

# Newly Identified Breast Luminal Progenitor and Gestational Stem Cell Populations Likely Give Rise to HER2-Overexpressing and Basal-Like Breast Cancers

James R. W. McMullen

Loma Linda University

Ubaldo Soto (✉ [usoto@llu.edu](mailto:usoto@llu.edu))

Loma Linda University

---

## Research Article

**Keywords:** Mammary Gland Development, Breast cancer, Basal-like breast cancer, HER2 breast cancer, Breast neoplasms

**Posted Date:** March 21st, 2022

**DOI:** <https://doi.org/10.21203/rs.3.rs-1449625/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Abstract

Breast Cancer (BrC) is a common malignancy with genetically diverse subtypes. There is evidence that specific BrC subtypes originate from particular normal mammary cell populations. However, the cell populations that give rise to most BrC subtypes are unidentified. Several human breast scRNAseq datasets are available. In this research, we utilized a robust human scRNAseq dataset to identify population-specific marker genes and then identified the expression of these marker genes in specific BrC subtypes. In humans, several BrC subtypes, HER2-enriched, basal-like, and triple-negative (TN), are more common in women who have had children. This observation suggests that cell populations that originate during pregnancy give rise to these BrCs. The current human datasets have few normal parous samples, so we supplemented this research with mouse datasets, which contain mammary cells from various developmental stages. This research identified two novel normal breast cell populations that may be the origin of the basal-like and HER2-overexpressing subtypes, respectively. A stem cell-like population, SC, that expresses gestation-specific genes has similar gene expression patterns to basal-like BrCs. A novel luminal progenitor cell population and HER2-overexpressing BrCs are marked by S100A7, S100A8, and S100A9 expression. We bolstered our findings by examining SC gene expression in TN BrC scRNAseq datasets. We discovered that several potential cancer stem cell populations highly express most of the SC genes. In summary, normal SC and the novel luminal progenitor cell population likely give rise to basal-like and HER2-overexpressing BrCs, respectively. Characterizing these normal cell populations may facilitate a better understanding of specific BrCs subtypes.

## 1 Introduction

Breast cancer (BrC) is the most common cancer and the second most common cause of cancer-related death in women [1, 2]. BrCs are categorized as Luminal A, Luminal B, HER2-enriched, and triple-negative (TN) for treatment purposes [3]. TN BrCs are heterogeneous, and researchers have used gene expression patterns to subdivide them into seven categories, including basal-like BrC [3, 4].

There is evidence that specific cell populations give rise to distinct BrC subtypes, but the origin of most BrC subtypes is unknown. Two research papers identified that luminal progenitor cells give rise or are associated with basal-like TN BrC [5, 6]. In mice, BRCA1 mutant mammary luminal progenitor cells can generate a tumor closely resembling the basal-like BrC subtype [5]. In human BRCA1 mutant basal-like cancers, there is an aberrant increase in the number of luminal progenitor cells [6]. The normal cell populations that become the other BrC subtypes are unknown.

Multiple scRNAseq datasets have provided plenty of information on mammary cell populations in humans and mice [7–11]. The human datasets are limited since they primarily focus on characterizing normal adult nulliparous cell populations [7, 8]. The mouse datasets examine the mammary tissue in multiple crucial developmental stages, including fetal and adult nulliparous, gestational, lactating, and post-involution [9, 10]. The mouse datasets illustrate a considerable change in mammary cell populations with pregnancy. One surprising finding in the mouse is that nulliparous and post-involution mammary cell

populations are different [9]. This feature is important because these different cell populations may become specific BrC subtypes in humans.

Nulliparous and parous women have different risks for BrC subtypes. Parous women have an increased risk of HER2-enriched, basal-like BrCs, and TN BrCs but reduced Luminal A and Luminal B tumor risk compared to nulliparous women [12–17]. The change in breast cancer risk led researchers to hypothesize that pregnancy changes human breast cell populations in normal development [18]. We now know this phenomenon occurs in mice, but it remains unproven in humans [9]. From the above data, we conclude that Luminal A and Luminal B cancers likely originate from nulliparous cell populations, while HER2-enriched and TN BrCs likely originate from post-pregnancy cell populations.

Since parous human breast cells are not well characterized while mouse mammary cells are, we wanted to use both the human and mouse datasets, but we needed to confirm that the mouse mammary gland is a good model of the human breast [9, 10]. Both mice and human mammary cells share the same eight general cell populations: stem cells, common progenitors, luminal progenitors, differentiated ductal cells, alveolar progenitors, differentiated alveolar cells, myoepithelial progenitors, and myoepithelial cells [19–21]. Further, the mouse and human mammary cell populations share multiple markers that identify these populations, including ITGA6, ESR1, PGR, KIT, CD29, and EpCAM, and even share differentiation patterns [8, 19–23].

We characterized normal human adult mammary cell populations from published data and identified marker genes for these populations. We then identified marker gene expression in human BrC subtypes and in the mouse mammary datasets. Lastly, based on shared gene expression patterns, we identified two potential cell populations of origin for human BrC subtypes.

## 2 Material And Methods

### 2.1 scRNA-seq bioinformatics

Patient samples from the GSE161529 dataset from Pal and colleagues were downloaded from the Gene Expression Omnibus (GEO) server [7]. We generated a folder containing a barcode, features, and matrix file for each patient sample. We imported seven normal patient sample datasets with the Read10X Seurat command in R. Patient samples N-N280-Epi, N-N1105-Epi, N-MH0064-Epi are premenopausal and nulliparous, N-MH0023-Epi is premenopausal and parous, N-PM0342-Epi is postmenopausal and nulliparous, and N-PM0372-Epi and N-MH275-Epi are postmenopausal and parous [7]. We removed potentially low-quality cells with  $\geq 20\%$  mitochondrial genes,  $< 500$  genes,  $< 1,500$  unique molecular identifiers (UMI), or  $< 0.8 \log_{10} \text{GenesPerUMI}$ . We removed genes expressed in less than ten cells. SCTransform was used to normalize, scale, and find variable features in the data. Anchor-based integration was performed. We used 100 principal components for PCA analysis, followed by t-SNE and UMAP analysis. SCTransform is highly effective, so 100 principle components contribute to robust analysis. We performed a clustering analysis with several cluster resolution values and generated

heatmaps, t-SNE, and dot plots. The top differentially expressed genes were used as population-specific marker genes and examined in detail. We also examined eight scRNAseq TN BrC datasets from GSE161529 using the above protocols, with the additional step of removing cell populations that highly express CD31 (endothelial cell marker) and CD45 (immune cell marker). Four samples had BRCA mutations: TN-B1-MH4031, TN-B1-MH0131, TN-B1-Tum0554, TN-B1-MH0177; “B1” indicates BRCA mutant. Four samples had normal BRCA: TN-MH0126, TN-MH0135, TN-SH0106, TN-MH0114-T2.

