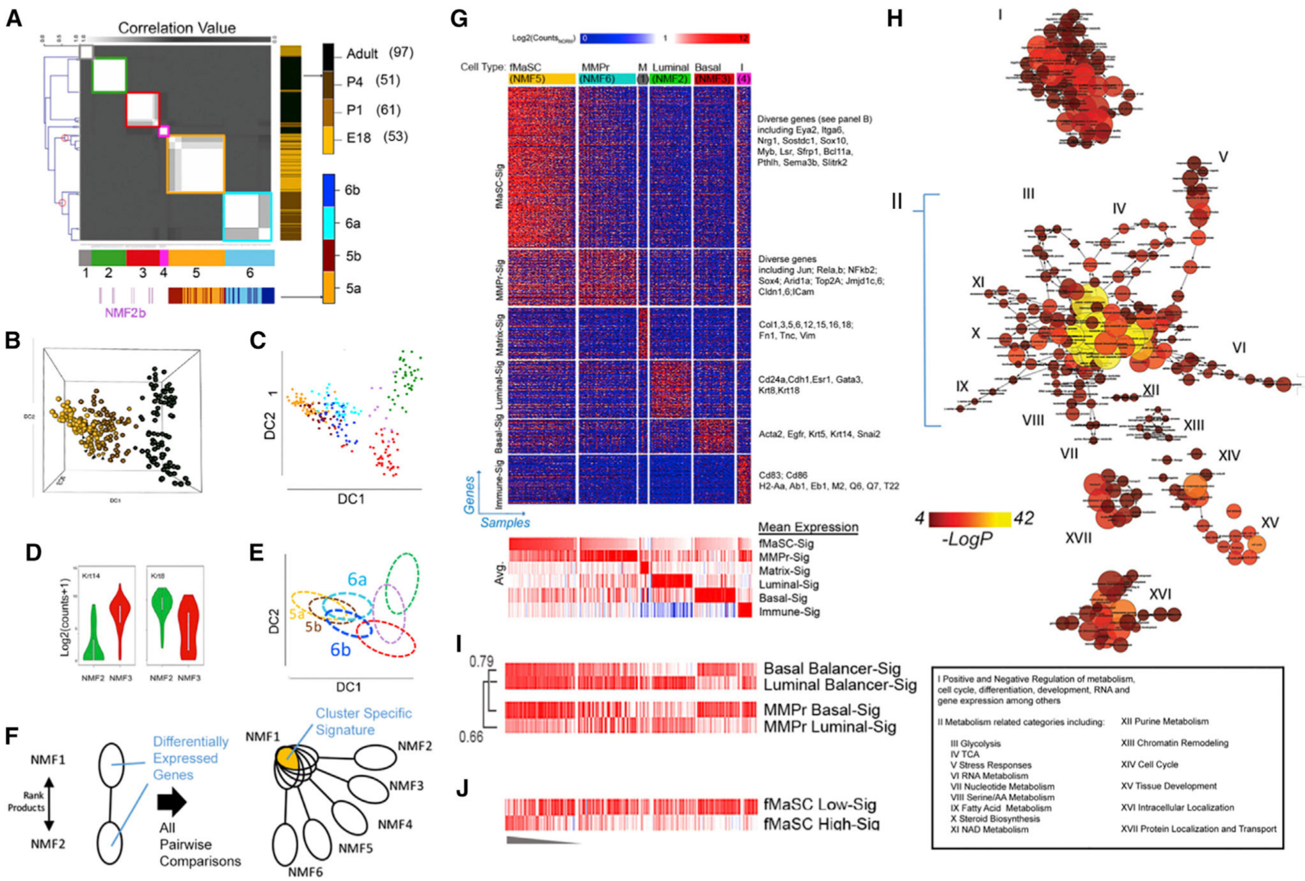


Figure 3. Identification of Cell Types and Signatures across Early Mammary Epithelial Development

(A) NMF correlation matrix of single cells (clusters/rank = 6) with NMF sample groups 1–6 numbered and color-coded. Also shown and color-coded are the developmental stages of isolation for each cell, and subclusters (a and b) derived for adult cells and NMF groups 5 and 6 independently of the correlation matrix shown.

(B) Relationships between single-cell profiles graphed according to the first three diffusion components, and color-coded by stage of origin, $\sigma = 129$.

(C) DC1 and two positions for the epithelial NMF groups, $\sigma = 55$.

(D)Violin plots showing expression levels of luminal-associated Krt8 and basal-associated Krt14 among cells of NMF groups 2 and 3.

