Single-cell analysis identifies TCF4 and ID3 as a molecular switch of mammary epithelial stem cell differentiation.

Koon-Kiu Yan^{1*}, Erin Nekritz^{2*}, Bensheng Ju¹, Xinran Dong¹, Rachel Werner², Dayanira Alsina-Beauchamp², Celeste Rosencrance¹, Partha Mukhopadhyay², Qingfei Pan¹, Andrej Gorbatenko², Liang Ding¹, Yanyan Wang³, Chenxi Qian¹, Hao Shi³, Bridget Shaner¹, Sivaraman Natarajan¹, Hongbo Chi³, John Easton¹, Jose Silva^{2†}, Jiyang Yu^{1†},

1. Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
2. Department of Pathology, Icahn School of Medicine at Mount Sinai Hospital, NY, NY 10029, USA
3. Department of Immunology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

*These authors contributed equally to this work

†Correspondence to

-Jiyang Yu Ph.D.
Assistant Member
Department of Computational Biology
jiyang.yu@stjude.org

St. Jude Children's Research Hospital
262 Danny Thomas Pl.,
Memphis, TN 38105

-Jose Silva Ph.D.
Associate Professor
Department of Pathology
Jose.silva@mssm.edu

Icahn School of Medicine at Mount Sinai Hospital,
1 Gustave L. Levi Pl.,
NY, NY 10029

ABSTRACT

It is well known that the expansion of the mammary epithelium during the ovarian cycles in female mammals is supported by the transient increase in mammary epithelial stem cells (MaSCs). However, dissecting the molecular mechanisms that govern MaSC function and differentiation is poorly understood due to the lack of standardized methods for their identification and isolation.

The development of robust single-cell mRNA sequencing () technologies and the computational methods to analyze them provides us with novel tools to approach the challenge of studying MaSCs in a completely unbiased way without. Here, we have performed the largest scRNA-seq analysis of individual mammary epithelial cells (~70,000 cells). Our study identified a distinct cell population presenting molecular features of MaSCs.

Importantly, further purification and additional in-depth single-cell analysis of these cells revealed that they are not a fully homogenous entity. Instead, we identified three subpopulations representing early stages of lineage commitment. By tracking their molecular evolution through single-cell network analysis we found that one of these subpopulations represents bipotent MaSCs from which luminal and basal lineages diverge. Importantly, we also confirmed the presence of these cells in human mammary glands. Finally, through expression and network analysis studies, we have uncovered transcription factors that are activated early during lineage commitment. These data identified E2-2 (Tcf4) and ID3 as a potential molecular switch of mammary epithelial stem cell differentiation.

INTRODUCTION

The mammary gland is a dynamic organ that undergoes extensive tissue remodeling over the female lifespan¹⁻⁴. In the adult female, expansion and regression of the basal and luminal bilayers occur regularly during the ovarian cycles. Additionally, massive epithelial expansion and generation of specialized secretory cells occur during pregnancy, resulting in the generation of specialized secretory glands. In all of these stages, epithelial expansion is supported by a transient increase in the mammary epithelial stem cell (MaSC) population⁵⁻⁷.

Although the existence of MaSCs is well known^{8,9}, two main limitations have handicapped the dissection of the molecular mechanisms that govern their function and differentiation. The first issue is the limited ability to identify and isolate MaSCs. For this, fluorescence-activated cell sorting (FACS) using antibodies against membrane markers followed by transplantation into cleared mammary fat pads has been used to examine MaSC potential. However, these assays represent non-physiological conditions where lineage-committed cells demonstrate facultative stem cell activity that does not occur *in vivo*^{8,10}. Thus, during recent years, the study of MaSCs has turned to lineage tracing experiments. Here, cells of interest are labeled *in vivo* at an initial time point, and the progeny derived from these cells is tracked over time^{5,11-13}. Unfortunately, such experiments have generated conflicting results. Although some studies have reported the existence of luminal and basal unipotent MaSCs¹⁴, others have found bipotent MaSCs that replenish both lineages^{5,11-13}. Importantly, both FACS and lineage tracing experiments rely on the use of predetermined markers or promoters and, consequently, are subject to bias.

The second limitation in studying MaSCs involves the methods used to identify genes or pathways involved in lineage commitment. Generally, molecular features (such as mRNA profiling) of a MaSC-enriched population are compared with those of differentiated luminal and basal cells^{12,14-16}. Importantly, however, this type of comparison is

handicapped by the very nature of fate specification. Molecularly, differentiation is a continuous process in which a few signals start a cascade that leads to profound changes observed between uncommitted cells and the fully differentiated progeny. However, the initiating and subsequent signals do not necessarily accumulate, and instead can appear and disappear without being detected at the time when cells are being processed. Thus, delineating the molecular hierarchy that regulates lineage commitment of the mammary epithelium requires the identification of pure MaSC populations and elucidation of the molecular changes that occur early during the process.