## **2.2 Using GENT2 to identify marker gene expression in human breast cancers**

We then examined the expression of the identified population-specific marker genes in human BrCs. The GENT2 database facilitated this analysis [24]. GENT2 collates and runs statistics on breast cancer microarray datasets from multiple groups in the GEO database.

## **2.3 Identifying human population-specific marker gene expression patterns in mouse mammary datasets**

Chung and colleagues and Bach and colleagues created user-friendly online resources to explore the RNA expression in mouse mammary gland cell populations in fetal and adult nulliparous, gestational, lactating, and post-involution glands [9, 10]. [https://wahl-lab-salk.shinyapps.io/Mammary\\_snATAC](https://wahl-lab-salk.shinyapps.io/Mammary_snATAC) and <https://marionilab.cruk.cam.ac.uk/mammaryGland> are the respective web addresses [9, 10]. We identified the mouse mammary cell populations that expressed the human population-specific marker genes.

# **3 Results**

## **3.1 Human mammary gland cells have population-specific gene markers**

Pal and colleagues robustly analyzed human breast tissue in pre and postmenopausal states [7]. We reexamined seven normal patient samples from their dataset, GSE161529, looking for population-specific markers at several clustering resolutions. The cells from different patient samples are shown in a t-SNE plot with dots representing individual cells and dot color representing individual tissue samples (Fig. 1A). We tested multiple clustering resolutions and decided to focus on 0.05 and 0.14 clustering resolutions based on strong marker gene expression patterns (Fig. 1B, 1C). At 0.05 clustering resolution, we used known mammary epithelial cell population markers to identify and group the human cell populations into luminal progenitor (LP), luminal differentiated (LD), basal (B), and contaminate stromal (Str) cells (Fig. 1B). When we reexamined the cells at 0.14 cluster resolution, subpopulations were identified, including three luminal progenitors (LP1-3), two luminal differentiated (LD1-2), a basal (B1), a transition (T), a

stem cell-like (SC), and a stromal cell (Str) population (Fig. 1C). The expression of the marker

genes used to identify LP, LD, and B populations are shown in Fig. 1D. FOXA1, PGR (progesterone receptor), ESR1 (estrogen receptor) identify luminal differentiated cells [22]. ALDH1A3 and KIT marks luminal progenitor cells [22, 25]. ACTA2 (smooth muscle actin) and KRT5 mark basal cells [26]. Lastly, KRT18 is a general luminal marker identifying LP and LD [26]. We removed stromal cells from further analysis and made heatmaps for the top 20 and 10 differentially expressed genes at 0.05 and 0.14 cluster resolutions, respectively (Fig. 1E).

Figure 2A shows the expression of the marker genes from Fig. 1D in the gene clusters in a dot plot. The dot plot shows the percentage of cells within a population that express the gene, based on the dot size, and the relative gene expression level, indicated by the dot color. We identified a transitional population (T) positioned between the luminal and basal cells with low luminal and basal marker expression (Fig. 2A). As expected, the rest of the markers are localized to the respective cell subpopulations. We also identified the “SC” population as a stem cell-like population because it expresses multiple stem cell-related genes, including ALDH1A3, BIRC5, CDK6, HMGB2, and STMN1 (Fig. 1D, 1E, 2B). Interestingly, ALDH1A3 identifies mammary luminal progenitor and stem cells [22, 27]. BIRC5 is an oncofetal gene highly expressed in embryonic stem cells [28]. CDK6 regulates exit from quiescence and is required for normal function in hematopoietic stem cells and maintains self-renewal in leukemia stem cells [29–31]. HMGB2 is expressed in proliferating neural stem cells, regulates telomerase activity, and maintains stemness in hematopoietic and mesenchymal stem cells [32–35]. STMN1 is also associated with cycling stem cells [36]. Because of the expression of these genes in “SC,” we

identified SC as a stem cell-like population. Genes that were highly expressed in SC were placed

in Table 1. In Fig. 2B, we show the cluster-specific expression of the SC genes (Table 1) as well as a luminal (KRT18), basal (KRT5), epithelial (EPCAM), and mesenchymal (VIM) marker gene. Interestingly, SC has luminal, basal, epithelial, and mesenchymal properties based on gene expression. We also show the expression of LP2 and LP3 marker genes and recorded these genes in Table 1 (Fig. 2C and 2D).

Table 1

Genes highly expressed in the SC, LP2, and LP3 populations, and their expression in mouse mammary populations and human breast cancer

Gene Marker	Identified Human Cell Populations	Identified Mouse Cell Populations	GENT2 data- BrC subtype with highest gene expression	GENT2 data- highest Log2 fold change between BrC subtypes
SC genes				
BIRC5	SC	Very hi in fMaSC, med in AD1, AP1, B1	Basal	1.994***
CDK6	SC	Medium in all LP, AP, AD, B1 B2, B4, and B5; low in fMaSC	Basal	1.831***
CENPF	SC	AD1, AP1, B1; low in fMaSC	Basal	1.858***
CENPW	SC	medium in fMaSC, AD1, AP1, B1, low in rest	Basal	2.248***
FDCSP	Hi in LP3, low in rest	-	Basal	4.458***
HIST1H4C	SC, low in rest	Low in fMaSC	Triple-negative	3.214***
HMGB2	SC, low in rest	Very hi in fMaSC, AP1, and AD1, medium in rest	Triple-negative (second highest in Basal)	1.833***
STMN1	Hi in SC, low in rest	Very hi in fMaSC, hi in AP1, AD1, AP2, B1, low in rest	Triple-negative (second highest in Basal)	2.208***
TOP2A	SC	Hi in fMaSC, medium in AP1, AD1, B1	Basal	1.634***
TPX2	SC	fMaSC, AP1, AD1, B1	Basal	2.264***
TYMS	SC	Hi in fMaSC medium in AP1, AD1, B1, low in rest	Basal	1.899***
UBE2C	SC	Very hi in fMaSC, hi in AP1, AD1, B1	Basal	1.949***
UBE2S	Hi in SC, low in rest	Medium-low in all adult cell populations	Basal	1.988***
LP2 genes				
CXCL5	LP2	Hi in LP2, medium in LP1 and LP4	Triple-negative	2.142***

