© 2022. Published by The Company of Biologists Ltd | Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

TOOLS AND RESOURCES

Single-cell analysis reveals the Comma-1D cell line as a unique model for mammary gland development and breast cancer

Rachel L. Werner[1,*], Erin A. Nekritz[1,*], Koon-Kiu Yan[2,*], Bensheng Ju[2], Bridget Shaner[2], John Easton[2], Jiyang Yu[2,‡] and Jose Silva[1,‡]

ABSTRACT

The mammary gland epithelial tree contains two distinct cell populations, luminal and basal. The investigation of how this heterogeneity is developed and how it influences tumorigenesis has been hampered by the need to perform studies on these populations using animal models. Comma-1D is an immortalized mouse mammary epithelial cell line that has unique morphogenetic properties. By performing single-cell RNA-seq studies, we found that Comma-1D cultures consist of two main populations with luminal and basal features, and a smaller population with mixed lineage and bipotent characteristics. We demonstrated that multiple transcription factors associated with the differentiation of the mammary epithelium *in vivo* also modulate this process in Comma-1D cultures. Additionally, we found that only cells with luminal features were able to acquire transformed characteristics after an oncogenic HER2 (also known as ERBB2) mutant was introduced in their genomes. Overall, our studies characterize, at a single-cell level, the heterogeneity of the Comma-1D cell line and illustrate how Comma-1D cells can be used as an experimental model to study both the differentiation and the transformation processes *in vitro*.

KEY WORDS: Comma-1D, Mammary epithelium differentiation, Stem cells, Oncogenesis, Breast cancer

INTRODUCTION

The adult female mammary consists of a core epithelial tree organized into ducts and terminal lobules that are surrounded by a variety of stromal cells such as fibroblast and adipocytes (Macias and Hinck, 2012; Ip and Asch, 2000). The mammary epithelium is organized into a bilayer of basal and luminal cells that are molecularly and functionally different. Although terminally differentiated luminal and basal mammary epithelial cells have been well characterized molecularly (Visvader and Lindeman, 2006; Visvader and Stingl, 2014), how the differentiation process is orchestrated, and what are the molecular determinants that mediate differentiation are much less understood.

[1]Graduate School, Department of Pathology, Icahn School of Medicine at Mount Sinai Hospital, New York, NY 10029, USA. [2]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.
*These authors contributed equally to this work

‡Authors for correspondence (jose.silva@mssm.edu; jiyang.yu@stjude.org)

E.A.N., 0000-0002-1916-5193; B.S., 0000-0002-2312-2196; J.Y., 0000-0003-1244-4429; J.S., 0000-0001-5147-8950

Handling Editor: Andrew Ewald
Received 2 September 2021; Accepted 11 April 2022

The past decade has brought extraordinary advancements in genomic approaches that have allowed us to compare cell types at a genome-wide level (e.g. expression profiling and chromatin epigenetics). These technologies generate data connecting hundreds of genes to a specific cell type. However, potentially relevant genes need to be further investigated functionally to assign specific biological roles. When investigating the biology of the mammary gland, one of the major barriers is the need to perform these studies *in vivo*. Although *in vivo* validation is critical to fully confirm the function of a particular gene, it also imposes a strong limitation on the number of genes that can be investigated at a time. On the other hand, *in vitro* systems represent tractable alternatives that can be used as testing platforms to easily interrogate multiple candidates at once before transitioning interesting candidates to *in vivo* settings. Dozens of transformed cell lines have been established *in vitro* and fully characterized (Ghandi et al., 2019). However, these lines are a clonal expansion of cancer cells and are not suitable for the study of differentiation processes. A limited number of immortalized, but non-transformed, human (Qu et al., 2015; Kumar et al., 2018) and mouse (Vaidya et al., 1978; Howard et al., 1983; Anderson et al., 1979) cell lines of mammary epithelial origin have been also established. Unfortunately, almost all of these models represent lineage-committed cells without the ability to generate a heterogeneous progeny. The only exception is the Comma-1D cell line.

The Comma-1D cell line was originally established from a normal mouse mammary gland at mid-pregnancy (Danielson et al., 1984). It shows remarkable morphological heterogeneity in culture, with cells that specifically express luminal and basal cytokeratins. When transplanted into cleared mammary fat pads, these cells also show unique morphogenetic properties *in vivo* and form outgrowths that resemble a fully functional mammary epithelial tree, including milk-producing alveoli during pregnancy (Danielson et al., 1984; Kittrell et al., 2011). Overall, these characteristics indicate the presence of multipotent cells among Comma-1D cultured cells. In this regard, multipotent cells have been found enriched in Comma-1D cells expressing high levels of the stem cell antigen (Sca-1^{High}; Sca-1 is also known as Ly6a) (Ibarra et al., 2007; Deugnier et al., 2006). However, precise identification and molecular characterization of the multipotent cells embedded in Comma-1D cultures has not been performed.

Here, we utilized single-cell RNA-sequencing (scRNA-seq) (Hwang et al., 2018; Kumar et al., 2017) to dissect the heterogeneity of the Comma-1D cells at the individual cell level. Our studies found that the bulk of Comma-1D culture consists of cells expressing markers and networks of lineage-committed luminal or basal subpopulations. After purification by flow cytometry, these cells were only able to generate homogeneous cultures containing cells of the same lineage (unipotent). We also identified a small population of cells that express bilineage markers. Interestingly,

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

analysis tracking the evolutionary trajectory of Comma-1D cells positioned this population at the top of the differentiation hierarchy. In contrast to lineage-committed luminal or basal cells, these cells were able to fully regenerate heterogeneous cultures containing all original Comma-1D cell subpopulations (multipotent). Importantly, these cells only represent a small number of the Sca-1^{High} cells. We also performed network analysis coupled with functional studies to study the molecular determinants that impact the differentiation of Comma-1D cells. These studies identified multiple genes that have been previously found to be important during the differentiation process of mammary epithelial cells *in vivo*.

Increasing evidence indicates that the molecular alterations found in human breast cancers do not accumulate in random cells. Instead, specific alterations commonly affect specific subpopulations, and this specificity determines the breast cancer subtypes seen in the clinic (Latil et al., 2017; Visvader, 2011; Gilbertson, 2011; Skibinski and Kuperwasser, 2015). Thus, we also investigated whether unipotent and bipotent Comma-1D cells can be transformed by the bona fide breast cancer oncogene *Her2* (also known as *Erbb2*) (Banerji et al., 2012; Curtis et al., 2012). Remarkably, we found that only Comma-1D cells with luminal features acquired transformed properties.

Overall, the studies presented here characterize, at a single-cell level, the heterogeneity of the Comma-1D cell line and illustrate how Comma-1D cells can be used as a relevant experimental model to study both differentiation and transformation processes *in vitro*.

RESULTS

Single-cell RNA-seq analysis reveals the presence of three subpopulations in Comma-1D cells

scRNA-seq is a powerful method to identify cell subtypes and track the trajectories of cell lineages (Hwang et al., 2018; Kumar et al., 2017). To study the heterogeneity of Comma-1D cells, we performed scRNA-seq on ∼20,000 cells growing exponentially in cell culture using a droplet-based platform (10X Genomics) (Hwang et al., 2018; Ding et al., 2018). Clustering analysis of all sequenced cells based on *t*-distributed stochastic neighbor embedding (tSNE plots) (Kumar et al., 2017; Hwang et al., 2018) found two major clusters comprising ∼70% and ∼29% of the cells (Fig. 1A). To discover the identity of these clusters, we first collected the expression profiles of mammary luminal and basal cells obtained by scRNA sequencing generated by the Tabula Muris consortium (Tabula Muris, 2020). Next, we used the tool SciBet (Li et al., 2020) to obtain a gene list with the most-specific basal and luminal epithelial genes (see details in Materials and Methods section and Table S1). Finally, we calculated a luminal and basal score for every cell in our single-cell Comma-1D data by computing the average expression of these basal and luminal markers. This study revealed that each of the larger clusters represents mammary epithelial cells with specific luminal or basal characteristics (Fig. 1B). Thus, the luminal Comma-1D population expressed high levels of *Krt8*, *Krt18*, *Cd14* and *Epcam*, whereas the basal Comma-1D population expressed showed higher expression of *Krt5*, *Krt14* and *Trp63* (Fig. 1C; Fig. S1A). For simplification, in this manuscript, we will call these populations Comma-1D-luminal (C1D-L) and Comma-1D-basal (C1D-B). Additionally, a smaller population accounting for ∼1–2% of all total cells was also identified (Fig. 1A). Remarkably, this small cluster showed mixed features of both luminal and basal cells (Fig. 1B,C; Fig. S1A). Here, we will call this population Comma-1D bilinage (C1D-bi).

Comma-1D cells with the ability to reconstitute a functional mammary gland upon transplantation *in vivo* have been associated with high levels of Sca-1 protein (*Ly-6A* and *Ly-6E* genes) (Ibarra et al., 2007; Deugnier et al., 2006). Interestingly, the expression of both genes was higher in C1D-B and C1D-bi than in C1D-L (Fig. S1B).

Intrigued by these results, we investigated the lineage relationship of the three clusters by generating a diffusion plot (Fig. 1D). Here, we observed that the C1D-L and C1D-B relocated in well-defined groups separated from each other. In contrast, the small subpopulation was positioned forming a vertex between the two clusters. Finally, we also dissected the differentiation trajectories of the three Comma-1D clusters by pseudotime analysis (Haghverdi et al., 2016; Haghverdi et al., 2015). This computational method embeds scRNA-seq profiles in a low-dimensional space where the distance between adjacent cells represents the progression through a continuous differentiation process. This showed a gradual bifurcation from the cells belonging to the small cluster into two different branches of C1D-L and C1D-B (Fig. 1D; Fig. S1C).

Overall, these results indicate that Comma-1D cultures contain at least three distinct populations. Two of these populations express markers of luminal or basal mammary epithelium, whereas the third one shows an expression pattern of both lineages. Remarkably, our computational studies suggested that the small cluster with mixed lineage features represents bipotent cells that can generate C1D-L and C1D-B.