The rapid development of robust single-cell mRNA sequencing (scRNA-seq) technologies and the computational methods to analyze them provides us with novel tools to approach the challenge of studying MaSCs in a completely unbiased way without any selection of predetermined markers or perturbations. Thus, by coupling scRNA-seq with state-of-the-art systems biology^{17,18} we have analyzed over 70,000 individual mammary epithelial cells. Our study clearly identified a distinct population of cells characterized by the expression of the protein C receptor, which has recently been shown to be expressed in multipotent cells in several tissues^{13,19,20}. Importantly, additional in-depth single-cell analysis of these cells identified three subpopulations representing an early stage of lineage commitment. By tracking their molecular evolution through single-cell network analysis we found that one of these subpopulations represents bipotent MaSCs from which luminal and basal lineages diverge. Importantly, we also confirmed the presence of these cells in human mammary glands after performing similar studies. Finally, through expression and network analysis studies, we have uncovered transcription factors that are activated early during lineage commitment. These data identified E2-2 (*Tcf4*) and ID3 as a molecular switch of mammary epithelial stem cell differentiation.

RESULTS

Single-cell studies of the mouse mammary gland identify distinct cell subpopulations.

ScRNA-seq is a powerful method to identify cell subtypes and track the trajectories of cell lineages^{21,22}. We performed scRNA-seq on mammary epithelial cells (MECs) isolated at different stages of the female mouse development: puberty (6 weeks), virgin adult (4 months) and old (24 months) (Fig. 1A). Because of the well-known impact of the ovarian hormones during the estrous cycle^{4,6,7}, we further separated adult females into estrus and diestrus groups. For isolation, non-epithelial cells were removed by antibodies against CD31⁺, CD45⁺, Ter119⁺ and BP1⁺²³⁻²⁵ and EpCAM⁺ epithelial cells were isolated by FACS. In total, we evaluated ~20,000 MECs per time point (>70,000 cells total) using a droplet-based platform (10X Genomics)^{21,26}, which is one of the largest scRNA-seq studies ever performed. Clustering analysis of all sequenced cells based on t-distributed stochastic neighbor embedding (tSNE) plotting^{21,22} identified three of the major clusters as basal, luminal estrogen receptor-positive (ER⁺) and luminal estrogen receptor-negative (ER⁻) cells (Fig. 1A-1C). Remarkably, an additional small population also emerged (Fig. 1A). This population did not present high levels of basal or luminal markers but instead showed the distinct expression of the protein C receptor (PROCR) (Fig. 1B-C) which has been linked to multipotent stem cells in endothelium¹⁹, keratinocytes²⁰, and mammary epithelial cells¹³. In agreement with previous reports¹³, we found that PROCR⁺ cells express high levels of mesenchymal genes such as *Vim*, *Cdh5*, *Zeb1*, *Zeb2*, *Twist1*, and low levels of epithelial genes such as *Epcam*, *Cdh1*, *Krt14*, *Krt5*, *Krt8*, and *Krt18*. The size of the PROCR⁺ population

ranged from 0.3 to 3% of the total mammary epithelium, and it was more prevalent during puberty (Fig. 1C and data not shown).

PROCR+ cells consist of three subpopulations representing MaSC differentiation routes. To investigate the PROCR+ population at a very high resolution we isolated the PROCR+ population by FACS as previously described¹³ and performed scRNA-seq on ~2,000 PROCR+ cells from each time point (Fig. 1A). Remarkably we found that PROCR+ cells are not a single entity but rather consist of three distinct cell clusters (Fig. 1A) that are seen in all developmental stages (Fig. 2A). To investigate the molecular characteristics of these clusters, we first defined the basal and luminal signatures by performing differential expression analysis of the scRNA-seq profiles of differentiated basal and luminal subpopulations. Then, we used the NetBID algorithm^{18,27} to infer the activities for basal or luminal signatures in each cell across all populations. We found that one of the three PROCR+ clusters presented higher luminal signature activity while another presented higher basal activity, whereas the last was undefined (Fig. 2B). For simplicity, we have designated these MaSC subtypes luminal-MaSCs, basal-MaSCs, and core-MaSCs, respectively.