Gene Marker	Identified Human Cell Populations	Identified Mouse Cell Populations	GENT2 data- BrC subtype with highest gene expression	GENT2 data- highest Log2 fold change between BrC subtypes
LCN2	Hi in LP2, low in rest of LP	Very hi in LP2, medium-low in rest of adult cell populations	Basal	2.587***
S100A7	LP2	Low in LP1 & B3	HER2	5.443***
S100A8	LP2	LP1, LP2	HER2	4.348***
S100A9	Hi in LP2, low in rest of LP	none	HER2	4.637***
SAA1	Hi in LP2, low in rest of LP, B1	AD2, LP2	Triple-negative	2.659***
SAA2	LP2, low in LP3	AD2	-	-
SERPINB4	LP2	-	Luminal	1.503***
SLPI	Hi in LP2, low in rest of LP and SC	Hi in B2, low in rest of adult cell populations	Basal	2.681***
LP3 genes				
CHI3L1	LP2, LP3	-	Basal	2.780***
CYP1B1	LP3, low in LP1 and LP2	-	Triple-negative	1.410***
FDCSP	LP3	-	Basal	4.458***
LTF	LP3, low in LP1 and LP2	Hi in AP2, LP2, medium in LP1, LP3, and LP4	HER2	2.825***
RARRES1	LP2, LP3, low in LP1	none	Basal	3.189***
*** p < 0.001				
GENT2 data comes from n = 2164 microarrayed patient samples in the GEO database				
fMaSC fetal mammary stem cells, B basal, AD differentiated alveolar cells, AP alveolar progenitor cells, LP luminal progenitor cells, LD differentiated luminal cells, SC stem cell, - gene not in dataset				

## 3.2 GENT2 identifies breast cancers associated with normal breast population-specific markers

We utilized the tool GENT2 to identify if the population-specific markers from differential expressed gene analysis were strongly associated with specific BrC subtypes. GENT2 runs statistical analysis on a compilation of microarray data from 2,164 BrC patient tumor samples [24]. We recorded the resulting analyses in Table 1. Table 1 shows the BrC subtype that most expresses the gene marker and the largest log 2-fold gene expression difference between subtypes. Significant fold change differences between BrC subtypes indicate that the cell population marker is also BrC subtype-specific and can be a promising marker for BrC diagnosis. LD2, T, and B1 related genes are not strongly associated with specific human BrC subtypes or generally have low log 2-fold change values in the GENT2 database (Table S1). LP1 marker genes had considerable expression overlap with the other LP populations and SC, and LD1 had considerable overlap with LD2, so we did not look into these or the above populations further (Table S1).

The best LP3 marker gene, FDCSP, is highly expressed specifically in Basal-like BrC, and the less specific LP3 marker genes are also highly expressed in Basal-like BrC (Table 1). However, the LP3 marker genes are not all expressed primarily in one cancer subtype, and this suggests that it either does not give rise to cancer subtypes or can give rise to several cancer subtypes. Interestingly, in LP2, the most specific marker genes, S100A7, S100A8, and S100A9, are explicitly expressed in HER2 BrCs in the GENT2 database (Table 1). We hypothesize that LP2 may be the cell population that becomes HER2-overexpressing BrC.

### **3.3 Expression of human population-specific marker genes in mouse mammary gland cell populations**

Having extracted many exciting data from this human dataset, we then examined mouse mammary RNAseq/scRNAseq datasets using graphical user interfaces designed by the data's creators [9, 10]. As stated in the introduction, the mouse datasets are more fleshed out in terms of cell populations in mammary developmental stages than the current human datasets, having fetal cells and adult nulliparous, gestational, lactating, and post-involution cells. We thought these datasets could contribute to our study. In the mouse datasets, there are fetal mammary stem cells (fMaSC), five basal cell populations (B1-B5), two differentiated alveolar cell populations (AD1-2), two alveolar progenitor cell populations (AP1-2), four luminal progenitor cell populations (LP1-4), and two differentiated luminal cell populations (LD1-2). Figure 3A shows a representation of the adult mouse mammary epithelial cell populations based on the work of Bach and colleagues [9]. Figure 3B shows the mouse mammary cell populations in the nulliparous, gestation, lactation, and post involution states. We identified the expression of the Table 1 genes in the mouse datasets, determined which mouse cell population expressed the gene, and recorded these populations in Table 1. We found some exciting results.

In the mouse data, the majority of the specific markers for the human SC population (9 of 12), BIRC5, CENPF, CENPW, HMGB2, STMN1, TOP2A, TPX2, TYMS, and UBE2C, are expressed in gestational-specific mouse cell populations AP1, AD1, and B1 (Fig. 3B, Table 1) [9]. Further, most of the SC genes (11 of 12) were expressed at some level, specifically in fMaSCs (Table 1), reinforcing our theory that SC is indeed a mammary stem cell population [10]. The majority of the SC population is from tissue sample PM0372-Epi, a parous postmenopausal woman. Further, in the GENT2 data, most of these markers are highly

expressed primarily in basal-like BrCs, and two of the three exceptions, HMGB2 and STMN1, have the second-highest expression in basal-like BrCs. We hypothesize that SC is a remnant of a human gestational cell population. Further, we hypothesize that SC is the cell of origin for many basal-like BrCs.