Bipotent C1D-bi cells can regenerate the original heterogeneity of Comma-1D cells

Based on the data described above, we hypothesized that the presence of cells with luminal and basal characteristics in the Comma-1D cultures was maintained by the C1D-bi population. To test this hypothesis, we designed a fluorescence-activated cell sorting (FACS) strategy based on the differential expression of Epcam and Sca-1 among the three Comma-1D clusters. The C1D-B population was separated from the rest based on their low expression of Epcam. C1D-bi was separated from C1D-L based on its high expression of Sca-1. Noticeably, a vast majority of the C1D-B cells had high levels of Sca-1. Thus C1D-B was Epcam^{Low}/Sca1^{High}, C1D-L was Epcam^{High}/Sca1^{Low} and C1D-bi was Epcam^{High}/Sca1^{High} (Fig. 2A). After FACS, we performed RNA-seq of the purified populations and compared them with the expression data obtained from scRNA-seq. As expected, hierarchical clustering and principal component analysis revealed that the purified cells faithfully represent the three populations seen by single-cell analysis (Fig. 2B; Fig. S2A).

Next, we looked at the cellular heterogeneity of these populations. For this, we combined microscopic visualization of the purified cells with forward scatter (FSC) and side scatter (SSC) flow cytometry analysis. This study revealed clear morphological differences (Fig. 2C). C1D-B are larger cells presenting distinctive lamellipodia all around the cell. C1D-L are smaller cells without lamellipodia, with decreased cytoplasmic content and that had reduced granularity. C1D-bi are of intermediate size, with low granularity and presenting membrane protrusions that give them a stellate morphology.

Matrigel is a mixture of proteins, including laminins, collagens, and other components commonly found in extracellular matrices. When epithelial cells are embedded and cultured in this mix, they generate 3D organoids that resemble the epithelial organization found in normal tissues more closely than standard 2D cultures (Lee et al., 2007). Thus, we cultured the three purified Comma-1D populations in Matrigel to compare their morphogenetic capabilities (Fig. 2D). Remarkably, clear differences were observed. Whereas

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

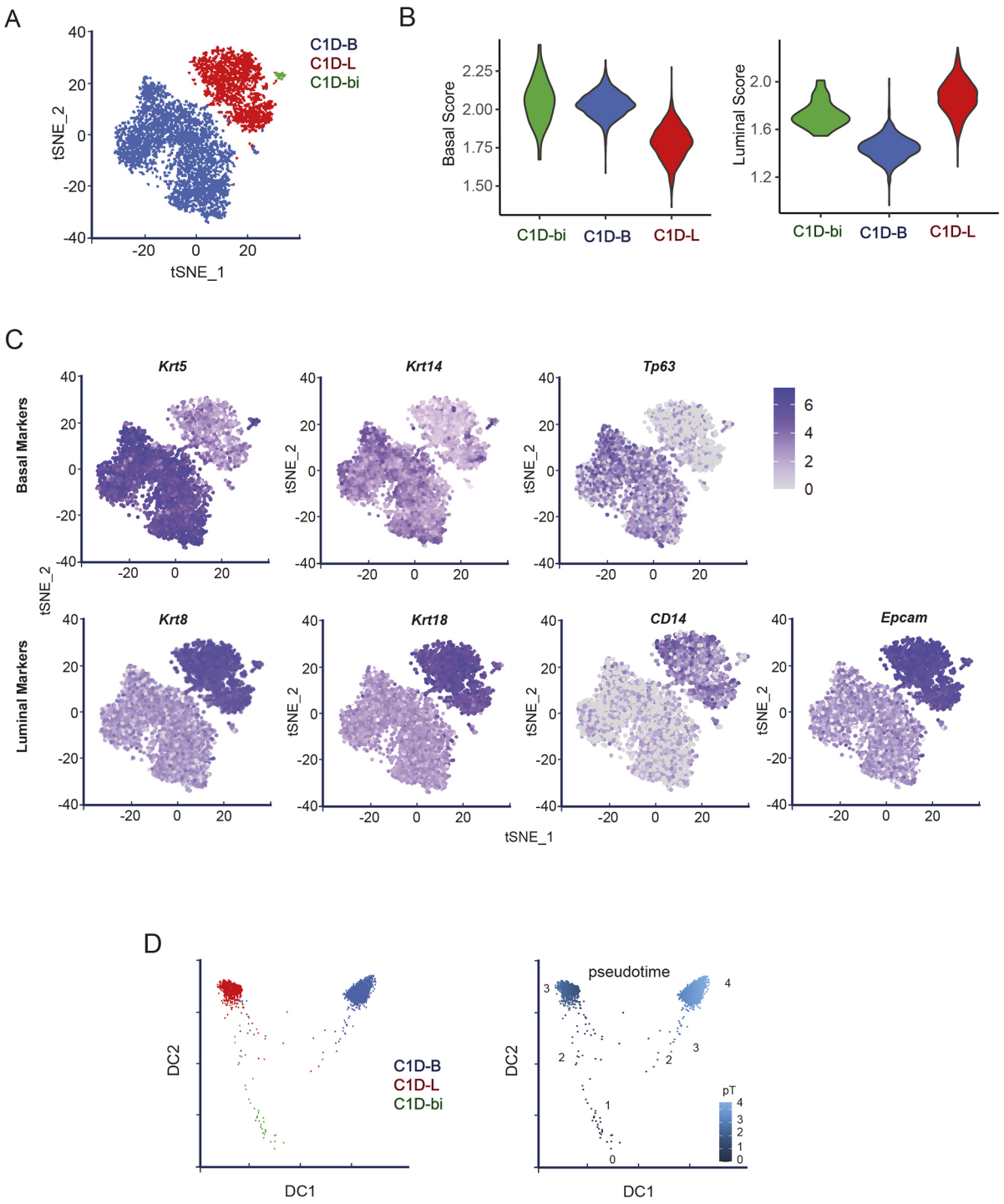

Fig. 1. Single-cell analysis of Comma-1D cells. (A) tSNE plot revealing that comma-1D cultures contain three populations. (B) Expression profiles of the three subpopulations from A revealed that C1D-bi cells express both luminal and basal genes. Basal and luminal genes were grouped in a single score. Results are given as a violin plot. (C) tSNE plot with the overlaid expression of lineage-specific genes. (D) Diffusion (left) and pseudotime (right) plots indicating the lineage relationship and divergent trajectory from C1D-bi to C1D-L and CiD-L. Results shown are representative of $n \geq 3$ experiments.

C1D-B efficiently regenerated spherical organoids in less than a week, C1D-L formed a large number of underdeveloped structures. Although some organoids also emerged from C1D-L, these predominantly presented irregular shapes. Mixed phenotypes were seen in C1D-bi cultures containing spherical and non-spherical organoids, as well as underdeveloped structures.

To further test the functional capabilities of FACS-purified cells we maintained them in culture for a few weeks and performed flow cytometry analysis of the cultures at different time points (Fig. 2E). This study showed that both C1D-L and C1D-B cultures largely maintained homogeneous populations through time. Remarkably, C1D-bi cultures progressively increased the content in C1D-L and C1D-B cells until the original percentages found in parental Comma-1D cultures were reached. Interestingly, when grown independently, C1D-B proliferated faster than C1D-L (Fig. S2). It was not possible to test the growth of

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

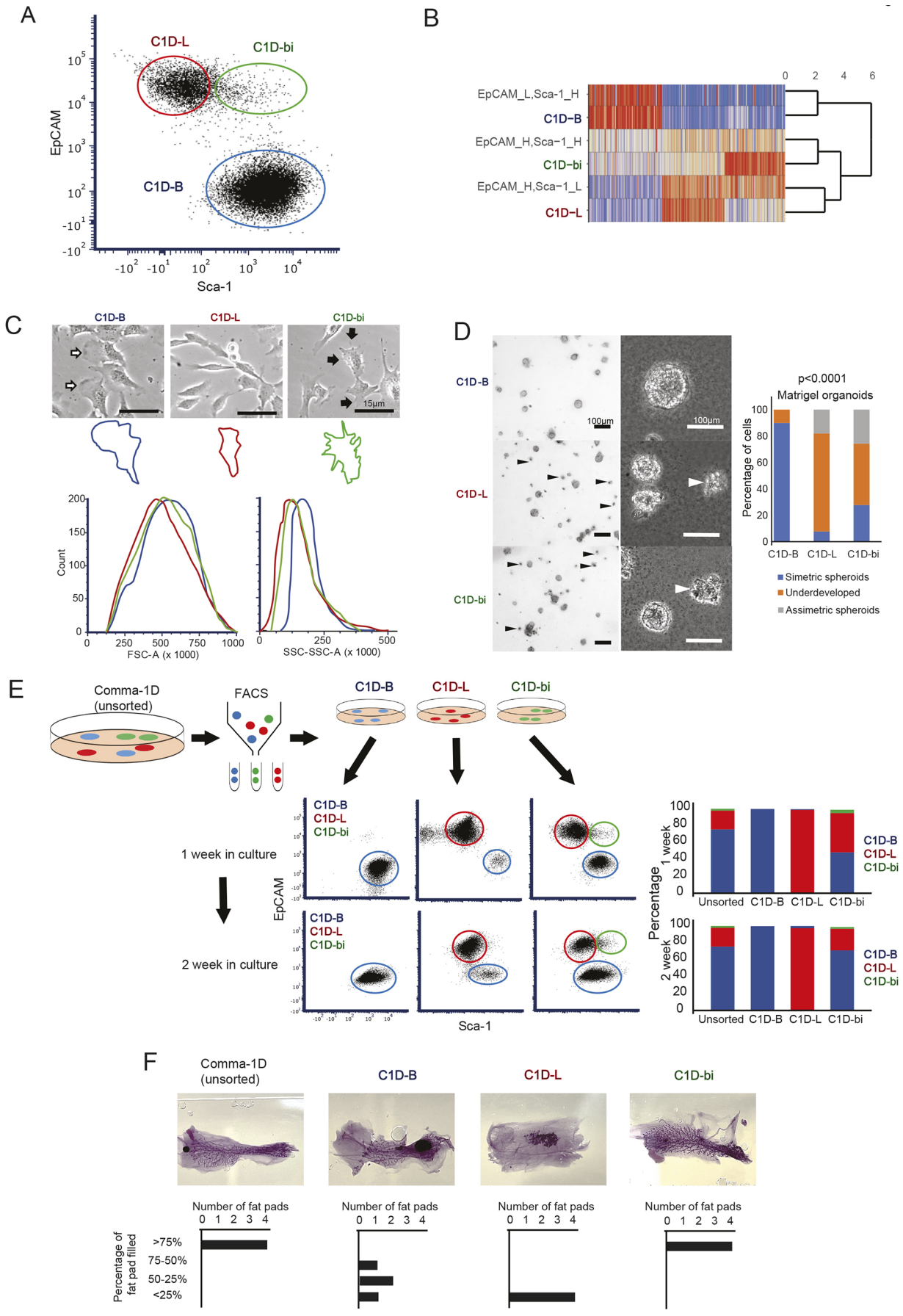

Fig. 2. See next page for legend.