It has been reported that the differentiation potential of a cell can be approximated by computing the signaling promiscuity, or entropy, of the cell's transcriptome, thus providing a means to identify stem cell phenotypes²⁸. Interestingly, the core-MaSC subpopulation presented the highest entropy, indicative of more plasticity and therefore the highest differentiation potential (Fig. 2C). Next, we investigated the lineage relationship of these cells by generating a diffusion plot (Fig. 2D). The differentiated populations relocated to well-defined groups forming a triangle-shaped pattern. In contrast, the core-MaSCs were positioned at the geometrical center of the triangle while the basal and luminal-MaSCs extended from the core-MaSCs toward the corresponding lineage-committed cells. Finally, we dissected the differentiation trajectories of MaSCs by pseudotime analysis^{29,30}. This computational method embeds scRNA-seq profiles in a low-dimensional space where the distance between adjacent cells represents the progression through a continuous but stochastic differentiation process. This analysis showed a gradual bifurcation from core-MaSCs into two branches: basal-MaSCs and luminal-MaSCs (Fig. 2E). Overall, our studies position core-MaSCs at the very top of the MaSC hierarchy.

Single-cell studies of the human breast identify PROCR+ cells analogous to mouse MaSCs. Next, we obtained normal healthy tissue from mastectomies and performed scRNA-seq on the dissociated human MECs. Remarkably, we observed the presence of a small subpopulation highly similar to the mouse PROCR+ MaSCs (Fig. 3A). To obtain enough PROCR+ cells for further analysis, we used a FACS strategy similar to the one used in mice (Fig. 3B) and performed scRNA-seq. Here, we found that human PROCR+ cells contained three subpopulations that are analogous to the mouse core-, luminal and basal-MaSCs (Fig. 3C).

Molecular dissection of the differentiation routes of bipotent MaSCs. ScRNA-seq analysis provides expression data for ~2,000 to 3,000 of all the expressed genes^{31,32}. Although this is enough to sort cell heterogeneity, complete transcriptomic profiles are necessary for studying cell regulation. Thus, we devised a FACS strategy to isolate the core-MaSC, basal-MaSC, and luminal-MaSC subpopulations and subjected them to

bulk RNA-seq (Fig. 4). To this end, we searched our scRNA-seq data to select membrane proteins specifically expressed in each of the MaSC subpopulations. We found that the core- and basal-MaSCs express high levels of CD36 and podoplanin (Pdpn) respectively, while the luminal-MaSCs do not express either of these genes (Fig. 4A). Thus, we purified PROCR⁺ (EpCAM⁺/CD49^{high}/PROCR⁺) cells and used these markers to sort the MaSC subpopulations (CD36⁺ to obtain core-MaSCs, Pdpn⁺ for basal-MaSCs and CD36⁻/Pdpn⁻ for luminal-MaSCs) (Fig. 4B). As expected, PCA analysis of the sc- and the bulk RNA profiles were highly correlated, validating our FACS strategy (Fig. 4C).

To identify key regulators of the differentiation process we used a powerful method that analyzes expression profiling called inference of transcriptional regulatory networks (TRNs)^{26,33-37}. Briefly, a TRN consists of a web of connections (network) between transcription factors (TFs) and their targets. A TF-gene connection can show a positive (activator) or a negative (repressor) correlation. The activity of a TF is calculated based on the expression of its targets and is represented numerically. Subsequently, the TRN is inferred by integrating all the TF-genes hubs. To delineate TRNs, we used the SJARACNe¹⁷ algorithm (an improved implementation of ARACNe³⁸ that has been used extensively in normal and cancer cells^{26,33-37}). Then, we used the MaSC-specific TRN to identify the Master Regulators (MRs) of each MaSC subpopulation. A MR is defined as a TF that is differentially active among the populations studied^{35-37,39-45}. For this, we used NetBID²⁷ to calculate the activity scores of each TF in each MaSC subpopulation. To quantify the significance of each activity score we performed gene set enrichment analysis (GSEA) to assess the enrichment of its predicted targets. Finally, we kept TFs with adjusted p-values <0.05 and further identified MRs based on the ranking of the normalized enrichment score (NES). Our analysis found that a small series of TFs such as *Tcf4*, *Foxp4*, and *Mndal* emerged as top-MRs when core-MaSCs progress to luminal-MaSCs, whereas *Zeb2*, *Id3*, *Twist2*, and *Prrx1* are MRs for the transition from core-MaSCs to basal-MaSCs (Fig.5A and 5B). Interestingly, while some of these TFs are known to be expressed in MaSCs, others represent novel MRs in MaSC fate specification. Of particular interest are *Tcf4* and *Id3*, which have been implicated in development in other systems but not in MaSCs.