### **3.4 Expression of normal breast stem cells markers in scRNAseq TN human breast cancer**

Having examined SC gene expression in GENT2 BrC data and in mouse datasets, we then examined eight TN BrC scRNAseq datasets from GSE161529. Figure 4A shows each dataset's cells (dots) in a t-SNE plot. Notably, most TN BrCs formed individual non-overlapping groups, demonstrating the variability within the TN subtype. Figure 4B shows the cells clustered at a 0.25 resolution. We then examined the SC marker genes in these clusters. The resultant dot plot shows that two cancer cell populations highly express many SC genes (Fig. 4C). Populations 1 and 7 have high SC gene expression. Notably, population 7 is from a BRCA mutant dataset while 1 is BRCA normal. Based on the SC gene expression, it appears that populations 1 and 7 are potential cancer stem cell (CSC) populations. Populations 1 and 7 both have adjacent cell populations, 0 and 2, from the same respective patient samples. These last two cell populations are likely composed of cells that differentiated from the respective CSC populations.

Overall, this research suggests that a gestation-specific stem-cell population exists in adult human breast tissue and that this population could be the origin of Basal-like BrCs and Basal-like CSCs. Further, the identified marker genes for this population are highly population-specific in the mammary gland, are related to stem cell function, and are, therefore, promising targets for treatment in Basal-like BrC. We also identified a luminal progenitor population marked by S100A7, S100A8, and S100A9 that may be the origin of HER2 enriched BrCs.

## **4 Discussion**

The cell populations of origin for specific BrC subtypes are generally unknown. If these specific cell populations are identified, we can identify cellular pathways that contribute to subtype-specific carcinogenesis and potentially screen patients for malignant mutations in these cell populations that would predispose patients to BrC. This research compares normal breast cell populations to human BrC subtypes and normal mouse mammary gland populations. We identified shared gene expression patterns and predicted two cell populations that may transform into Basal-like and HER2-overexpressing BrCs, respectively. Basal BrC is a subtype of TN, and we went on to examine TN BrC scRNAseq datasets [3]. From this research, we identify two potential CSC populations within a subset of TN BrCs. These CSCs highly expressed many of the SC genes.

The identified normal SC population and the corresponding CSC populations have known mammary gland stemness properties. SC and the CSCs have luminal, basal, epithelial, and mesenchymal marker expression. Mammary stem cells are known to express both luminal and basal markers and even vimentin, a mesenchymal marker [27, 37]. Also, partial EMT, which is characterized by co-expression of epithelial and mesenchymal markers, identifies BrC cells with stem cell properties [38, 39].

The SC population we identified expressed twelve marker genes, BIRC5, CDK6, CENPF, CENPW, HIST1H4C, HMGB2, STMN1, TOP2A, TPX2, TYMS, UBE2C, and UBE2S. The scientific literature not only corroborates that many of these genes are stem cell related, but also that they are specifically associated with the basal-like BrC subtype. BIRC5, CENPF, FDCSP, HIST1H4C, HMGB2, STMN1, TYMS, UBE2C, and UBE2S are upregulated in BrCs, specifically in basal-like BrCs for BIRC5, CENPF, STMN1, TYMS [4, 40–53]. In several papers where TN BrCs were not subdivided to include a separate section for the basal-like subtype, the TN subtype showed high expression of SC marker genes. CENPW, TPX2, and UBE2C are overexpressed in TN BrCs [54–56]. UBE2C expression is upregulated in HER2 expressing and TN BrCs [47, 48].

Many of the SC marker genes are correlated with poor BrC patient survival statistics in the literature. BrCs expressing BIRC5, CENPF, or UBE2C have reduced disease-free, metastasis-free, and overall survival [41, 48, 50, 57, 58]. HIST1H4C expression is associated with worse overall and metastasis-free survival in BrCs [44]. HMGB2, STMN1, and TPX2 expression is correlated with worse disease-free and overall survival [45, 52, 56, 59, 60]. TYMS expressing BrCs have low overall survival [46].

Several SC marker genes have known stem cell-related properties in BrCs. STMN1 expression is associated with the CD44+/CD24- BrC stem cell phenotype [59]. TYMS maintains BrC spheroid formation efficiency and CD24- status [46]. TPX2 and UBE2C knockdown reduce colony formation efficiency in TN BrCs [55]. Lastly, UBE2S knockdown suppresses anchorage-independent growth in BrCs [49].

Interestingly, several breast cancer papers identify co-expression of the SC marker genes. CENPF expression correlates with BIRC5 expression [4, 58]. TPX2 and UBE2C were highly expressed in the same TN BrC cell populations and cell lines [55]. Overall, the scientific literature corroborates and greatly bolsters the association of the SC markers with basal-like BrC.

Most SC marker genes are specifically upregulated in basal-like or TN BrC subtypes, and we hypothesize that malignant SC cells create basal-like BrCs. These marker genes were almost exclusively in the SC population in the normal human breast epithelium; therefore, these genes could be promising targets for targeted treatment of basal-like BrC with minimal local normal tissue damage. For instance, the SC gene BIRC5 is an onco-fetal protein rarely expressed in adult tissues and BIRC5 inhibitors have been shown to be effective in *in vitro* BrC treatment [61, 62]. A CENPF inhibitor also shows promise in BrC treatment *in vitro* [63].

Besides the SC population, SC marker genes, and potential TN CSC populations, we also identified a S100A7, S100A8, and S100A9 expressing normal mammary luminal progenitor cell population. We determined that these three genes are strongly associated with the HER2 BrC subtype and hypothesized that transformed cells from this population become HER2-overexpressing BrCs. Many scientific papers have shown that these three genes are associated with HER2 BrCs and are often correlated with poor BrC patient outcomes. S100A7 expression is negatively correlated with ESR1 and PGR in human BrCs and correlated with decreased disease-free and overall survival [64, 65]. High S100A8 expression is positively correlated with HER2 expression and negatively correlated with ESR1 and PGR expression [66–68]. Further, high S100A8 expression is associated with increased cancer relapse and lower overall and

disease-free survival [66]. S100A9 has also been correlated with HER2 BrCs and poor overall survival [67–69]. Further, a Luminal A cell line treated with 100A8/A9 had a marked decrease in ESR1 expression, suggesting that S100A8/A9 may have a causal role in the HER2 BrC phenotype [67]. Together, there is strong evidence in the literature that S100A7, S100A8, and S100A9 are negative BrC prognostic markers and are associated with HER2 BrCs, as we suggest in this research.