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

Fig. 2. Cellular and functional characterization of Comma-1D populations. (A) FACS strategy used to isolate Comma-1D populations. (B) Unsupervised cluster analysis using the top differentially expressed genes showing that purified cells represent the populations identified in sc-seq analysis. (C) Morphology in standard 2D culture of the Comma-1D subpopulations. White arrows indicate the presence of lamellipodia and black arrows indicate acicular membrane protrusions. Representative morphologies are also sketched below the pictures. Differences in cell size (FSC) and granularity (SSC) distribution among the populations are shown in the graphs. Results in A–C are representative of $n \geq 3$ experiments. (D) Morphology in 3D Matrigel culture of the Comma-1D populations. Quantification of organoid formation efficacy is shown in the bar graph. The result indicates the combination of $n=3$ different experimental replicates with >100 spheroids counted for each population in each of the replicates. The *P*-value was calculated by one-way ANOVA with Chi-squared test. (E) The panel shows the percentage of each Comma-1D subpopulation (evaluated by FACS) that is generated by C1D-B, C1D-L and C1D-bi after these cells are FACS purified and cultured for 1 and 2 weeks. (F) Carmine Red staining showing the reconstitution of mammary epithelium after orthotopic fat pad transplantation of FACS-purified Comma-1D populations. The width of the images shown represents 2 inches. The bar graph indicates the percentage of fat pad filled ($n=4$).

C1D-bi because they rapidly transitioned into more committed populations.

Finally, we assayed the reconstitution potential of each of these populations *in vivo*. Thus, 250,000 FACS purified cells were transplanted orthotopically in cleared mammary fat pads of syngeneic 3-week-old female mice as previously described (Danielson et al., 1984; Kittrell et al., 2011), and the extent of the reconstituted mammary tree was evaluated by Carmine Red staining 5 weeks after transplantation. Remarkably, only C1D-bi was able to phenocopy the behavior of unsorted parental Comma-1D cells and fill the entire mammary fad pad (Fig. 2F). In contrast, C1D-B partially filled ∼50% of the gland and only minor cell growth was observed in C1D-L transplanted glands (Fig. 2F). As mentioned above, parental Comma-1D cells have been reported to respond to lactogenic stimuli (Danielson et al., 1984; Kittrell et al., 2011). Thus, we also investigated whether C1D-bi reconstituted mammary glands mentioned this potential. As expected based on their ability to phenocopy the parental behavior, we observed that the outgrowths generated by C1D-bi cells generated milk-producing epithelium in transplanted animals during pregnancy (Fig. S2C,D).

Overall, all the above shows that Comma-1D populations present cellular, molecular and functional heterogeneity. Importantly, these results reveal that the heterogeneity of Comma-1D cultures originates in a small fraction of cells with the ability to generate both luminal and basal committed cells.

Cellular and molecular dissection of bipotent Comma-1D

To investigate the characteristics of the C1D-bi cell more deeply, we first used FACS to obtain a purified population of $Epcam^{High}$/$Sca1^{High}$ cells and, then we performed scRNA-seq. This analysis revealed that these cells can be further divided into three clusters (Fig. 3A). Expression profiling of these clusters revealed that two of them are very similar to C1D-B and C1D-L cells indicating some degree of lineage specialization (Fig. 3A). Remarkably, diffusion and pseudotime analysis indicated that the last cluster was the origin of the other two (Fig. 3B; Fig. S3). Based on the above, for simplification, we will call these C1D-bi subpopulations C1D-bi luminal (C1D-biL), C1D-bi basal (C1D-biB) and C1D-bi-stem, respectively. Next, we aimed to test the ability of these cells to regenerate heterogeneous Comma-1D cultures. For this, we first designed a FACS strategy to sort individual subpopulations. Based on the expression data from the sc-RNAseq studies we identified podoplanin as a cell surface marker that was differentially expressed in C1D-bi clusters (Fig. 3C). Additionally, we also observed that C1D-biL expressed the highest levels of Epcam (Fig. 3C). Thus, C1D-biL was $Epcam^{Highest}$/$Sca1^{High}$/$Podoplanin^{Neg}$, C1D-biB was $Epcam^{High}$/$Sca1^{High}$/$Podoplanin^{High}$ and C1D-bi-stem was $Epcam^{High}$/$Sca1^{High}$/$Podoplanin^{low}$ (Fig. 3D). After sorting using these markers, C1D-bi subpopulations were individually cultured, and interestingly, all of them were able to regenerate the heterogeneity of parental Comma-1D cultures (Fig. 3D). However, some differences were noticed. After 1 week, cultures from C1D-biL and C1D-biB contained larger numbers of C1D-L and C1D-B cells, respectively, and additional passes were required until the percentages for other subpopulations increased. This suggests that although bipotency is preserved in all C1D-bi subpopulations, some degree of lineage specialization, luminal or basal, is already present. Remarkably, this allowed us to investigate the molecular determinants that define early lineage specification of Comma-1D cells by comparing their expression profiles. For this, we decided to specifically look at the transcription factors (TFs) that increased their activity when C1D-bi-stem cells progress toward luminal and basal expression patterns.

To identify the key TF that regulates the differentiation process we used a powerful method that analyzes expression profiles called inference of transcriptional regulatory networks (TRNs) (Ding et al., 2018; Alvarez et al., 2016; Putcha et al., 2015; Aytes et al., 2014; Bisikirska et al., 2013; Carro et al., 2010). Briefly, a TRN consists of a web of connections (network) between 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. Then, the cell network is inferred by integrating all the TF–gene hubs. To delineate TRNs, we used the SJARACNe (Khatamian et al., 2019) algorithm [an improved implementation of ARACNe (Margolin et al., 2006), which has been used extensively in normal and cancer cells (Ding et al., 2018; Alvarez et al., 2016; Putcha et al., 2015; Aytes et al., 2014; Bisikirska et al., 2013; Carro et al., 2010)]. Then, we used the C1D-bi-specific TRN to identify the master regulators (MRs) of each C1D-bi subpopulation. A MR is defined as a TF that is differentially active among the populations studied (Aytes et al., 2014; Bisikirska et al., 2013; Carro et al., 2010). For this, we used NetBID (Wang et al., 2019) to calculate the activity scores of each TF in each C1D-bi subpopulation. Finally, to quantify the significance of the activity score we performed gene set enrichment analysis (GSEA) to assess the enrichment of its predicted targets. Notably, multiple TFs for which a relevant role for basal (e.g. *Trp63, Sox2, Twist1, Twist2* or *Klf5*) and luminal epithelial cell differentiation (e.g. *Gata3*, *Klf2, Klf6, Klf7, Elf5* or *Sox9*) has been discovered *in vivo* (Inman et al., 2015; Zhou et al., 2019; Pellacani et al., 2019) were found in our analysis (Fig. 4A; Fig. S4, Table S2).

To further confirm the critical role of these MRs in Comma-1D cells, we performed gain-of-function studies in TFs with a well-known role in the differentiation of mammary epithelial cells, namely, *Trp63* (basal) and *Gata3* (luminal). First, we purified C1D-bi by FACS as described above, and then we overexpressed the TFs in these cells by viral transduction of cDNA (Fig. 4B). For *Trp63*, we overexpressed the DeltaNp63 isoform as it has been shown to play a critical role in epithelial development and differentiation (Romano et al., 2012). Finally, we analyzed how the overexpression of these TFs impacted the regeneration of heterogeneous Comma-1D cultures using flow cytometry. As expected, overexpressing the basal MR *Trp63* increased the fraction of C1D-B cells in the culture (Fig. 4C). In contrast, overexpressing the luminal MR *Gata3* greatly increased the percentage of C1D-L cells (Fig. 4C).