Straight *Tcf4* knockout animals are not viable⁴⁶, consequently, whether they present mammary gland abnormalities is unknown. Homozygous *Id3*-KO mice are viable, and mammary abnormalities have been reported in mice lacking expression of other Id proteins. *Id4*-null mice exhibited reduced ductal and alveolar expansion⁴⁷ and *Id2*-null mice showed impaired alveolar differentiation during pregnancy⁴⁸. Individual *Id1* and *Id3* null strains do not present any phenotype due to compensatory mechanisms⁴⁹. Unfortunately, *Id1* and *Id3* double knockout mice are embryonic lethal⁵⁰ preventing further studies. Because these data illustrate the importance of ID proteins in the mammary gland we examined the expression of *Id* genes in our single-cell data (Fig. 5C). Remarkably, we found that although *Id4* expression is very high in mature basal cells, the expression of *Id3* is almost exclusive to basal-MaSCs. Furthermore, *Tcf4* is only expressed in non-basal-MaSCs.

DISCUSSION

The mammary epithelium undergoes continuous expansion and regression during the female lifetime¹⁻⁴. These cycles of expansion need to be tightly regulated in time and space to ensure proper organ function and avoid

pathological consequences. Although the existence of MaSCs supporting mammary epithelial expansion is well documented⁵⁻⁹, the mechanisms regulating homeostasis and differentiation of adult MaSCs are poorly understood. Characterizing these processes is critical for understanding the dynamics of the mammary epithelium and preventing and treating diseases such as breast cancer. However, conflicting results exist regarding the differentiation potential of MaSCs and while some studies have reported the existence of luminal and basal unipotent MaSCs¹⁴, others have found bipotent MaSCs that replenish both lineages^{5,11-13}.

While previous studies have described the expression of protein C receptor (PROCR) as a marker of bipotent MaSCs¹³, our study goes one step further. Through the use of unbiased single-cell RNA sequencing (scRNA-seq) we have found that the PROCR+ MaSC population is composed of a core-MaSC subpopulation located at the top of the mammary epithelial hierarchy as well as two more lineage-committed subpopulations presenting basal and luminal molecular characteristics. Importantly, we have also delineated the activation pattern of the TFs during the early phases of lineage specification. Our studies indicate that *Tcf4* and *Id3* are two master regulators involved in the early acquisition of lineage-specific molecular features in MaSCs. The protein product of *Tcf4* called E2-2, belongs to the bHLH family of transcription factors that form homo- and heterodimers^{51,52}. ID3 is also an HLH member; however, it works as an inhibitor of other bHLH proteins (including E2-2) by blocking their transcriptional activity^{51,52}. Based on their anti-correlative activity observed in our preliminary data, we propose a model in which the acquisition of luminal features may be triggered by the expression of genes controlled by E2-2 whereas inhibition of E2-2 by ID3 leads to basal differentiation. Additionally, while ID3 and E2-2 are well-known partners, it is also possible that both of them work independently of each other. Thus, inhibition of other bHLH partners by ID3 and interaction of E2-2 with additional transcriptional regulators should also be considered. Of course, these alternatives are not mutually exclusive, and most likely, MaSC fate is influenced by a network coordinated by these two bHLH proteins consisting of common and specific downstream gene targets (Fig. 6). Thus, while our studies have unveiled *Tcf4* and *Id3* as some of the initial TF controlling lineage-specification, additional studies are needed to address the extension of this network and the importance of each hub.

Individual author contributions

JS and JY designed the research plan. KKY, EN, JS, and JY wrote the manuscript. EN, RW, DA coordinated the animal experiments and performed the mammary gland dissection and purification of mammary epithelial cells by FACS. KKY performed the computational and systems biology analysis. XD and LD developed the data portal. QP and CQ assisted single-cell RNA-seq data analysis. PM and AG provided additional support for animal care and mammary gland dissection. BJ, CR, BS, SN, and JE performed single-cell and bulk genomics profiling. YW, HS, and HC coordinated animal processing for single-cell studies.