Interestingly, S100A7, S100A8, and S100A9 are also associated with stem-cell properties in BrC. In BrC, expression of these three genes is associated with effective mammosphere formation and inhibition of these genes stunted mammosphere growth and xenograft tumor growth [70].

As with the SC marker genes, S100A7, S100A8, and S100A9 are promising targets for treatment. We found that these genes are almost exclusively in one luminal progenitor cell population, suggesting that targeted treatment for these genes in HER2 BrC would not cause significant off-target damage to most breast cells, contributing to better patient outcomes.

## **5 Conclusion**

In conclusion, we identified a normal stem cell-like cell population and its marker genes in adult human breast tissue. The marker genes are expressed almost exclusively in gestation-specific adult mouse cell populations, suggesting that it is a human gestation-specific stem cell population. Further, many of these genes are specifically expressed in basal-like and TN BrC. We identified two potential CSC populations within a subset of TN BrCs. These populations highly express many of the SC genes and may represent normal stem cells that have become cancerous. We also identified a novel normal luminal progenitor cell population marked by three genes (S100A7, S100A8, S100A9) that are also explicitly overexpressed in HER2 breast cancer. We hypothesize that these two cell populations give rise to the basal-like/TN and HER2-overexpressing breast cancer subtypes, respectively.

## **Declarations**

### **Acknowledgments**

The Loma Linda University Department of Basic Sciences and the Center for Health Disparities and Molecular Medicine provided the resources used in this project. The authors thank Dr. Jiang Zhong (Loma Linda University) for his critical reading of the paper and helpful comments. The authors also thank Dr. Isaac Kremsky (Loma Linda University) for his advice on scRNAseq data processing.

### **Funding**

The Loma Linda University Department of Basic Sciences provided funding for this project.

### **Conflicts of interest/Competing interests**

The authors declare that they have no conflicts of interest.

## Data availability

Publicly available datasets were used for this research. The GSE161529 dataset was produced by Pal and colleagues and is available in the Gene Expression Omnibus (GEO) archive [7]. GENT2 data was from the GENT2 server at <http://gent2.appep.kr/gent2/> [24]. [https://wahl-lab-salk.shinyapps.io/Mammary\\_snATAC](https://wahl-lab-salk.shinyapps.io/Mammary_snATAC) and <https://marionilab.cruk.cam.ac.uk/mammaryGland> are the website addresses that display the mouse datasets in a user-friendly research format provided by their respective authors.

## Code availability

The workflows we used with R are stored as supplementary word files.

## Author's contributions

Paper concept and design was done by JM. Bioinformatics analysis and manuscript writing were performed by JM. JM and US contributed to the reviewing and editing of the manuscript prior to article submission.

## References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *Cancer J Clin.* 2020;70(1):7–30.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin.* 2018;68(6):394–424.
3. Dai X, Cheng H, Bai Z, Li J. Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. *J Cancer.* 2017;8(16):3131–41.
4. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Investig.* 2011;121(7):2750–67.
5. Molyneux G, Geyer FC, Magnay FA, McCarthy A, Kendrick H, Natrajan R, Mackay A, Grigoriadis A, Tutt A, Ashworth A, et al. BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell Stem Cell.* 2010;7(3):403–17.
6. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med.* 2009;15(8):907–13.
7. Pal B, Chen Y, Vaillant F, Capaldo BD, Joyce R, Song X, Bryant VL, Penington JS, Di Stefano L, Tubau Ribera N, et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* 2021;40(11):e107333.
8. Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, Phung AT, Willey E, Kumar R, Jabart E, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity.

- Nat Commun. 2018;9(1):2028.
9. Bach K, Pensa S, Grzelak M, Hadfield J, Adams DJ, Marioni JC, Khaled WT. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun.* 2017;8(1):2128.
  10. Chung CY, Ma Z, Dravis C, Preissl S, Poirion O, Luna G, Hou X, Girardi RR, Ren B, Wahl GM. Single-Cell Chromatin Analysis of Mammary Gland Development Reveals Cell-State Transcriptional Regulators and Lineage Relationships. *Cell Rep.* 2019;29(2):495–510.e496.
  11. Hu L, Su L, Cheng H, Mo C, Ouyang T, Li J, Wang T, Fan Z, Fan T, Lin B, et al. Single-Cell RNA Sequencing Reveals the Cellular Origin and Evolution of Breast Cancer in BRCA1 Mutation Carriers. *Cancer Res.* 2021;81(10):2600–11.
  12. Fortner RT, Sisti J, Chai B, Collins LC, Rosner B, Hankinson SE, Tamimi RM, Eliassen AH: **Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: results from the Nurses' Health Studies.** *Breast cancer research: BCR* 2019, **21**(1):40.
  13. Millikan RC, Newman B, Tse C-K, Moorman PG, Conway K, Smith LV, Labbok MH, Geradts J, Bensen JT, Jackson S, et al. Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat.* 2008;109(1):123–39.
  14. Palmer JR, Viscidi E, Troester MA, Hong CC, Schedin P, Bethea TN, Bandera EV, Borges V, McKinnon C, Haiman CA, et al: **Parity, lactation, and breast cancer subtypes in African American women: results from the AMBER Consortium.** *Journal of the National Cancer Institute* 2014, **106**(10).
  15. Chen L, Li CI, Tang MT, Porter P, Hill DA, Wiggins CL, Cook LS. **Reproductive Factors and Risk of Luminal, HER2-Overexpressing, and Triple-Negative Breast Cancer Among Multiethnic Women.** *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research.* cosponsored by the American Society of Preventive Oncology. 2016;25(9):1297–304.
  16. Gaudet MM, Gierach GL, Carter BD, Luo J, Milne RL, Weiderpass E, Giles GG, Tamimi RM, Eliassen AH, Rosner B, et al. Pooled Analysis of Nine Cohorts Reveals Breast Cancer Risk Factors by Tumor Molecular Subtype. *Cancer Res.* 2018;78(20):6011–21.
  17. Lambertini M, Santoro L, Del Mastro L, Nguyen B, Livraghi L, Ugolini D, Peccatori FA, Azim HA Jr. Reproductive behaviors and risk of developing breast cancer according to tumor subtype: A systematic review and meta-analysis of epidemiological studies. *Cancer Treat Rev.* 2016;49:65–76.
  18. Russo J, Moral R, Balogh GA, Mailo D, Russo IH. The protective role of pregnancy in breast cancer. *Breast cancer research: BCR.* 2005;7(3):131–42.
  19. Visvader JE. Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes Dev.* 2009;23(22):2563–77.
  20. Gross KM, Zhou W, Breindel JL, Ouyang J, Jin DX, Sokol ES, Gupta PB, Huber K, Zou L, Kuperwasser C. Loss of Slug Compromises DNA Damage Repair and Accelerates Stem Cell Aging in Mammary Epithelium. *Cell Rep.* 2019;28(2):394–407.e396.
  21. Arendt LM, Keller PJ, Skibinski A, Goncalves K, Naber SP, Buchsbaum RJ, Gilmore H, Come SE, Kuperwasser C. Anatomical localization of progenitor cells in human breast tissue reveals