(E) A simplified schematic of the approximate positions of NMF subclusters in the two-dimensional (2D) diffusion map from (C).

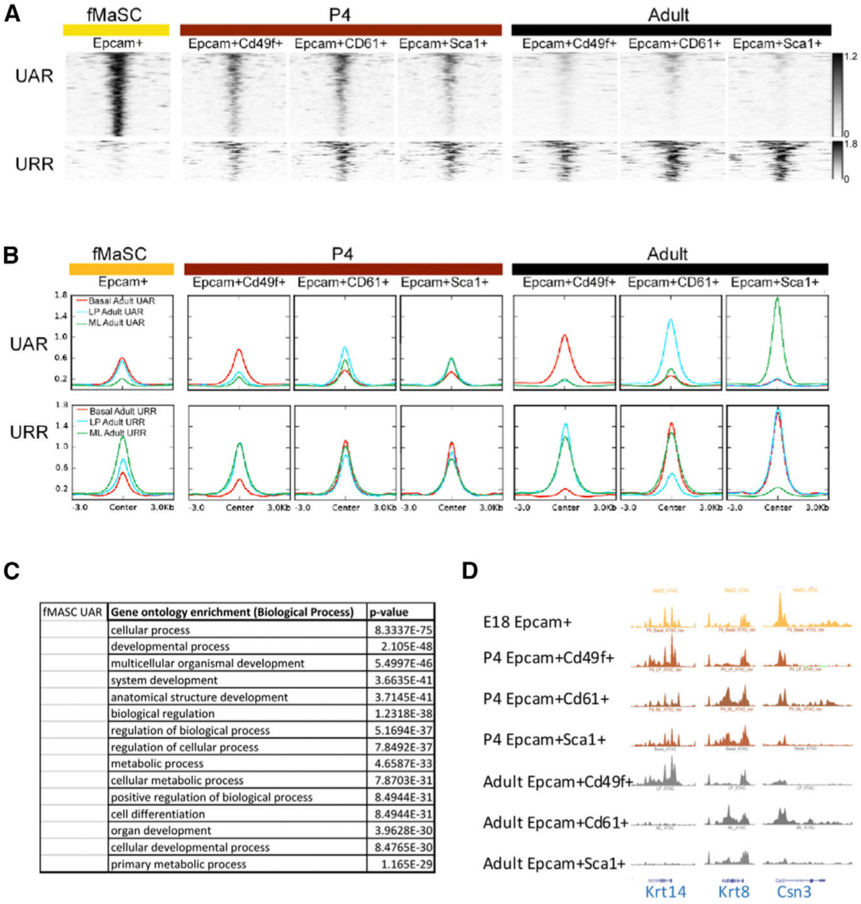
(F) A schematic describing the rank product-based procedure for defining group-specific signatures.

(G) Heatmap of signature expression with cells sorted by NMF group with annotation of select genes. Rows, Gene expression values. Columns, Samples (i.e., single cells). Mean signature enrichment per cell is also given (bottom).

(H) Graphical representation of GO categories for the fMASC signature.

(I) Mean expression in individual cells of a luminal balancer signature and a basal balancer signature. Also shown is the mean expression of differentially expressed gene lists delineating the prominent subdivision of MMPr cells. Brackets, Pearson correlation.

(J) Mean expression of the differentially expressed gene lists for NMF5a versus NMF5b (i.e., fMaSC subdivision). Ordering of cells from (G) is maintained in (I) and (J).



fMASC UAR	Gene ontology enrichment (Biological Process)	p-value
	cellular process	8.3337E-75
	developmental process	2.105E-48
	multicellular organismal development	5.4997E-46
	system development	3.6635E-41
	anatomical structure development	3.7145E-41
	biological regulation	1.2318E-38
	regulation of biological process	5.1694E-37
	regulation of cellular process	7.8492E-37
	metabolic process	4.6587E-33
	cellular metabolic process	7.8703E-31
	positive regulation of biological process	8.4944E-31
	cell differentiation	8.4944E-31
	organ development	3.9628E-30
	cellular developmental process	8.4765E-30
	primary metabolic process	1.165E-29

Figure 4. Differential Chromatin Regulation between Primitive and Mature Mammary Epithelial Cells

(A) Tornado plots of uniquely accessible regions (UARs) (n = 1,222) and uniquely repressed regions (URRs) (n = 242) for fMaSCs relative to differentiated mammary epithelia as indicated. Loci, represented by the y axes, are held consistent between stages, while their intensity representing ATAC-seq signal is differential.

(B) Averaged accessibility for UARs and URRs derived from adult mammary epithelial lineage comparisons.