References

- 1 Howard, B. A. & Gusterson, B. A. Human breast development. *J Mammary Gland Biol Neoplasia* **5**, 119-137 (2000).
- 2 Ip, M. M. & Asch, B. B. Introduction: an histology atlas of the rodent mammary gland and human breast during normal postnatal development and in cancer. *J Mammary Gland Biol Neoplasia* **5**, 117-118 (2000).
- 3 Macias, H. & Hinck, L. Mammary Gland Development. *Wiley Interdiscip Rev Dev Biol* **1**, 533-557, doi:10.1002/wdev.35.
- 4 Macias, H. & Hinck, L. Mammary Gland Development. *Wiley interdisciplinary reviews. Developmental biology* **1**, 533-557, doi:10.1002/wdev.35 (2012).
- 5 van Amerongen, R., Bowman, A. N. & Nusse, R. Developmental stage and time dictate the fate of Wnt/beta-catenin-responsive stem cells in the mammary gland. *Cell stem cell* **11**, 387-400, doi:10.1016/j.stem.2012.05.023 (2012).
- 6 Asselin-Labat, M. L. *et al.* Control of mammary stem cell function by steroid hormone signalling. *Nature* **465**, 798-802, doi:10.1038/nature09027 (2010).
- 7 Joshi, P. A. *et al.* Progesterone induces adult mammary stem cell expansion. *Nature* **465**, 803-807, doi:10.1038/nature09091 (2010).
- 8 Lloyd-Lewis, B., Harris, O. B., Watson, C. J. & Davis, F. M. Mammary Stem Cells: Premise, Properties, and Perspectives. *Trends Cell Biol* **27**, 556-567, doi:10.1016/j.tcb.2017.04.001 (2017).
- 9 Visvader, J. E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev* **28**, 1143-1158, doi:10.1101/gad.242511.114 (2014).
- 10 Wahl, G. M. & Spike, B. T. Cell state plasticity, stem cells, EMT, and the generation of intra-tumoral heterogeneity. *NPJ Breast Cancer* **3**, 14, doi:10.1038/s41523-017-0012-z (2017).
- 11 Wuidart, A. *et al.* Quantitative lineage tracing strategies to resolve multipotency in tissue-specific stem cells. *Genes Dev* **30**, 1261-1277, doi:10.1101/gad.280057.116 (2016).
- 12 Prater, M. D. *et al.* Mammary stem cells have myoepithelial cell properties. *Nat Cell Biol* **16**, 942-950, 941-947, doi:10.1038/ncb3025 (2014).
- 13 Wang, D. *et al.* Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* **517**, 81-84, doi:10.1038/nature13851 (2015).
- 14 Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189-193, doi:10.1038/nature10573 (2011).
- 15 Visvader, J. E. & Lindeman, G. J. Mammary stem cells and mammapoiesis. *Cancer Res* **66**, 9798-9801, doi:10.1158/0008-5472.CAN-06-2254 (2006).
- 16 Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993-997, doi:nature04496 [pii]10.1038/nature04496 (2006).
- 17 Khatamian, A., Paull, E. O., Califano, A. & Yu, J. SJARACNe: a scalable software tool for gene network reverse engineering from big data. *Bioinformatics*, doi:10.1093/bioinformatics/bty907 (2018).
- 18 Du, X. *et al.* Hippo/Mst signalling couples metabolic state and immune function of CD8alpha(+) dendritic cells. *Nature* **558**, 141-145, doi:10.1038/s41586-018-0177-0 (2018).
- 19 Yu, Q. C., Song, W., Wang, D. & Zeng, Y. A. Identification of blood vascular endothelial stem cells by the expression of protein C receptor. *Cell research* **26**, 1079-1098, doi:10.1038/cr.2016.85 (2016).
- 20 Xue, M., Dervish, S., Chan, B. & Jackson, C. J. The Endothelial Protein C Receptor Is a Potential Stem Cell Marker for Epidermal Keratinocytes. *Stem Cells* **35**, 1786-1798, doi:10.1002/stem.2630 (2017).
- 21 Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 96, doi:10.1038/s12276-018-0071-8 (2018).
- 22 Kumar, P., Tan, Y. & Cahan, P. Understanding development and stem cells using single cell-based analyses of gene expression. *Development* **144**, 17-32, doi:10.1242/dev.133058 (2017).
- 23 Shehata, M. *et al.* Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland. *Breast cancer research : BCR* **14**, R134, doi:10.1186/bcr3334 (2012).
- 24 Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84-88, doi:nature04372 [pii]10.1038/nature04372 (2006).