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

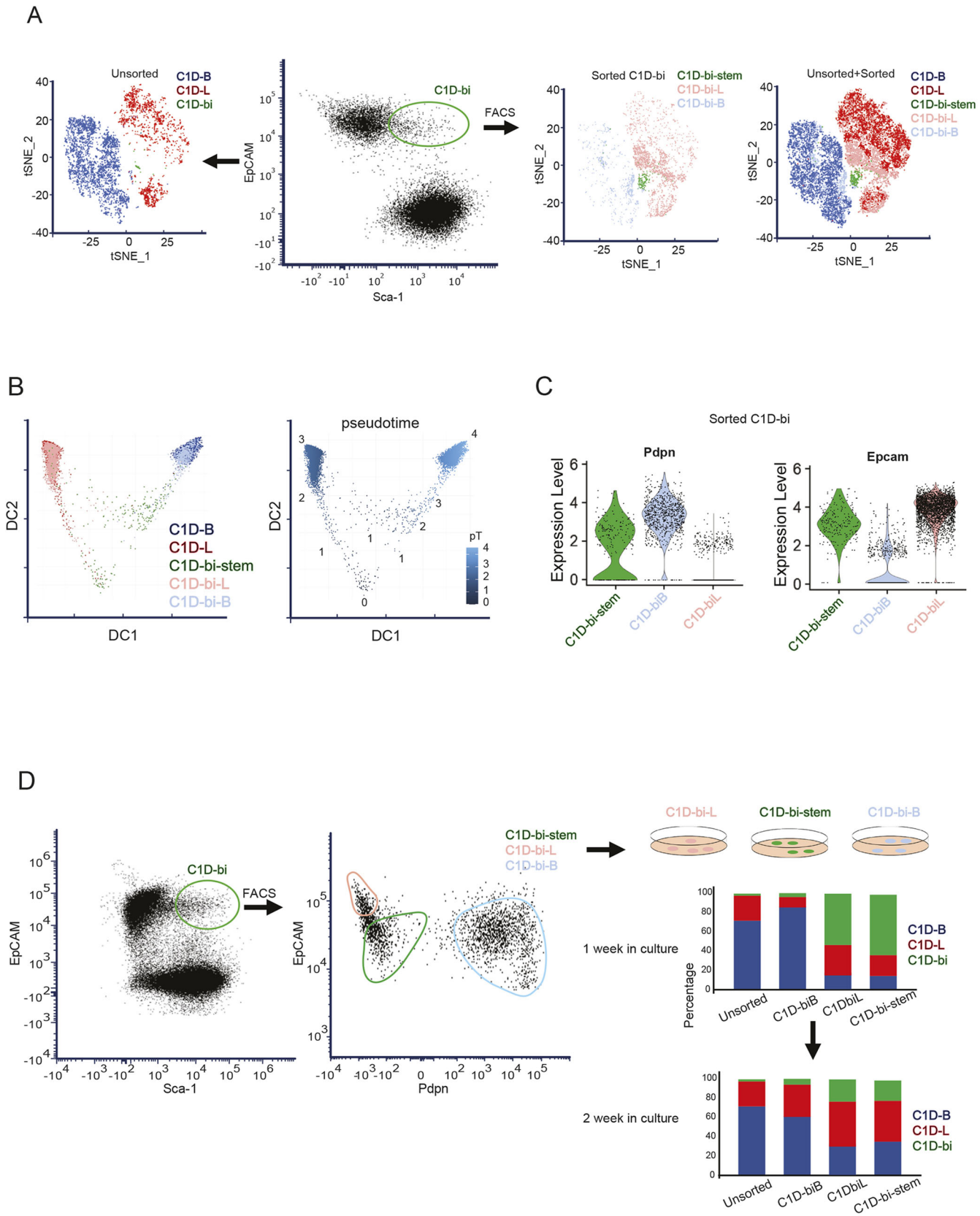

Fig. 3. Single-cell and functional analysis of Comma-1D-bi cells. (A) The tSNE plot shows that FACS purified Comma-1D-bi cells can be subdivided into three subpopulations. (B) Diffusion (left) and pseudotime (right) plots indicating the lineage relationship and divergent trajectory from C1D-bi-stem to the rest of the subpopulations. Results shown in A, B are representative of $n \geq 3$ experiments. (C) Expression level of *Pdpn* and *Epcam* in the Comma-1D-bi subpopulations. Results are given as a violin plot, with individual data points shown. (D) Reconstitution of parental cultures from FACS-purified Comma-1D-bi subpopulations as revealed by determining the percentage of cells belonging to each population. Results are representative of $n \geq 3$ experiments.

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

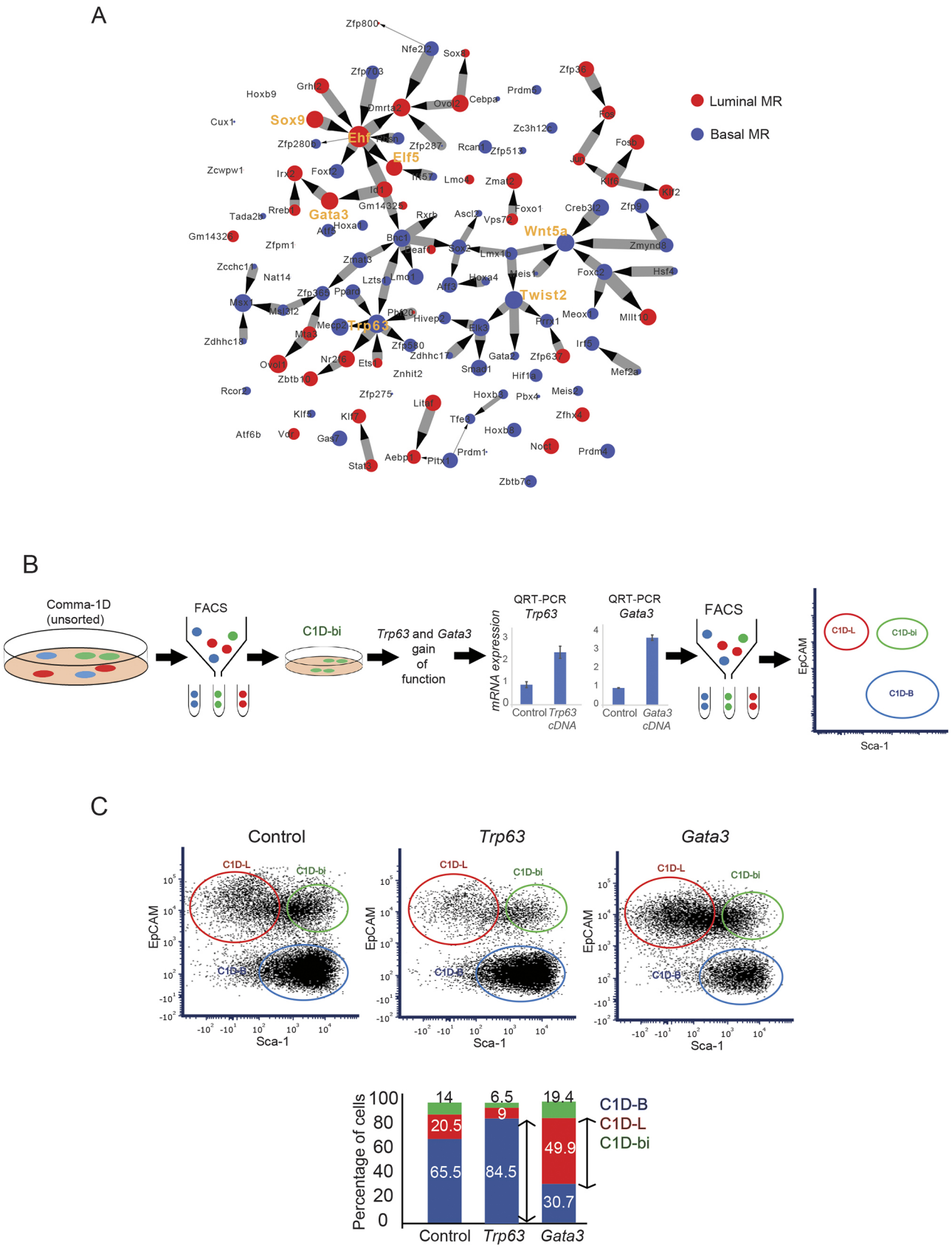

Fig. 4. Molecular characterization of Comma-1D populations. (A) The figure shows network connections of luminal and basal transcription factors in Comma-1D subpopulations. (B) The figure shows the strategy to generate and analyze Comma-1D populations overexpressing *Trp63* and *Gata3*. The qRT-PCR data in the schematic representation shows real results (mean±s.d.; n=3). (C) The FACS profiles show how overexpression of T*rp63* and *Gata3* affect the reconstitution pattern of sorted C1D-bi cells as revealed by determining the percentage of cells belonging to each population. Results are representative of $n \geq 3$. FACS profiles were collected 3 weeks after viral transduction of C1D-bi cells with the corresponding c-DNAs.

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

Overall, these results confirm that important factors and regulatory networks that govern the differentiation process of mammary epithelial cells *in vivo* function, at least in part, in Comma-1D cells.

Comma-1D populations present different transformation sensitivities

Increasing evidence is showing that not all cells are equally sensitive to transformation (Latil et al., 2017; Visvader, 2011). Thus, neither the same genetic alterations can transform all cells nor will cells that get transformed with a particular set of alterations necessarily get transformed with another (Visvader, 2011; Gilbertson, 2011; Skibinski and Kuperwasser, 2015). This has important clinical consequences and generates diverse pathological subtypes even among cells within the same tissue of origin. Large sequencing cancer projects such as BRCA-TCGA (Banerji et al., 2012) and METABRIC (Curtis et al., 2012) have revealed a comprehensive list of genetic alterations that are associated with each of the breast cancer subtypes. However, which the cell of origin is for each of these subtypes is still an open question (Zhang et al., 2017).

Comma-1D are immortalized but not transformed cells. They harbor two distinct *Tp53* mutations that result in loss of its function (Jerry et al., 1994). Thus, we decided to investigate how the three main Comma-1D populations respond to the bona fide breast cancer oncoprotein HER2. For this, we sorted C1D-B, C1D-L, and C-1Dbi by FACS as described above. Sorted populations were then transduced with virus-containing constructs co-expressing oncogenic HER2 ($HER2^{mut}$; a mutant hyperactivated formed of HER2, G776YVMA; Wang et al., 2006) and GFP. At 1 week after transduction, the brightest 25% of the GFP-expressing cells were sorted to obtain homogeneous populations expressing similar levels of the oncogenes (Fig. 5A). Finally, oncogene-expressing Comma-1D variants and control counterparts were compared through several assays to address how the transforming abilities of $HER2^{mut}$ depended on the recipient population.

First, we tested whether the expression of oncogenic HER2 modified the phenotypes of the Comma-1D populations in Matrigel. Although all $HER2^{mut}$ organoids presented overgrowth phenotypes after 10 days in culture, we observed clear differences among the Comma-1D populations despite them expressing comparable amounts of $HER2^{mut}$ (Fig. 5B). $HER2^{mut}$ expressing C1D-B organoids were still of regular shape reaching a maximum of 3 to 4 times the size of controls. In contrast, C1D-L organoids were massive irregular structures ∼10 times larger than their control counterparts. Intermediate phenotypes were found in C1D-bi.

One of the hallmarks of epithelial transformation is the ability of these cells to grow in attachment-independent conditions. Normal epithelial cells require attachment to grow in cell culture and they will die by apoptotic cell death termed anoikis if the attachment is prevented (Taddei et al., 2012). To investigate this feature, we plated Comma-1D variants in a semisolid agar culture system that prevents attachment (Fig. 5C). As expected, none of the control variants generated any colonies after 1 month in culture. Remarkably, only C1D-L cells transformed with HER2 were able to efficiently generate colonies in agar (Fig. 5C).

When transplanted into syngeneic mice, parental comma-1D cells transformed with HER2 form mammary tumors (Xiang et al., 2008). Thus, we wondered whether C1D-L recapitulates this phenotype. Not surprisingly, overexpressing HER^{mut} in unsorted parental cells and C1D-L generated tumors compatible with poorly differentiated mammary carcinomas and that were indistinguishable from each other (Fig. S5A).

To investigate how HER2 impacts the different Comma-1D populations, first we compared the expression profiles of HER2-expressing cells with their corresponding control counterparts. GSEA revealed that HER2 induces multiple signatures associated with activation of canonical mitogen-activated protein kinase signaling, increased glycolysis, and enhanced migration and epithelial–mesenchymal transition specifically in C1D-L (Fig. 5D). Next, we compared the expression profiles of the different Comma-1D populations with the expression profiles of human breast cancers. For this, we use PAM50 signatures to identify the most likely molecular human counterpart for each of the parental Comma-1D populations and their corresponding $HER2^{mut}$-expressing variants. Remarkably, whereas parental Comma-1D populations were found to be related to normal and basal molecular subtypes of human breast cancer overexpression of oncogenic HER2 induced a massive shift towards the HER2+ molecular subtype exclusively in C1D-L (Fig. 5E). Numerous murine models of breast cancer have been created to mimic the genetic aberrations found in human tumors. In particular, gene expression profiles of 385 tumors representative of 27 different genetically engineered mouse models (GEMMs) of breast cancer have been described and compared with human counterparts (Pfefferle et al., 2013). Thus, we also used murine mammary tumor profiles representing well-defined human breast cancer molecular subtypes to further investigate how HER2 impacts the different Comma-1D populations. Consistent with our previous results, we found that overexpression of HER2 in C1D-L cells upregulates a gene signature associated with HER2-enriched murine tumors (Fig. 5F), whereas only a minor or no effect is seen in C1D-B or C1D-bi cells (Fig. S5B).