(C) The top 15 GO categories for fMaSC UARs with Bonferroni-corrected p values.

(D) Proximal chromatin accessibility at three lineage-associated loci compared across indicated differentiation states.

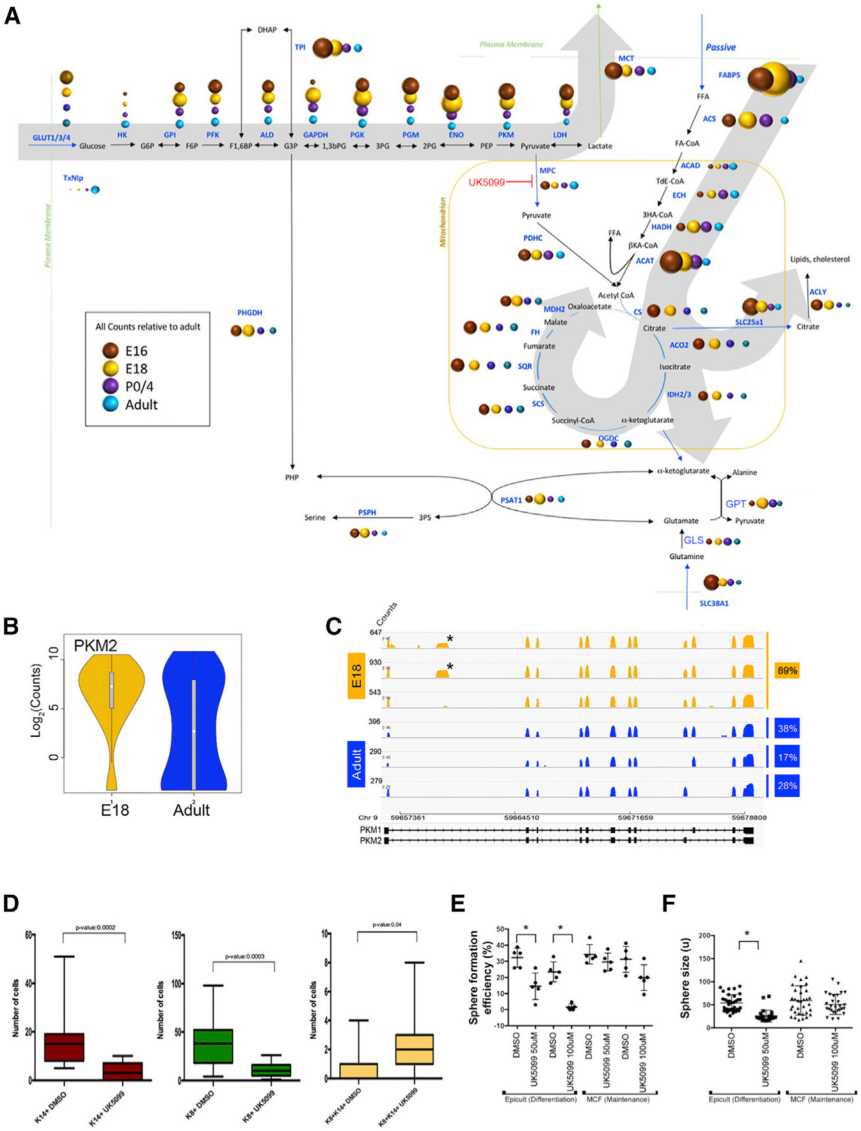


Figure 5. Metabolism-Related Gene Expression Profile in fMaSC

(A) Mean expression levels (sphere size) of genes encoding glycolytic, Krebs cycle, and fatty acid metabolism enzymes across four developmental stages. Gray arrows, Modeled metabolic flux.

(B) Expression levels of the Pkm2 in fetal and adult mammary epithelial cells.

(C) Aligned reads matching Pkm1 and −2 in three representative adult and fetal epithelial cells with the percentage of similar cells at each stage. *Occasional reads are seen in introns and non-coding regions across the genome adjacent to encoded poly-thymine tracts, but these are not enumerated in the quantification of transcripts.

(D) Enumeration of Krt8 and Krt14 single-positive and double-positive cells in organoids.

(E and F) The number (E) and size (F) of organoid structures produced under treatment with the mitochondrial pyruvate transport inhibitor, UK5099 (n = 5). *p < 0.1, Student’s t test.