- 25 Asselin-Labat, M. L. *et al.* Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nat Cell Biol* **9**, 201-209, doi:10.1038/ncb1530 (2007).
- 26 Ding, H. *et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat Commun* **9**, 1471, doi:10.1038/s41467-018-03843-3 (2018).
- 27 Wang, Y. *et al.* LKB1 orchestrates dendritic cell metabolic quiescence and anti-tumor immunity. *Cell research* **29**, 391-405, doi:10.1038/s41422-019-0157-4 (2019).
- 28 Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* **8**, 15599, doi:10.1038/ncomms15599 (2017).
- 29 Haghverdi, L., Buttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845-848, doi:10.1038/nmeth.3971 (2016).
- 30 Haghverdi, L., Büttner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989-2998, doi:10.1093/bioinformatics/btv325 (2015).
- 31 Pal, B. *et al.* Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat Commun* **8**, 1627, doi:10.1038/s41467-017-01560-x (2017).
- 32 Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun* **8**, 2128, doi:10.1038/s41467-017-02001-5 (2017).
- 33 Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838-847, doi:10.1038/ng.3593 (2016).
- 34 Putcha, P. *et al.* HDAC6 activity is a non-oncogene addiction hub for inflammatory breast cancers. *Breast cancer research : BCR* **17**, 149, doi:10.1186/s13058-015-0658-0 (2015).
- 35 Aytes, A. *et al.* Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer cell* **25**, 638-651, doi:10.1016/j.ccr.2014.03.017 (2014).
- 36 Bisikirska, B. C. *et al.* STK38 is a critical upstream regulator of MYC's oncogenic activity in human B-cell lymphoma. *Oncogene* **32**, 5283-5291, doi:10.1038/onc.2012.543 (2013).
- 37 Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318-325, doi:10.1038/nature08712 (2010).
- 38 Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7, doi:1471-2105-7-S1-S7 [pii]10.1186/1471-2105-7-S1-S7 (2006).
- 39 Basso, K. *et al.* Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells. *Blood* **115**, 975-984, doi:10.1182/blood-2009-06-227017 (2010).
- 40 Wang, K. *et al.* Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol* **27**, 829-839, doi:nbt.1563 [pii]10.1038/nbt.1563 (2009).
- 41 Tomljanovic, Z., Patel, M., Shin, W., Califano, A. & Teich, A. F. ZCCHC17 is a master regulator of synaptic gene expression in Alzheimer's disease. *Bioinformatics* **34**, 367-371, doi:10.1093/bioinformatics/btx608 (2018).
- 42 Rajbhandari, P. *et al.* Cross-Cohort Analysis Identifies a TEAD4-MYC Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. *Cancer discovery* **8**, 582-599, doi:10.1158/2159-8290.CD-16-0861 (2018).
- 43 Dutta, A. *et al.* Identification of an NKX3.1-G9a-UTY transcriptional regulatory network that controls prostate differentiation. *Science* **352**, 1576-1580, doi:10.1126/science.aad9512 (2016).
- 44 Ikiz, B. *et al.* The Regulatory Machinery of Neurodegeneration in In Vitro Models of Amyotrophic Lateral Sclerosis. *Cell Rep* **12**, 335-345, doi:10.1016/j.celrep.2015.06.019 (2015).
- 45 Aubry, S. *et al.* Assembly and interrogation of Alzheimer's disease genetic networks reveal novel regulators of progression. *PLoS One* **10**, e0120352, doi:10.1371/journal.pone.0120352 (2015).
- 46 Flora, A., Garcia, J. J., Thaller, C. & Zoghbi, H. Y. The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proc Natl Acad Sci U S A* **104**, 15382-15387, doi:10.1073/pnas.0707456104 (2007).
- 47 Dong, J. *et al.* ID4 regulates mammary gland development by suppressing p38MAPK activity. *Development* **138**, 5247-5256, doi:10.1242/dev.069203 (2011).
- 48 Mori, S., Nishikawa, S. I. & Yokota, Y. Lactation defect in mice lacking the helix-loop-helix inhibitor Id2. *EMBO J* **19**, 5772-5781, doi:10.1093/emboj/19.21.5772 (2000).

- 49 de Candia, P., Benera, R. & Solit, D. B. A role for Id proteins in mammary gland physiology and tumorigenesis. *Adv Cancer Res* **92**, 81-94, doi:10.1016/S0065-230X(04)92004-0 (2004).
- 50 Lyden, D. *et al.* Id1 and Id3 are required for neurogenesis, angiogenesis and vascularization of tumour xenografts. *Nature* **401**, 670-677, doi:10.1038/44334 (1999).
- 51 Wang, L. H. & Baker, N. E. E Proteins and ID Proteins: Helix-Loop-Helix Partners in Development and Disease. *Dev Cell* **35**, 269-280, doi:10.1016/j.devcel.2015.10.019 (2015).
- 52 Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol* **20**, 429-440 (2000).

Figure Legend

Fig. 1 Single-cell transcriptomics of mammary epithelium: A) Purification and scRNA-seq profile of Balb/cj virgin females as described in the text. The tSNE plot shows all the different time points merged and an enlarged plot of the MaSC subpopulations. B) tSNE plot from A showing the expression of specific genes (purple). C) The bar plot shows the expression of lineage-specific markers in mammary epithelial subpopulations. Cell subpopulations are color-coded.