Overall, these results confirm that not all Comma-1D cells respond equally to oncogenic stimuli and that the C1D-L population is the most sensitive to transformation with HER2.

DISCUSSION

Experimental models are essential for the laboratory-based investigation of those processes that occur in nature. For this, the ability to grow tissue-derived cells *in vitro* is one of the most useful methodologies. These represent tractable systems that are much easier to handle than animal models. The overwhelming majority of all cell cultures are a clonal expansion of a small number of cells that, although can be cultured indefinitely and are easy to manipulate, do not capture the cellular heterogeneity that occurs *in vivo*. This is a major limitation when investigating processes such as lineage commitment and differentiation. Cultures of primary cells extracted fresh from tissues provide an alternative to obtaining heterogeneous cultures. However, these cultures are unstable and finite due to cellular mechanisms like senescence that limit the time for which they can be maintained (Herranz and Gil, 2018). In this context, Comma-1D is a unique mouse mammary epithelial cell line. These cells were spontaneously immortalized and contain a heterogeneous mix of cells with diverse molecular and functional lineage features (Danielson et al., 1984; Kittrell et al., 2011; Chen et al., 2007). Importantly, this includes the presence of a multipotent group of cells that is preserved during culture (Kittrell et al., 2011; Ibarra et al., 2007). Thus, Comma-1D presents intrinsic characteristics that make it a unique model to study multipotency and lineage differentiation *in vitro*.

Here, we report that single-cell analysis of Comma-1D cells reveals how the heterogeneity of this cell line is maintained for a small population of bipotent cells that can generate populations with specific luminal and basal molecular characteristics. Previous studies

Journal of Cell Science (2022) 135, jcs259329. doi:10.1242/jcs.259329

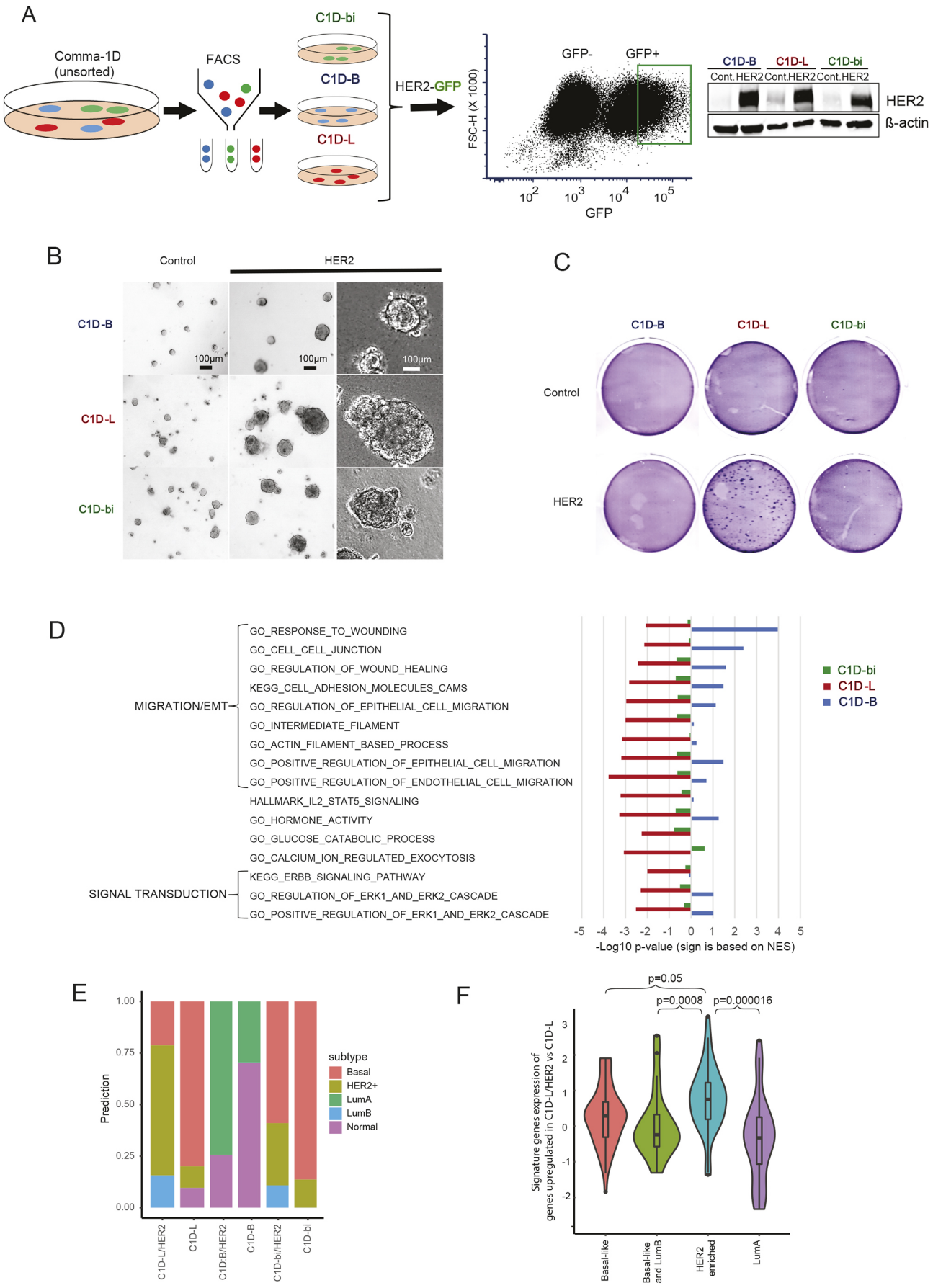

Fig. 5. See next page for legend.

Fig. 5. Impact of oncogenic HER2 in Comma-1D cell populations. (A) The figure shows the strategy to generate Comma-1D populations expressing equivalent levels of an oncogenic HER2 protein. Morphology in 3D Matrigel culture (B) and Colony formation in semisolid agar (C) of parental Comma-1D populations and the corresponding HER2 expressing variants. (C) of cells HER Comma-1D populations. Results shown in A–C are representative of $n \geq 3$ experiments. (D) Gene set enrichment analysis comparing parental Comma-1D populations versus the corresponding HER2 expressing variant. The bar graph indicates significant *P*-values as well as the direction of the changes. NES, Normalized enrichment score. (E) Prediction of the corresponding breast cancer molecular subtype of parental and HER2 overexpressing Comma-1D populations using PAM50 signatures. (F) Upregulation of genes associated with different molecular subtypes of mammary tumors from mouse transgenic models in C1D-L overexpressing oncogenic HER2. Results are given as a violin plot with a box-and-whisker plot. The box in the violin plot shows the median, the 1st and the 3rd quartile. The whiskers are drawn down to the 10th percentile and up to the 90th. Points below and above the whiskers are drawn as individual dots. (n=3).

have indicated that the expression of Sca-1 membrane protein can be used to capture a population of Comma-1D cells enriched in multipotent cells. Because a large number of Comma-1D cells that express high levels of Sca-1 also express markers of basal epithelium, it was assumed that the bipotent population was of basal origin. Remarkably, our results reveal a different picture. Our studies confirm that bipotent Comma-1D cells express high levels of Sca-1 but in contrast to what was originally thought, these cells also express luminal markers presenting an intermediate phenotype. Molecular characterization of the three main Comma-1D populations revealed multiple TFs that are well known to be key regulators of the differentiation process *in vivo*. Importantly, we have demonstrated that these master regulators also exert similar functions in Comma-1D cells *in vitro* and that the experimental manipulation of these factors impacts the generation of the lineage-committed Comma-1D populations C1D-L and C1D-B. This is important as it indicates that the differentiation of Comma-1D cells *in vitro* conserves a degree of molecular similarities with the processes described *in vivo*. Thus, Comma-1D cells represent an easy-to-use model that provides a source of unlimited material to study the differentiation of the mammary epithelium *in vitro*. This will be especially relevant for studies that require a significant amount of starting material such as biochemical or metabolomic studies.

Another important finding from our data is that not all Comma-1D cells respond the same way to oncogenic stimuli. As mentioned above, tumorigenesis is the result of a perfect storm where specific alterations in tumor suppressors and oncogenes accumulate in specific cells (Visvader, 2011; Gilbertson, 2011; Skibinski and Kuperwasser, 2015). A hierarchy exists in the mammary epithelium where embryonic bipotent mammary epithelial stem cells (MaSCs) generate unipotent luminal and basal progenitors that are responsible for the generation and maintenance of the various differentiated cells that form the mammary epithelial tree (Skibinski and Kuperwasser, 2015). However, understanding how normal mammary gland heterogeneity influences breast tumor heterogeneity is poorly understood. Because of its multipotent characteristics, it was initially postulated that MaSCs could be the 'cell-of-origin' for breast cancers (Li et al., 2003). However, emerging data is showing a different picture. Thus, for instance, conditional deletion of BRCA1 in luminal precursor cells generates tumors that do not present a luminal phenotype, instead, they resemble the basal-like phenotype found in $BRCA^{mut}$ carriers in human cancers (Molyneux et al., 2010). Similarly, because of its complex heterogeneity, the origin of HER2+ tumors remains challenging to resolve (Godoy-Ortiz et al., 2019). Interestingly, in our studies, oncogenic HER2 was only able to promote transformed characteristics in C1D-L cells (see Fig. 5C). Noticeably, breast cancers emerging in the commonly used HER2 mouse model MMTV-neu (Guy et al., 1992) present an expression pattern similar to the luminal human breast cancer (Pfefferle et al., 2013).

Overall, these results support that the heterogeneity of Comma-1D cells can be used to study how the response to cancer-promoting mutations is influenced by the cellular fingerprint of cell subtypes and how different alterations impact the tumor phenotype.