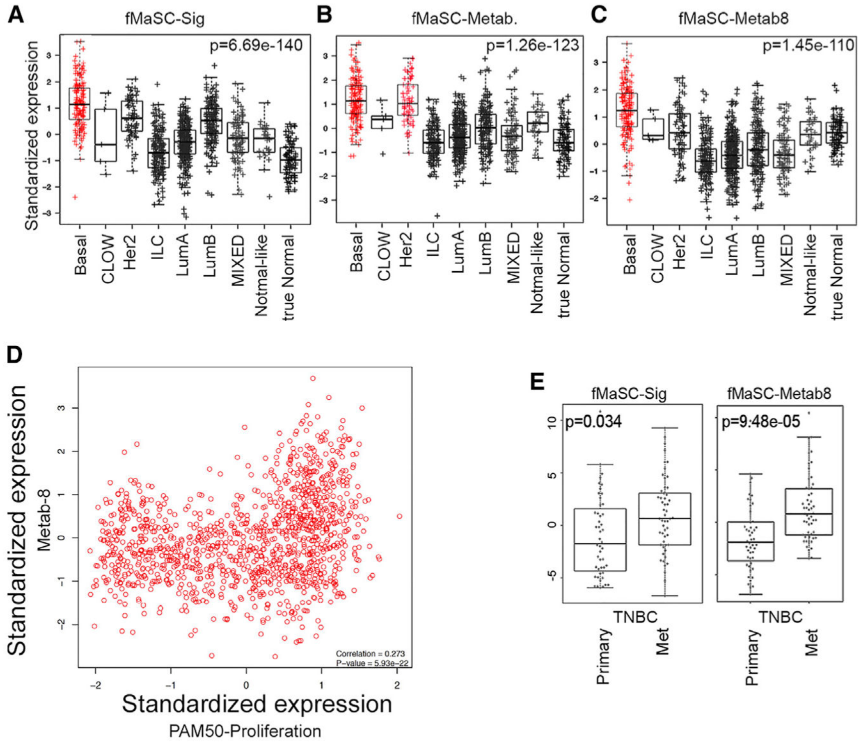


Figure 6. Human Breast Cancers Exhibit fMaSC-Related Biology

(A-C) Expression levels of single-cell-derived fMaSC signatures in human breast cancers of varying molecular subtype. (A) fMaSC-Signature. (B) fMaSC-metabolism Signature. (C) Reduced 8 gene fMaSC-metabolism signature. p value, ANOVA between groups.

(D) TCGA breast cancer expression data plotted for median expression of a proliferations signature (PAM50, × axis) versus median expression of the Metab-8 fMaSC subsignature.

(E) fMASC-signature genes are overexpressed in metastatic triple-negative breast cancer. p value, Kolmogorov-Smirnov.

KEY RESOURCES TABLE

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Antibodies		
Alexa Fluor 647 anti-CD326 (Epcam)	Biolegend	Cat# 118212; RRID:AB_1134101
FITC anti-Cd49f (ITGA6)	Stem Cell Technologies	Cat# 60037FI.1; RRID:AB_2734790
Anti-Krt14	Biolegend	Cat# 905304; RRID:AB_2616896
Anti-Krt8(TROMA-1)	DSHB Univ. of Iowa	Cat# AB-531826; RRID:AB_531826
Anti-H3K27me3	Millipore	Cat# 07–449; RRID:AB_310624
FITC-anti rat lgG2	Thermo Fisher	Cat# PA1–84761; RRID:AB_933936
PE-anti rabbit IgG	Santa Cruz Biotechnology	Cat# SC-3739; RRID:AB_649004
Anti-BrdU	Bio-Rad	Cat# MCA2060GA; RRID:AB_10545551
Chemicals, Peptides, and Recombinant Proteins		
UK5099	Sigma Aldrich	Cat# PZ0160
bFGF	Stem Cell Technologies	Cat# 78003
EGF	Sigma Aldrich	Cat# E4127
DAPI	Thermo Scientific	Cat# 62248
Urea	Sigma	Cat# U5378
N.N.N’N’-tetrakis (2-hydroxypopryl) ethylenediamine	Sigma	Cat# 122262
Polyethylene glycol mono-p-isooctylphenyl ether/Triton X-100	Sigma	Cat# 93443
2,2,2’-nitrilotriethanol	Sigma	Cat# 90279
Critical Commercial Assays		
Chromium prep	10x Genomics	Cat# 120237
Nextera DNA library kit	Illumina	Cat# FC-121–1030
TapeStation DNA high sensitivity kit (D1000)	Agilent	Cat# 5067–5585
NEBNext High Fidelity 2x PCR mix	NEB	Cat# M0541
RNAscope Multiplex Fluorescent v2	ACD Bio, Newark CA	Cat# 323110
SMARTer Ultra Low RNA kit	Clontech	Cat# 634936
SMART-Seq® v4 Ultra® Low Input RNA Kit for Sequencing	Clontech	Cat# 634888
Advantage 2 PCR Kit	Clontech	Cat# 634206
C1 Single-Cell auto Prep Reagent Kit for mRNA Seq	Fluidigm	Cat# 1006201
Bioanalyzer RNA Pico Kit	Agilent	Cat# 5067–1513
Quanti-it Pico-green dsDNA assay kit	Thermo Fisher	Cat# P11496
LIVE/DEAD kit	Invitrogen	Cat# MP03224
AM Pure XP beads	Agencourt	Cat# A63880
C1 Single-Cell mRNA Seq IFC, 10–17 μm	Fluidigm	Cat# 100–6041
ERCC RNA spike in mix	Thermo Fisher	Cat# 4456740
Illumina Nextera XT DNA sample preparation kit	Illumina	Cat# FC-131–1096
Illumina Nextera XT DNA sample preparation index kit	Illumina	Cat# FC-131–1002
Deposited Data		
FastQ sequencing files	NCBI sequence read archive and GEO https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111113 and https://trace.ncbi.nlm.nih.gov/Traces/sra/	SAMN07138894; GSE111113
Experimental Models: Organisms/Strains		
C57BL/6	Charles River	Strain Code: 027
CB17/lcr-Prkdcscid/lcrlcoCrl (SCID)	Charles River	Strain Code: 236
Software and Algorithms		
Bowtie	http://bowtie-bio.sourceforge.net/	Langmead et al., 2009
MACS2	http://liulab.dfci.harvard.edu/MACS/	Zhang etal., 2008