Fig. 2 Molecular characterization of MaSC populations: A) The tSNE plot and the bar plots show the distribution of MaSCs populations in all developmental stages evaluated. B) Basal and luminal activity of MaSC subpopulations and differentiated cells and C) entropy (transcriptional plasticity) calculated as described in the text. D) Diffusion plot showing the lineage relationship of the six subpopulations. E) Pseudotime analysis showing a divergent trajectory from core-MaSCs to basal-MaSCs and luminal-MaSCs. In A and D the colored circles indicate the position of the MaSC subpopulations.

Fig. 3 Purification of PROCR+ human cells: A) tSNE plots showing the human mammary epithelial subpopulations. The violin plot shows the expression of PROCR in each subpopulation. B) FACS profiles showing the purification of PROCR+ cells. C) tSNE-plots showing the subpopulations as well as a diffusion map and the basal and luminal activity of the purified PROCR+ MaSCs from B.

Fig. 4 Purification of MaSC subpopulations: A) tSNE plots showing the expression of *Cd36* and *Pdpr* in MaSCs subpopulations. B) FACS plots showing the sorting strategy for isolation of MaSCs subpopulations. C) PCA plots showing the correlation between scRNA-seq and bulk RNA-seq data.

Fig. 5 Master Regulators of MaSC subpopulations: A) NetBID activity analysis for TFs specifically active in basal-MaSC vs luminal-MaSCs. The dot size is proportional to the normalized enrichment score (NES) of the TF targets. B) Pseudotime analysis showing the change in the activity of MRs during the progression from core-MaSCs to basal-MaSCs or luminal-MaSCs. C) Bar plot showing the expression of specific genes in mammary subpopulations (color-coded).

Fig. 6 Proposed E2-2(Tcf4)/ID3 switch model: Bipotent adult core-MaSCs express high levels of the E2-2(Tcf4) transcription factor. During the early stages of lineage commitment, increasing expression of the HLH partner ID3 in some of the core-MaSCs inhibits Tcf4, and potentially other partners, leading to the acquisition of basal characteristics. In contrast, core-MaSCs that enter the differentiation process and do not express ID3 will acquire luminal lineage characteristics.

MATERIALS AND METHODS

Mammary gland dissociation and epithelial enrichment (mouse).

Experimental animals

Wild-type BALB/cJ virgin female mice were obtained from The Jackson Laboratory. The mice were sacrificed at the specified age using a CO₂ chamber. Animal maintenance and all experiments were performed under the Institutional Animal Care and Use Committee (IACUC) guidelines.

Determination of estrous cycle stage

Vaginal smears were taken by flushing with 15 µl of sterile PBS which were then spread on a slide and allowed to dry. The cells were visualized after staining with crystal violet and washing twice with deionized water. The stages were then identified as described (Byers *et al.*, 2012).

Mammary gland dissociation and epithelial enrichment

The 3rd and 4th mammary glands were dissected from female mice, pooled, and minced followed by enzymatic digestion in DMEM/F12 media containing 2 mg/ml Collagenase A (Roche #11088793001) and 100 U/ml Hyaluronidase (Sigma #H3506) for 2 hours at 37°C. Single-cell suspensions were obtained via sequential incubations in pre-warmed 0.25% Trypsin-EDTA followed by 5 mg/ml Dispase II (Life Technologies #17105-041) in PBS with 0.1 mg/ml DNase I (Stemcell #07900), pipetting up and down at each step. Finally, the cells were incubated in Red Blood Cell Lysing Buffer (Sigma #R7757) and filtered through a 40 µm cell strainer. The samples were enriched for Lin⁻ epithelial cells using the EasySep Mouse Epithelial Cell Enrichment Kit II (Stemcell #19758) according to the protocol.

Fluorescence-activated cell sorting (FACS)

The epithelial cell-enriched cell suspensions were incubated with the following antibodies: EpCAM-PerCP/Cy5.5, CD49f-APC, PROCR-PE, and Sca-1-BV421 (BioLegend #118220, #313616, #141504, and BD #562729). The LIVE/DEAD™ Fixable Aqua Dead Cell Stain (ThermoFisher #L34957) was used to determine cell viability. Antibody incubations were performed for 15 minutes on ice. The cells were sorted using a FACSaria II (BD) cell sorter using the following method: following size and doublet discrimination and viability screening, EpCAM⁺ cells were sorted for the Total Epithelium samples. To enrich for MaSCs, EpCAM⁺CD49fhighPROCR⁺ cells were sorted.