An additional area of intrigue is how the constant equilibrium between the Comma-1D populations is preserved in culture. Our data have shown that C1D-L and C1D-B can grow as individual subpopulations (Fig. 2E) and that C1D-B are the fastest-growing cells (Fig. S2). Thus, in standard conditions, C1D-B cells would take over the entire culture over time. Additionally, due to the small percentage of C1D-bi in Comma-1D cultures, these cells would be progressively diluted even if the growth rate were identical to C1D-B. One possibility to maintain the heterogeneity is that some kind of dedifferentiation (Jopling et al., 2011) naturally occurs. We have observed that some of our sorted C1D-L cultures generate a small number of C1D-B cells. However, the number of cells formed was very small (<2%), inconsistent among purifications, and we were able to culture pure C1D-L cultures for long periods (>1 month) (data not shown). Thus, it is unclear whether these cells are sporadic contaminant C1D-B cells or are generated under certain conditions from C1D-L. Another possibility is that when all Comma-1D populations are together they interact with each other to maintain the equilibrium. Heterotypic communication between luminal and basal cells is known to play important roles in mammary epithelial homeostasis *in vivo* (Centonze et al., 2020). Interestingly, the expression profiles of Comma-1D cells reveal strong differential expression of some ligands and secreted molecules. For instance, CSF3 and calprotectin (an S100A8–S100A9 complex) were highly expressed in C1D-L cells, whereas multiple insulin-like growth factor-binding proteins (IGFBP-2, -3, -4, -6 and -7) were expressed in C1D-B cells. Although investigating this question is out of the scope of this manuscript, additional studies may set up the basis to use the Comma-1D model to investigate heterotypic communication between cells.

Finally, there are limitations in the Comma-1D model that are worth discussing. The data shown here illustrates how the heterogeneity of the Comma-1D mimics, to some extent, what is observed *in vivo*. However, not all processes or pathways seen *in vivo* are present in this culture model. For instance, these cells do not express estrogen or progesterone receptors (Stingl, 2011) or the *Foxa1* (Theodorou et al., 2013) TF, which are critical regulators of mammary epithelial homeostasis.

In summary, our single-cell studies of Comma-1D cells provide a molecular and functional frame to utilize the unique heterogeneity of this model as a tractable and easy-to-use platform for mammary epithelial differentiation and tumorigenesis.

MATERIALS AND METHODS

Cell culture

The Comma1-D cell line was obtained from Dr Greg Hannon (Cambridge Institute, UK) as an early passage of the original line generated by Dr Medina (Danielson et al., 1984). Comma1-D cells were grown in Dulbecco's modified Eagle's medium (DMEM):F12 (Gibco) supplemented with 2% fetal calf serum (FCS), 5 ng/ml murine epidermal growth factor (EGF; Sigma), 10 µg/ml human insulin (Sigma) and 50 µg/ml gentamicin (Gibco). They were used for an average of 10–15 passes before thawing a new batch. HEK-293T Phoenix cells were obtained from the ATCC and maintained in DMEM supplemented with 10% FBS (Corning) and 1% penicillin-streptomycin (Gibco). All cells were tested

for contamination before starting the experiments. RNA-seq-characterized cultures will be shared with the scientific community through collaboration.

Mammary fat pad transplantation

Transplantation of Comma-1D populations and HER2 transformed variants was performed as previously described (Danielson et al., 1984; Xiang et al., 2008). Briefly, 2.5×10^5 cells were surgically injected in 10 µl of PBS in cleared fat pad (mammary sprout located between the nipple and the lymph node was excised) of 21-day-old female syngeneic Balb/c mice. Reconstitution of the normal mammary epithelium was evaluated by standard Carmine Red staining after 5–6 weeks. Evaluation of tumor growth in HER2 transformed variants was performed 6–8 weeks after transplantation by visual inspection and hematoxylin and eosin (H&E) staining (see methods in Lewis and Porter, 2009). For pregnancy studies, transplanted animals were mated at 2 months old and mammary glands were evaluated 1–3 days after delivery of pups. All animal experiments were approved by the institutional IACUC committee.

Western blotting

Cells were lysed in RIPA buffer (50 mM Tris-HCl pH 7.4, 150 mM, NaCl, 5 mM EDTA, pH 8.0, 30 mM NaF, 1 mM Na_3VO_4, 40 mM β-glycerophosphate, protease inhibitors, 10% glycerol and 1% Nonidet-P40). Protein concentrations were determined by using the Protein Assay Kit (Bio-Rad #500-0006). Equal amounts of proteins were subjected to SDS-PAGE and transferred to nitrocellulose membranes (GE Healthcare #10401197). Non-specific binding was blocked by incubation with 5% non-fat milk in phosphate-buffered saline (Sigma) with 0.05% Tween 20. Membranes were incubated with primary antibodies overnight at 4°C and fluorescent secondary antibodies for 1 h at room temperature. The primary antibodies used were anti-actin (1:500; Santa Cruz Biotehcnlogy, sc47778) and anti-ErbB2/c-Neu (1:500; Sigma #OP15). All antibody dilutions were used as recommended by the manufacturer. The unprocessed blot corresponding to Fig. 5 is shown in Fig. S6.

RT-PCR

Total RNA was extracted using Qaigen RNAeasy Mini Kit (#74104) according to the manufacturer's instructions. First-strand cDNA was generated from total RNA for each sample using the Roche Transcriptor First Strand cDNA Synthesis Kit (04 379 012 001) according to the manufacturer's instructions. cDNA was then amplified in triplicate for each sample with using PowerTrack SYBR Green Master Mix (Thermo #A46012). PCRs consisted of an initial denaturation step (3 min at 95°C) and 40 cycles of PCR (95°C for 10 s, and 59.5°C for 45 s). Relative expression was calculated by normalizing cycle threshold to that of *Gapdh*. The following primers were utilized (specific to mouse): forward hTrp63, 5′-GAGCAGCCTTGACCAGTCTC-3′; reverse hTrp63, 5′-GAGGAG-CCGTTCTGAATCTG-3′; forward mGapdh, 5′-AAGGGCTCATGACCA-CAGTC-3′; reverse mGapdh, 5′-GGATGCAGGGATGATGATGTTCT-3′; forward mGATA3, 5′-CGAATTCCGCATGGAGGTGACTG-3′; and reverse mGATA3, 5′-GACGGAGTTTCCGTAGTAGGACG-3′.

Comma1-D viral transduction

For ectopic HER2 expression, pBabe Her2 mutant GFP plasmid was used (a modified version of Addgene #40982). Virus production was achieved by transfecting 293 phoenix eco cells with jetPEI (Polyplus #101-10N). Medium was collected at two 24-h time points and concentrated overnight using virus Concentrator (Clontech #631231) and resuspended in Comma-1D medium. Comma1-D cells were infected for 24 h at which point the medium was changed. After 72 h, GFP-positive cells were selected via FACS (FACSAria II BD cell sorter). For *Gata3* overexpression, we utilized the LZRS-GATA3 (Addgene #34836) viral vector. For *Trp63* overexpression we subcloned ΔNp63beta (Addgene #27014) in a retroviral vector (pLPCX-GFP, Addgene #65433) using the following primers designed to add XhoI forward and BamHI reverse cut sites: forward, 5′-ATCGATCGCTCGAGATGGGCTGTGATCG-3′; reverse, 5′-ATCGATCGGGATCCTTATTTTCCATTCTTGGA-3′.

Flow cytometry and FACS

Cell suspensions were incubated with the following antibodies: EpCAM-PerCP/Cy5.5, CD49f-APC, Sca-1-BV421, Podoplanin- APC/Cy7 (BioLegend #118220, #313616, #108127, #127418; #118220 and #108127 were used at 0.15 µg per 10^6 cells; #313616 was used at 5 µl per 10^6 cells; #127418 was used at 0.05 µg per 10^6 cells). The LIVE/DEAD™ Fixable Aqua Dead Cell Stain (Thermo Fisher #L34957) was used to determine cell viability. Antibody incubations were performed for 15 min on ice. The cells were sorted using a FACSAria II (BD) cell sorter or analyzed with an Attune NxT Flow Cytometer. The gating strategy was the following: size discrimination by FSC-A/SSC-A followed by doublet discrimination by FSC-W/FSC-H. Singlets were assessed for viability and the viable cells were resolved into luminal, basal, and stem populations based on EpCAM/Sca-1.

Organoid culture

Sorted cells were seeded in 24-well ultra-low attachment plates at a density of 5000 or 10,000 cells/well in EpiCult-B Mouse Medium (StemCell #05610) supplemented with 5% FBS, 10 ng/ml EGF, 20 ng/ml basic fibroblast growth factor (bFGF), 4 µg/ml heparin, 5 µM Y-27632 and 5% Matrigel (Corning #354230). The organoid culture method was adapted from Guo et al. (2012). The result indicates the combination of three different experimental replicates with >100 spheroids counted for each population in each of them. Organoid dissociation was achieved by incubating organoids in 0.25% trypsin-EDTA in PBS for 5 min at 37°C followed by mechanical disruption with a P1000 pipette. Finally, the dissociated cells were filtered using a 40 µm cell strainer.

Colony-forming assay

Cells were seeded in triplicate (300 cells/well) in six-well tissue culture plates (Falcon) and grown without passaging for 21 days. Medium was aspirated and cells stained with 0.1% Crystal Violet solution for 1 h at room temperature with subsequent washes to remove the excess of stain before colonies were counted manually.

Agar assay

Cells were seeded in triplicate in six-well ultra-low attachment plates (Falcon) with each well containing a cell suspension of 5000 cells in 1.8 ml medium combined with 400 µl 2% melted sterile agar (w/v). After solidification, cells were maintained at 37°C. Medium was replaced twice a week. After 14 days, cells were stained with 0.01% Crystal Violet and the colonies were counted manually.

Growth curve assay

Cells were seeded in triplicate (2000/well) into a 96-well clear bottom tissue culture plate (Falcon) and Cell Titer Glo (Promega #G7570) applied for every 24 h-read as per the manufacturer's instructions.

Library preparation and single-cell sequencing

Library preparation was performed with Chromium Single Cell Gene Expression 3′ Kits (10x Genomics) and sequenced on NextSeq 2000 systems (Illumina). Read processing including demultiplexing, barcode assignment, and unique molecular identifier (UMI) quantification was performed by the Cell Ranger analysis pipeline. The reference genome mm10 was used. Quality control and data pre-processing were performed by the R package Seurat. In short, genes expressed in less than 1% of cells were filtered. A couple of criteria, including the number of genes detected, the total number of UMIs, and the percentage of molecules mapped to the mitochondrial genes, were used to determine the quality of cells. Specifically, cells in which either the number of genes or the total number of UMIs was less than three median absolute deviations (MAD) below the median of the distribution were filtered. The same for cells with a high percentage of molecules mapped to the mitochondrial genes (above the median by three MADs). The resultant gene-cell matrices were normalized such that the total expression of each cell was scaled by a factor of 100,000.