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bedtools	http://bedtools.readthedocs.io/en/latest/	Quinlan and Hall, 2010
Samtools	http://www.htslib.org/doc/samtools.html	Li etal., 2009
Deeptools	https://deeptools.readthedocs.io/en/latest/	Ramirez etal., 2014
GREAT	http://great.stanford.edu/public/html/index.php	McLean et al., 2010
SICER	https://home.gwu.edu/~wpeng/Software.htm	Zang etal., 2009
preprocessCore (R)	https://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html	Bolstad, 2018
Plotly	Plotly Technologies; https://plot.ly	N/A
rtsne	https://cran.r-project.org/web/packages/tsne/	van der Maaten and Hinton, 2008
NMF (R)	https://cran.r-project.org/web/packages/NMF/	Lee and Seung, 1999
SCENIC	https://github.com/aertslab/SCENIC	Aibar etal., 2017
R 3.3.0 (for Mac OsX)	https://cran.r-project.org/bin/macosx/old/	R-3.3.0.pkg
Destiny (Diffusion Maps)	https://bioconductor.org/biocLite.R	Haghverdi etal., 2015
TM4-MeV4.8	mev.tm4.org	Saeed etal., 2003
Non-Negative Matrix Factorization	mev.tm4.org	Lee and Seung, 1999
Rank products	mev.tm4.org	Breitling etal., 2004
Cytoscape 3.3.0	www.cytoscape.org	Shannon etal., 2003
BiNGO	http://apps.cytoscape.org/apps/bingo	Maere etal., 2005
FastQC 0.11.2	http://www.bioinformatics.babraham.ac.uk/projects/fastqc	Andrews, 2010
RSEM 1.2.29	https://github.com/deweylab/RSEM/releases	Li and Dewey, 2011
STAR 2.4.2a	https://github.com/alexdobin/STAR	Dobin etal., 2013
IGV2.3.83	https://software.broadinstitute.org/software/igv/	Robinson etal., 2011
ELDA	http://bioinf.wehi.edu.au/software/elda/	Hu and Smyth, 2009
Other		
Epicult-B Basal Medium (mouse)	Stem Cell Technologies	Cat# 05611
Epicult-B Proliferation Supplement (mouse)	Stem Cell Technologies	Cat# 0562
B27 Supplement	GIBCO	Cat# 17504044
Hydrocortisone	Sigma Aldrich	Cat# H4001
Collagen ase/Hyaluronidase	Stem Cell Technologies	Cat# 07912
Dispase	Stem Cell Technologies	Cat# 25300–054
Trypsin	GIBCO by Life Technologies	Cat# 25300–054
Fetal Bovine Serum	Serum Source	Cat# FB22–500
Matrigel (complete)	Corning	Cat# 354234
Matrigel (growth factor reduced)	Corning	Cat# 356231
DMEM-F12 (no phenol red)	Thermo Fisher	Cat# 21041025