Mammary gland dissociation and epithelial enrichment (Human)

Fresh tissue was obtained from mastectomy surgeries via the Biorepository Core at Mount Sinai according to an approved protocol. Samples were dissociated according to a published method (Shehata and Stingl, 2017). Briefly, the tissue was minced and enzymatically digested in DMEM/F12 media containing 2 mg/ml Collagenase A (Roche #11088793001), 100 U/ml Hyaluronidase (Sigma #H3506), 2% BSA, 5 µg/ml insulin, 50 µg/ml gentamycin, and 10mM HEPES overnight at 37°C. The samples were enriched for epithelial cells via centrifugation at 200xg for 5 minutes. Single-cell suspensions were obtained via sequential incubations in pre-warmed 0.25% Trypsin-EDTA followed by 5 mg/ml Dispase II (Life Technologies #17105-041) in PBS with 0.1 mg/ml DNase I (Stemcell #07900). The dissociated cells were subjected to FACS using the same method as above with the following antibodies: EpCAM-PerCP/Cy5.5, CD49f-APC, and PROCR-PE (BioLegend #324213, #313615, and #351903, respectively).

Bulk RNA library preparation and sequencing

For samples with limited amount of RNA, we used NEBnext Single Cell/Low RNA Library prep Kit for Illumina to generate sequencing libraries (New England BioLabs). Libraries were sequenced on Illumina NextSeq 500 system (Illumina).

For samples with abundant RNA, stranded RNA libraries were prepared with 100ng total RNA using the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche). RNA sequencing was performed on Illumina NovaSeq 6000 system (Illumina).

Single Cell RNA Sequencing (scRNA-Seq) sample processing

Flow-sorted single cells were captured with the 10xGenomics Chromium Controller (10xGenomics). Libraries were prepared with the Chromium Single Cell Gene Expression 3' v2 or v3 kits, and sequenced on either Illumina HiSeq 2500 or NextSeq 500 systems (Illumina). Sequencing data were run on Cell Ranger to generate feature barcode matrices and to perform clustering and gene expression analysis.

Single cell RNASeq data analysis

Sequencing data were processed and converted to counts by Cell Ranger, the software tool developed by 10X genomics. Counts data were then converted to gene-by-cell expression matrix by Seurat¹. In brief, cells were filtered based on the unique number of genes and UMIs. Cells with UMI number and the unique number of genes too high or too low were filtered. Cells with a high fraction of mitochondrial genes were further filtered. The number of counts in the remaining cells were normalized by a scaling factor 100000 and gene expression level was quantified as $E = \log_2(C + 1)$, where C is the normalized counts for gene i in cell j. The resultant expression matrix was used for clustering analysis by our in-house pipeline scMINER. To take into account of the non-linear similarity between cells, scMINER quantified the distance between cells using mutual information. Leading MDS components of the distance matrix were used for identifying clusters using consensus clustering. Clusters were then manually curated using marker genes from literature. Comprehensive identification of marker genes in each cluster was done by differential expression analysis using Seurat.

Developmental hierarchy and network inference

The developmental stages and pseudotime were inferred based on expression of individual genes using diffusion maps² and the tool Monocle 2³. The ordering of cells based on diffusion maps and pseudotime was supported by the entropy measure, a metric to quantify the developmental plasticity, defined as $-S$, where S is the relative expression⁴. To better differentiate different populations, activities of various gene sets or signatures were used on top of marker genes expression. Activities of specific gene sets were calculated using NetBID^{19,39}, which is essentially a weighted measure of expression of genes in the set. Network inference was performed using SJARACNE¹⁸.

In short, cluster-specific network was inferred based on the expression profile of the cells in the cluster based on mutual information dependency. Transcription factors with a high number of targets in the network were regarded as potential master regulators (MR). Gene set enrichment analysis (GSEA) and curated gene sets from MSigDB{Liberzon, 2011 #897 were used to identify potential biological functions of the targets. The activity of the potential regulators is calculated from the expression of its targets, and differential activities were inferred using NetBID. The regulators whose activities are correlated with the pseudotime were operationally defined as MR for the development.

References

- 1 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
- 2 Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989-2998, doi:10.1093/bioinformatics/btv325 (2015).
- 3 Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386, doi:10.1038/nbt.2859 (2014).
- 4 Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* **8**, 15599, doi:10.1038/ncomms15599 (2017).

Figure.1

Figure. 2

Figure 3

Figure 4.

Figure 5

Figure 6