Luminal and basal markers and scores

Gene expression in luminal and basal cells was obtained by single-cell RNA sequencing of the mammary gland of 3 month aged mice, generated by the

Tabula Muris Consortium. scRNA-seq and bulk profiling can be found in Gene expression omnibus (GEO) under accession #GSE182354. Informative genes for basal and luminal clusters were obtained in the pre-built training model of the tool SciBet (Li et al., 2020), which were identified by SciBet using their E-test. The top 100 luminal genes and the top 100 basal genes were selected, and the overlapping genes were filtered, arriving at 34 basal markers such as *Krt14* and *Krt5*, and 34 luminal markers such as *Krt18* and *Krt8*. Given the expression profile of a cell, the basal (luminal) scores were defined as the average expression of the basal (luminal) markers in the cell.

Clustering, diffusion mapping and pseudotime analysis of scRNA-seq data

Clustering was performed using an in-house mutual information-based clustering algorithm, scMINER (https://github.com/jyyulab/). Briefly, for each pair of cells, the pairwise mutual information was calculated based on the log-transformed expression profiles and was used as the distance between the cells. The mutual information-based metric was better in capturing the non-linearity between expression profiles. Multi-dimensional scaling was performed for dimension reduction, and the first 19 eigenvectors were used for further clustering. For robustness, consensus clustering was performed over ten independent k-means clustering. The visualization of clusters on tSNE plots in Figs 1A and 3A were generated using the tool Seurat (Satija et al., 2015). In Fig. 1A, only unsorted cells were used, whereas for Fig. 3A, both unsorted and sorted datasets were combined using the scRNA-seq integration pipeline described in the tool Seurat. The pipeline matches and align shared cell populations across datasets and correct for batch effects by identifying a set of anchors. Marker genes were used for creating the diffusion map and pseudo-time analysis. The diffusion map and the inference of pseudo-time were obtained by using the R package destiny (Angerer et al., 2016).

Inference of transcriptional regulatory networks

The cell-type-specific transcriptional regulatory network of C1D-bi was inferred from single-cell expression data using SJARACNe (Khatamian et al., 2019). Like the original mutual-information-based ARACNe algorithm, it eliminates the indirect interactions typically found by correlation-based analysis. The adaptive partitioning algorithm was used to estimate mutual information and bootstrapping was applied to ensure robustness. NetBID2 (https://github.com/jyyulab/NetBID; Du et al., 2018) was then used to calculate the activity of transcription factors. In short, the activity value summarizes the importance of the transcription factor. A high activity score means the targets of the transcription factors are highly expressed. TFs with high activity in either C1D-biL or C1D-biB were identified as the lineage-specific master regulators. The differential activity was estimated by a two-sample *t*-test. The network shown in Fig. 4A was constructed using HDMap (https://github.com/jyyulab/hdmap), and visualization by Cytoscape (Shannon et al., 2003). The size of nodes and edges scales with the betweenness centrality.

RNA-seq of purified Comma-1D populations, populations transduced with oncogenes and control counterparts

The Comma-1D populations were sorted and transduced with oncogenic HER2, and control counterparts (viruses carrying luciferase vectors). Stranded RNA libraries were prepared with 200 ng total RNA using the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche). RNA library sequencing was performed on Illumina NovaSeq 6000 system (Illumina). Expression quantification was performed at a gene level by Salmon (Patro et al., 2017). scRNA-seq and bulk profiling can be found in Gene expression omnibus (GEO) under accession #GSE182354.

Differential expression for each of the three HER2-transformed cell populations and their corresponding controls was estimated using the tool GFOLD (Feng et al., 2012). The enrichment of upregulated and downregulated genes among the MSigDB gene sets (v7.1) (Liberzon et al., 2015) were examined using the tool fgsea (https://github.com/ctlab/fgsea). Gene sets enriched with the up- and down-regulated genes will receive a positive and negative, respectively, normalized enrichment score (NES) with significant *P*-values.

The top 50 upregulated genes for each of the three HER2 transformed cell populations when compared to the corresponding controls were used as the signatures of the three populations. To confirm the resemblance of C1D-L with mouse and human HER2 tumors, the expression levels of the signature genes were examined in patient samples from the METABRIC dataset (Curtis et al., 2012) and samples of genetically engineered mouse models of breast cancer (Pfefferle et al., 2013). Prediction of the PAM50 molecular subtyping of the three HER2 transformed cell populations were performed using the R package *genefu* (Gendoo et al., 2016).

Acknowledgements

Comma-1D cells were obtained from Dr Greg Hannon. We would like to thank Dr Partha Mukhopadhyay and Zayd Rashid for their assistance with the fat pad transplantation experiments.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: R.L.W., E.A.N., P.J.Y., J.S.; Methodology: R.L.W., E.A.N., K.-K.Y., B.J., B.S., J.E., P.J.Y., J.S.; Software: K.-K.Y., P.J.Y.; Validation: R.L.W., E.A.N., K.-K.Y., P.J.Y., J.S.; Formal analysis: R.L.W., E.A.N., K.-K.Y., B.J., B.S., J.E., P.J.Y., J.S.; Investigation: R.L.W., E.A.N., K.-K.Y., P.J.Y., J.S.; Resources: B.J., B.S., J.E., P.J.Y., J.S.; Data curation: R.L.W., E.A.N., K.-K.Y., P.J.Y., J.S.; Writing - original draft: J.S.; Writing - review & editing: R.L.W., K.-K.Y., P.J.Y., J.S.; Visualization: R.L.W., E.A.N., K.-K.Y., P.J.Y., J.S.; Supervision: P.J.Y., J.S.; Project administration: J.S.; Funding acquisition: P.J.Y., J.S.

Funding

This work was in part supported by the National Institutes of Health (R01GM134382 to P.J.Y.) and by the American Lebanese Syrian Associated Charities. The views expressed in this article reflect the results of research conducted by the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government. Deposited in PMC for release after 12 months.

Data availability

scRNA-seq and bulk profiling can be found in Gene Expression Omnibus (GEO) under accession #GSE182354.

Peer review history

The peer review history is available online at https://journals.biologists.com/jcs/article-lookup/doi/10.1242/jcs.259329.

References

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H. and Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838-847. doi:10.1038/ng.3593

Anderson, L. W., Danielson, K. G. and Hosick, H. L. (1979). Epithelial cell line and subline established from premalignant mouse mammary tissue. *In Vitro* **15**, 841-843. doi:10.1007/BF02618037

Angerer, P., Haghverdi, L., Büttner, M., Theis, F. J., Marr, C. and Buettner, F. (2016). Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241-1243. doi:10.1093/bioinformatics/btv715

Aytes, A., Mitrofanova, A., Lefebvre, C., Alvarez, M. J., Castillo-Martin, M., Zheng, T., Eastham, J. A., Gopalan, A., Pienta, K. J., Shen, M. M. et al. (2014). 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

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., Lawrence, M. S., Sivachenko, A. Y., Sougnez, C., Zou, L., et al. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-409. doi:10.1038/nature11154

Bisikirska, B. C., Adam, S. J., Alvarez, M. J., Rajbhandari, P., Cox, R., Lefebvre, C., Wang, K., Rieckhof, G. E., Felsher, D. W. and Califano, A. (2013). 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

Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., Anne, S. L., Doetsch, F., Colman, H. et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318-325. doi:10.1038/nature08712

Centonze, A., Lin, S., Tika, E., Sifrim, A., Fioramonti, M., Malfait, M., Song, Y., Wuidart, A., Van Herck, J., Dannau, A. et al. (2020). Heterotypic cell-cell

communication regulates glandular stem cell multipotency. *Nature* **584**, 608-613. doi:10.1038/s41586-020-2632-y

Chen, M. S., Woodward, W. A., Behbod, F., Peddibhotla, S., Alfaro, M. P., Buchholz, T. A. and Rosen, J. M. (2007). Wnt/beta-catenin mediates radiation resistance of Sca1+ progenitors in an immortalized mammary gland cell line. *J. Cell Sci.* **120**, 468-477. doi:10.1242/jcs.03348

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352. doi:10.1038/nature10983

Danielson, K. G., Oborn, C. J., Durban, E. M., Butel, J. S. and Medina, D. (1984). Epithelial mouse mammary cell line exhibiting normal morphogenesis in vivo and functional differentiation in vitro. *Proc. Natl. Acad. Sci. USA* **81**, 3756-3760. doi:10.1073/pnas.81.12.3756

Deugnier, M. A., Faraldo, M. M., Teulière, J., Thiery, J. P., Medina, D. and Glukhova, M. A. (2006). Isolation of mouse mammary epithelial progenitor cells with basal characteristics from the Comma-Dbeta cell line. *Dev. Biol.* **293**, 414-425. doi:10.1016/j.ydbio.2006.02.007

Ding, H., Douglass, E. F., Jr, Sonabend, A. M., Mela, A., Bose, S., Gonzalez, C., Canoll, P. D., Sims, P. A., Alvarez, M. J. and Califano, A. (2018). 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

Du, X., Wen, J., Wang, Y., Karmaus, P. W. F., Khatamian, A., Tan, H., Li, Y., Guy, C., Nguyen, T. M., Dhungana, Y. et al. (2018). Hippo/Mst signalling couples metabolic state and immune function of CD8α+ dendritic cells. *Nature* **558**, 141-145. doi:10.1038/s41586-018-0177-0

Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X. and Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* **28**, 2782-2788. doi:10.1093/bioinformatics/bts515

Guo, W., Keckesova, Z., Donaher, J. L., Shibue, T., Tischler, V., Reinhardt, F., Itzkovitz, S., Noske, A., Zürrer-Härdi, U., Bell, G. et al. (2012). Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* **148**, 1015-1028. doi:10.1016/j.cell.2012.02.008

Gendoo, D. M., Ratanasirigulchai, N., Schroder, M. S., Paré, L., Parker, J. S., Prat, A. and Haibe-Kains, B. (2016). Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097-1099. doi:10.1093/bioinformatics/btv693

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., Mcdonald, E. R., 3RD, Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503-508. doi:10.1038/s41586-019-1186-3

Gilbertson, R. J. (2011). Mapping cancer origins. *Cell* **145**, 25-29. doi:10.1016/j.cell.2011.03.019