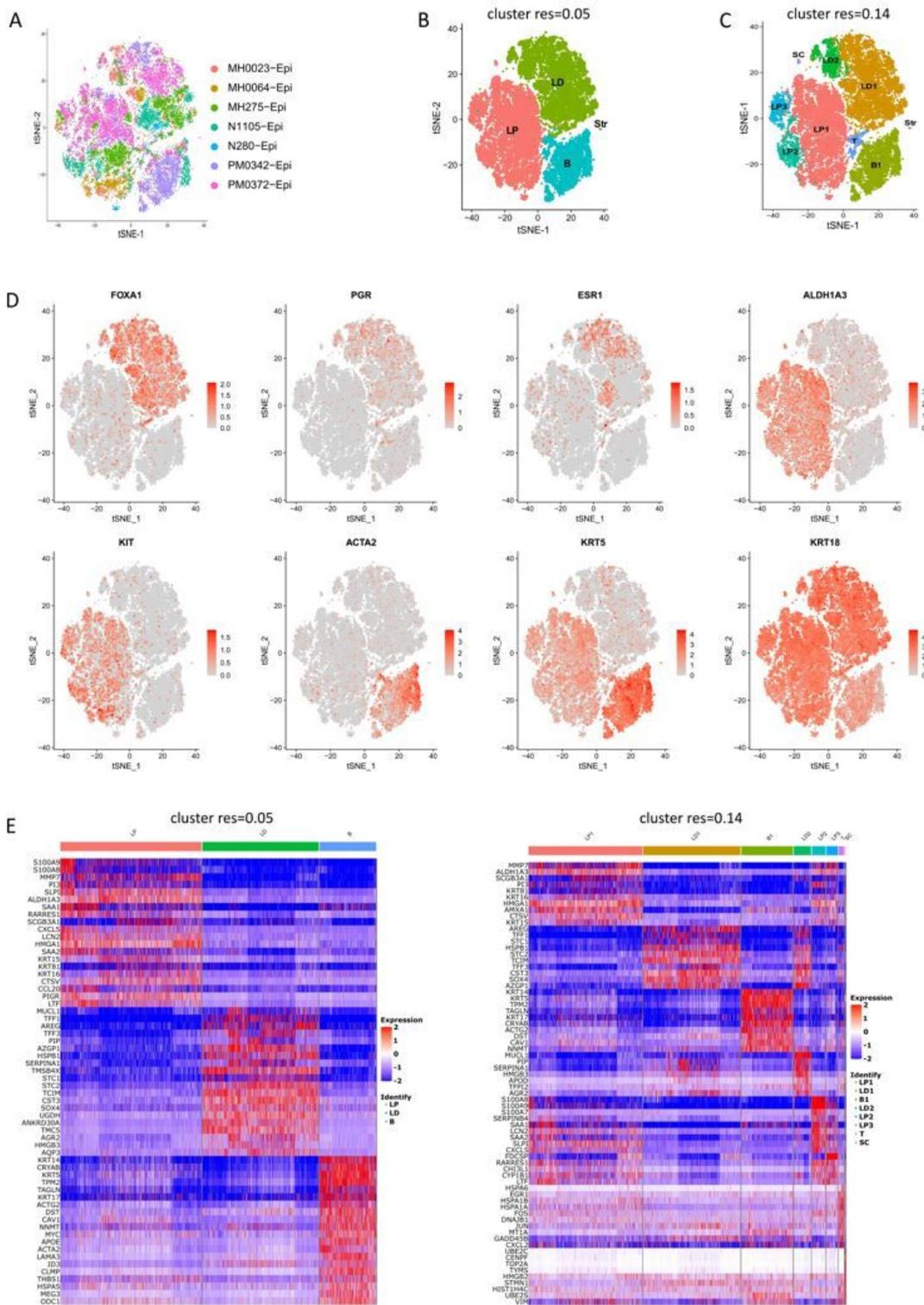
- enrichment of uncommitted cells within immature lobules. *Breast cancer research: BCR*. 2014;16(5):453.
22. Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE. Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast cancer research: BCR*. 2010;12(2):R21.
  23. Liu S, Cong Y, Wang D, Sun Y, Deng L, Liu Y, Martin-Trevino R, Shang L, McDermott SP, Landis MD, et al. Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem cell reports*. 2014;2(1):78–91.
  24. Park SJ, Yoon BH, Kim SK, Kim SY. GENT2: an updated gene expression database for normal and tumor tissues. *BMC Med Genom*. 2019;12(Suppl 5):101.
  25. Kim J, Villadsen R. Expression of Luminal Progenitor Marker CD117 in the Human Breast Gland. *J Histochem cytochemistry: official J Histochem Soc*. 2018;66(12):879–88.
  26. Abd El-Rehim DM, Pinder SE, Paish CE, Bell J, Blamey RW, Robertson JF, Nicholson RI, Ellis IO. Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol*. 2004;203(2):661–71.
  27. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, Brown M, Jacquemier J, Viens P, Kleer CG, Liu S, et al. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell*. 2007;1(5):555–67.
  28. Blum B, Bar-Nur O, Golan-Lev T, Benvenisty N. The anti-apoptotic gene survivin contributes to teratoma formation by human embryonic stem cells. *Nat Biotechnol*. 2009;27(3):281–7.
  29. Scheicher R, Hoelbl-Kovacic A, Bellutti F, Tigan AS, Prchal-Murphy M, Heller G, Schneckenleithner C, Salazar-Roa M, Zöchbauer-Müller S, Zuber J, et al. CDK6 as a key regulator of hematopoietic and leukemic stem cell activation. *Blood*. 2015;125(1):90–101.
  30. Laurenti E, Frelin C, Xie S, Ferrari R, Dunant CF, Zandi S, Neumann A, Plumb I, Doulatov S, Chen J, et al. CDK6 levels regulate quiescence exit in human hematopoietic stem cells. *Cell Stem Cell*. 2015;16(3):302–13.
  31. Li Z, Wang F, Tian X, Long J, Ling B, Zhang W, Xu J, Liang A. HCK maintains the self-renewal of leukaemia stem cells via CDK6 in AML. *J experimental Clin cancer research: CR*. 2021;40(1):210.
  32. Kimura A, Matsuda T, Sakai A, Murao N, Nakashima K. HMGB2 expression is associated with transition from a quiescent to an activated state of adult neural stem cells. *Dev dynamics: official publication Am Association Anatomists*. 2018;247(1):229–38.
  33. Kučírek M, Bagherpoor AJ, Jaroš J, Hampl A, Štros M. HMGB2 is a negative regulator of telomerase activity in human embryonic stem and progenitor cells. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*. 2019;33(12):14307–24.
  34. Zhang C, Fondufe-Mittendorf YN, Wang C, Chen J, Cheng Q, Zhou D, Zheng Y, Geiger H, Liang Y. Latexin regulation by HMGB2 is required for hematopoietic stem cell maintenance. *Haematologica*. 2020;105(3):573–84.