Godoy-Ortiz, A., Sanchez-Munoz, A., Chica Parrado, M. R., Álvarez, M., Ribelles, N., Rueda Dominguez, A. and Alba, E. (2019). Deciphering HER2 breast cancer disease: biological and clinical implications. *Front Oncol* **9**, 1124. doi:10.3389/fonc.2019.01124

Guy, C. T., Webster, M. A., Schaller, M., Parsons, T. J., Cardiff, R. D. and Muller, W. J. (1992). Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc. Natl. Acad. Sci. USA* **89**, 10578-10582. doi:10.1073/pnas.89.22.10578

Haghverdi, L., Buettner, F. and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989-2998. doi:10.1093/bioinformatics/btv325

Haghverdi, L., Buttner, M., Wolf, F. A., Büettner, F. and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845-848. doi:10.1038/nmeth.3971

Herranz, N. and Gil, J. (2018). Mechanisms and functions of cellular senescence. *J. Clin. Invest.* **128**, 1238-1246. doi:10.1172/JCI95148

Howard, D. K., Schlom, J. and Fisher, P. B. (1983). Chemical carcinogen-mouse mammary tumor virus interactions in cell transformation. *In Vitro* **19**, 58-66. doi:10.1007/BF02617995

Hwang, B., Lee, J. H. and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1-14. doi:10.1038/s12276-018-0071-8

Ibarra, I., Erlich, Y., Muthuswamy, S. K., Sachidanandam, R. and Hannon, G. J. (2007). A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells. *Genes Dev.* **21**, 3238-3243. doi:10.1101/gad.1616307

Inman, J. L., Robertson, C., Mott, J. D. and Bissell, M. J. (2015). Mammary gland development: cell fate specification, stem cells and the microenvironment. *Development* **142**, 1028-1042. doi:10.1242/dev.087643

Ip, M. M. and Asch, B. B. (2000). 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. doi:10.1023/A:1026435103940

Jerry, D. J., Medina, D. and Butel, J. S. (1994). p53 mutations in COMMA-D cells. *In Vitro Cell. Dev. Biol. Anim.* **30**, 87-89. doi:10.1007/BF02631398

Jopling, C., Boue, S. and Izpisua Belmonte, J. C. (2011). Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nat. Rev. Mol. Cell Biol.* **12**, 79-89. doi:10.1038/nrm3043

Khatamian, A., Paull, E. O., Califano, A. and Yu, J. (2019). SJARACNe: a scalable software tool for gene network reverse engineering from big data. *Bioinformatics* **35**, 2165-2166. doi:10.1093/bioinformatics/bty907

Kittrell, F. S., Carletti, M. Z., Kerbawy, S., Heestand, J., Xian, W., Zhang, M., Lamarca, H. L., Sonnenberg, A., Rosen, J. M., Medina, D. et al. (2011). Prospective isolation and characterization of committed and multipotent progenitors from immortalized mouse mammary epithelial cells with morphogenic potential. *Breast Cancer Res.* **13**, R41. doi:10.1186/bcr2863

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. doi:10.1242/dev.133058

Kumar, B., Prasad, M., Bhat-Nakshatri, P., Anjanappa, M., Kalra, M., Marino, N., Storniolo, A. M., Rao, X., Liu, S., Wan, J. et al. (2018). Normal breast-derived epithelial cells with luminal and intrinsic subtype-enriched gene expression document interindividual differences in their differentiation cascade. *Cancer Res.* **78**, 5107-5123. doi:10.1158/0008-5472.CAN-18-0509

Latil, M., Nassar, D., Beck, B., Boumahdi, S., Wang, L., Brisebarre, A., Dubois, C., Nkusi, E., Lenglez, S., Checinska, A. et al. (2017). Cell-type-specific chromatin states differentially prime squamous cell carcinoma tumor-initiating cells for epithelial to mesenchymal transition. *Cell Stem Cell* **20**, 191-204.e5. doi:10.1016/j.stem.2016.10.018

Lee, G. Y., Kenny, P. A., Lee, E. H. and Bissell, M. J. (2007). Three-dimensional culture models of normal and malignant breast epithelial cells. *Nat. Methods* **4**, 359-365. doi:10.1038/nmeth1015

Lewis, M. T and Porter, W W. (2009). Methods in mammary gland biology and breast cancer research: an update. *J. Mammary Gland Biol. Neoplasia* **14**, 365. doi:10.1007/s10911-009-9162-4

Li, Y., Welm, B., Podsypanina, K., Huang, S., Chamorro, M., Zhang, X., Rowlands, T., Egeblad, M., Cowin, P., Werb, Z. et al. (2003). Evidence that transgenes encoding components of the Wnt signaling pathway preferentially induce mammary cancers from progenitor cells. *Proc. Natl. Acad. Sci. USA* **100**, 15853-15858. doi:10.1073/pnas.2136825100

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X. and Zhang, Z. (2020). SciBet as a portable and fast single cell type identifier. *Nat. Commun.* **11**, 1818. doi:10.1038/s41467-020-15523-2

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P. and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst* **1**, 417-425. doi:10.1016/j.cels.2015.12.004

Macias, H. and Hinck, L. (2012). Mammary gland development. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 533-557. doi:10.1002/wdev.35

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* **7** Suppl. 1, S7. doi:10.1186/1471-2105-7-S1-S7

Molyneux, G., Geyer, F. C., Magnay, F. A., Mccarthy, A., Kendrick, H., Natrajan, R., Mackay, A., Grigoriadis, A., Tutt, A., Ashworth, A. et al. (2010). BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell Stem Cell* **7**, 403-417. doi:10.1016/j.stem.2010.07.010

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417-419. doi:10.1038/nmeth.4197

Pellacani, D., Tan, S., Lefort, S. and Eaves, C. J. (2019). Transcriptional regulation of normal human mammary cell heterogeneity and its perturbation in breast cancer. *EMBO J.* **38**, e100330. doi:10.15252/embj.2018100330

Pfefferle, A. D., Herschkowitz, J. I., Usary, J., Harrell, J. C., Spike, B. T., Adams, J. R., Torres-Arzayus, M. I., Brown, M., Egan, S. E., Wahl, G. M. et al. (2013). Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* **14**, R125. doi:10.1186/gb-2013-14-11-r125

Putcha, P., Yu, J., Rodriguez-Barrueco, R., Saucedo-Cuevas, L., Villagrasa, P., Murga-Penas, E., Quayle, S. N., Yang, M., Castro, V., Llobet-Navas, D. et al. (2015). HDAC6 activity is a non-oncogene addiction hub for inflammatory breast cancers. *Breast Cancer Res.* **17**, 149. doi:10.1186/s13058-015-0658-0

Qu, Y., Han, B., Yu, Y., Yao, W., Bose, S., Karlan, B. Y., Giuliano, A. E. and Cui, X. (2015). Evaluation of MCF10A as a reliable model for normal human mammary epithelial cells. *PLoS One* **10**, e0131285. doi:10.1371/journal.pone.0131285

Romano, R. A., Smalley, K., Magraw, C., Serna, V. A., Kurita, T., Raghavan, S. and Sinha, S. (2012). ΔNp63 knockout mice reveal its indispensable role as a master regulator of epithelial development and differentiation. *Development* **139**, 772-782. doi:10.1242/dev.071191

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495-502. doi:10.1038/nbt.3192

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., 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. doi:10.1101/gr.1239303

Skibinski, A. and Kuperwasser, C. (2015). The origin of breast tumor heterogeneity. *Oncogene* **34**, 5309-5316. doi:10.1038/onc.2014.475

Stingl, J. (2011). Estrogen and progesterone in normal mammary gland development and in cancer. *Horm. Cancer* **2**, 85-90. doi:10.1007/s12672-010-0055-1

Tabula Muris, C. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595. doi:10.1038/s41586-020-2496-1

Taddei, M. L., Giannoni, E., Fiaschi, T. and Chiarugi, P. (2012). Anoikis: an emerging hallmark in health and diseases. *J. Pathol.* **226**, 380-393. doi:10.1002/path.3000

Theodorou, V., Stark, R., Menon, S. and Carroll, J. S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res.* **23**, 12-22. doi:10.1101/gr.139469.112

Vaidya, A. B., Lasfargues, E. Y., Sheffield, J. B. and Coutinho, W. G. (1978). Murine mammary tumor virus (MuMTV) infection of an epithelial cell line established from C57BL/6 mouse mammary glands. *Virology* **90**, 12-22. doi:10.1016/0042-6822(78)90328-8

Visvader, J. E. (2011). Cells of origin in cancer. *Nature* **469**, 314-322. doi:10.1038/nature09781

Visvader, J. E. and Lindeman, G. J. (2006). Mammary stem cells and mammopoiesis. *Cancer Res.* **66**, 9798-9801. doi:10.1158/0008-5472.CAN-06-2254

Visvader, J. E. and Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev.* **28**, 1143-1158. doi:10.1101/gad.242511.114

Wang, S. E., Narasanna, A., Perez-Torres, M., Xiang, B., Wu, F. Y., Yang, S., Carpenter, G., Gazdar, A. F., Muthuswamy, S. K. and Arteaga, C. L. (2006). HER2 kinase domain mutation results in constitutive phosphorylation and activation of HER2 and EGFR and resistance to EGFR tyrosine kinase inhibitors. *Cancer Cell* **10**, 25-38. doi:10.1016/j.ccr.2006.05.023

Wang, Y., Du, X., Wei, J., Long, L., Tan, H., Guy, C., Dhungana, Y., Qian, C., Neale, G., Fu, Y. X. et al. (2019). LKB1 orchestrates dendritic cell metabolic quiescence and anti-tumor immunity. *Cell Res.* **29**, 391-405. doi:10.1038/s41422-019-0157-4

Xiang, B., Chatti, K., Qiu, H., Lakshmi, B., Krasnitz, A., Hicks, J., Yu, M., Miller, W. T. and Muthuswamy, S. K. (2008). Brk is coamplified with ErbB2 to promote proliferation in breast cancer. *Proc. Natl. Acad. Sci. USA* **105**, 12463-12468. doi:10.1073/pnas.0805009105

Zhang, M., Lee, A. V. and Rosen, J. M. (2017). The cellular origin and evolution of breast cancer. *Cold Spring Harb. Perspect Med.* **7**, a027128. doi:10.1101/cshperspect.a027128

Zhou, J., Chen, Q., Zou, Y., Zheng, S. and Chen, Y. (2019). Stem cells and cellular origins of mammary gland: updates in rationale, controversies, and cancer relevance. *Stem Cells Int.* **2019**, 4247168. doi:10.1155/2019/4247168