35. Taniguchi N, Caramés B, Hsu E, Cherqui S, Kawakami Y, Lotz M. Expression patterns and function of chromatin protein HMGB2 during mesenchymal stem cell differentiation. *J Biol Chem*. 2011;286(48):41489–98.
36. Han S, Fink J, Jörg DJ, Lee E, Yum MK, Chatzeli L, Merker SR, Josserand M, Trendafilova T, Andersson-Rolf A, et al. Defining the Identity and Dynamics of Adult Gastric Isthmus Stem Cells. *Cell Stem Cell*. 2019;25(3):342–56.e347.
37. Spike BT, Engle DD, Lin JC, Cheung SK, La J, Wahl GM. A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. *Cell Stem Cell*. 2012;10(2):183–97.
38. Kröger C, Afeyan A, Mraz J, Eaton EN, Reinhardt F, Khodor YL, Thiru P, Bieri B, Ye X, Burge CB, et al. Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc Natl Acad Sci USA*. 2019;116(15):7353–62.
39. Bieri B, Pierce SE, Kroeger C, Stover DG, Pattabiraman DR, Thiru P, Liu Donaher J, Reinhardt F, Chaffer CL, Keckesova Z, et al. Integrin-beta4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc Natl Acad Sci USA*. 2017;114(12):E2337–46.
40. Dai JB, Zhu B, Lin WJ, Gao HY, Dai H, Zheng L, Shi WH, Chen WX. **Identification of prognostic significance of BIRC5 in breast cancer using integrative bioinformatics analysis**. *Bioscience reports* 2020, 40(2).
41. Oparina N, Erlandsson MC, Fäldt Beding A, Parris T, Helou K, Karlsson P, Einbeigi Z, Bokarewa MI. **Prognostic Significance of BIRC5/Survivin in Breast Cancer: Results from Three Independent Cohorts**. *Cancers* 2021, 13(9).
42. Wang C, Zhou L, Li S, Wei J, Wang W, Zhou T, Liao S, Weng D, Deng D, Weng Y, et al. C4orf7 contributes to ovarian cancer metastasis by promoting cancer cell migration and invasion. *Oncol Rep*. 2010;24(4):933–9.
43. Wu M, Han L, Shi Y, Xu G, Wei J, You L, Chen Y, Zhu T, Li Q, Li S, et al. Development and characterization of a novel method for the analysis of gene expression patterns in lymphatic endothelial cells derived from primary breast tissues. *J Cancer Res Clin Oncol*. 2010;136(6):863–72.
44. Xie W, Zhang J, Zhong P, Qin S, Zhang H, Fan X, Yin Y, Liang R, Han Y, Liao Y, et al. Expression and potential prognostic value of histone family gene signature in breast cancer. *Experimental and therapeutic medicine*. 2019;18(6):4893–903.
45. Fu D, Li J, Wei J, Zhang Z, Luo Y, Tan H, Ren C. HMGB2 is associated with malignancy and regulates Warburg effect by targeting LDHB and FBP1 in breast cancer. *Cell communication and signaling: CCS*. 2018;16(1):8.
46. Siddiqui A, Gollavilli PN, Schwab A, Vazakidou ME, Ersan PG, Ramakrishnan M, Pluim D, Coggins S, Saatci O, Annaratone L, et al. Thymidylate synthase maintains the de-differentiated state of triple negative breast cancers. *Cell Death Differ*. 2019;26(11):2223–36.
47. Mo CH, Gao L, Zhu XF, Wei KL, Zeng JJ, Chen G, Feng ZB. The clinicopathological significance of UBE2C in breast cancer: a study based on immunohistochemistry, microarray and RNA-sequencing

- data. *Cancer Cell Int.* 2017;17:83.
48. Kim YJ, Lee G, Han J, Song K, Choi JS, Choi YL, Shin YK. UBE2C Overexpression Aggravates Patient Outcome by Promoting Estrogen-Dependent/Independent Cell Proliferation in Early Hormone Receptor-Positive and HER2-Negative Breast Cancer. *Front Oncol.* 2019;9:1574.
  49. Ayesha AK, Hyodo T, Asano E, Sato N, Mansour MA, Ito S, Hamaguchi M, Senga T. UBE2S is associated with malignant characteristics of breast cancer cells. *Tumour biology: the journal of the International Society for Oncodevelopmental Biology and Medicine.* 2016;37(1):763–72.
  50. Sun J, Huang J, Lan J, Zhou K, Gao Y, Song Z, Deng Y, Liu L, Dong Y, Liu X. Overexpression of CENPF correlates with poor prognosis and tumor bone metastasis in breast cancer. *Cancer Cell Int.* 2019;19:264.
  51. Ueda S, Kondoh N, Tsuda H, Yamamoto S, Asakawa H, Fukatsu K, Kobayashi T, Yamamoto J, Tamura K, Ishida J, et al. Expression of centromere protein F (CENP-F) associated with higher FDG uptake on PET/CT, detected by cDNA microarray, predicts high-risk patients with primary breast cancer. *BMC Cancer.* 2008;8:384.
  52. Askeland C, Wik E, Finne K, Birkeland E, Arnes JB, Collett K, Knutsvik G, Krüger K, Davidsen B, Aas T, et al. Stathmin expression associates with vascular and immune responses in aggressive breast cancer subgroups. *Sci Rep.* 2020;10(1):2914.
  53. Curmi PA, Noguès C, Lachkar S, Carelle N, Gonthier MP, Sobel A, Lidereau R, Bièche I. Overexpression of stathmin in breast carcinomas points out to highly proliferative tumours. *Br J Cancer.* 2000;82(1):142–50.
  54. Li Y, Tang XQ, Bai Z, Dai X. Exploring the intrinsic differences among breast tumor subtypes defined using immunohistochemistry markers based on the decision tree. *Sci Rep.* 2016;6:35773.
  55. Elango R, Vishnubalaji R, Shaath H, Alajez NM. Molecular subtyping and functional validation of TTK, TPX2, UBE2C, and LRP8 in sensitivity of TNBC to paclitaxel. *Mol therapy Methods Clin Dev.* 2021;20:601–14.
  56. Matson DR, Denu RA, Zasadil LM, Burkard ME, Weaver BA, Flynn C, Stukenberg PT. High nuclear TPX2 expression correlates with TP53 mutation and poor clinical behavior in a large breast cancer cohort, but is not an independent predictor of chromosomal instability. *BMC Cancer.* 2021;21(1):186.
  57. Psyrris A, Kalogeras KT, Kronenwett R, Wirtz RM, Batistatou A, Bournakis E, Timotheadou E, Gogas H, Aravantinos G, Christodoulou C, et al. Prognostic significance of UBE2C mRNA expression in high-risk early breast cancer. A Hellenic Cooperative Oncology Group (HeCOG) Study. *Annals of oncology: official journal of the European Society for Medical Oncology.* 2012;23(6):1422–7.
  58. O'Brien SL, Fagan A, Fox EJ, Millikan RC, Culhane AC, Brennan DJ, McCann AH, Hegarty S, Moyna S, Duffy MJ, et al. CENP-F expression is associated with poor prognosis and chromosomal instability in patients with primary breast cancer. *Int J Cancer.* 2007;120(7):1434–43.
  59. Obayashi S, Horiguchi J, Higuchi T, Katayama A, Handa T, Altan B, Bai T, Bao P, Bao H, Yokobori T, et al. Stathmin1 expression is associated with aggressive phenotypes and cancer stem cell marker expression in breast cancer patients. *Int J Oncol.* 2017;51(3):781–90.

60. Kuang XY, Chen L, Zhang ZJ, Liu YR, Zheng YZ, Ling H, Qiao F, Li S, Hu X, Shao ZM. Stathmin and phospho-stathmin protein signature is associated with survival outcomes of breast cancer patients. *Oncotarget*. 2015;6(26):22227–38.
61. Xiao M, Li W. Recent Advances on Small-Molecule Survivin Inhibitors. *Curr Med Chem*. 2015;22(9):1136–46.
62. Wang W, Zhang B, Mani AM, Wu Z, Fan Y, Li W, Wu ZH. Survivin Inhibitors Mitigate Chemotherapeutic Resistance in Breast Cancer Cells by Suppressing Genotoxic Nuclear Factor- $\kappa$ B Activation. *J Pharmacol Exp Ther*. 2018;366(1):184–93.
63. Brown HK, Ottewell PD, Coleman RE, Holen I. The kinetochore protein Cenp-F is a potential novel target for zoledronic acid in breast cancer cells. *J Cell Mol Med*. 2011;15(3):501–13.
64. Al-Haddad S, Zhang Z, Leygue E, Snell L, Huang A, Niu Y, Hiller-Hitchcock T, Hole K, Murphy LC, Watson PH. Psoriasin (S100A7) expression and invasive breast cancer. *Am J Pathol*. 1999;155(6):2057–66.
65. Emberley ED, Niu Y, Njue C, Kliever EV, Murphy LC, Watson PH. Psoriasin (S100A7) expression is associated with poor outcome in estrogen receptor-negative invasive breast cancer. *Clin cancer research: official J Am Association Cancer Res*. 2003;9(7):2627–31.
66. Wang D, Liu G, Wu B, Chen L, Zeng L, Pan Y. Clinical Significance of Elevated S100A8 Expression in Breast Cancer Patients. *Front Oncol*. 2018;8:496.
67. Bao YI, Wang A, Mo J. S100A8/A9 is associated with estrogen receptor loss in breast cancer. *Oncol Lett*. 2016;11(3):1936–42.
68. Song R, Struhl K. **S100A8/S100A9 cytokine acts as a transcriptional coactivator during breast cellular transformation**. *Science advances* 2021, 7(1).
69. Bergenfelz C, Gaber A, Allaoui R, Mehmeti M, Jirström K, Leanderson T, Leandersson K. S100A9 expressed in ER(-)PgR(-) breast cancers induces inflammatory cytokines and is associated with an impaired overall survival. *Br J Cancer*. 2015;113(8):1234–43.
70. Goh JY, Feng M, Wang W, Oguz G, Yatim SMJM, Lee PL, Bao Y, Lim TH, Wang P, Tam WL, et al. Chromosome 1q21.3 amplification is a trackable biomarker and actionable target for breast cancer recurrence. *Nat Med*. 2017;23(11):1319–30.

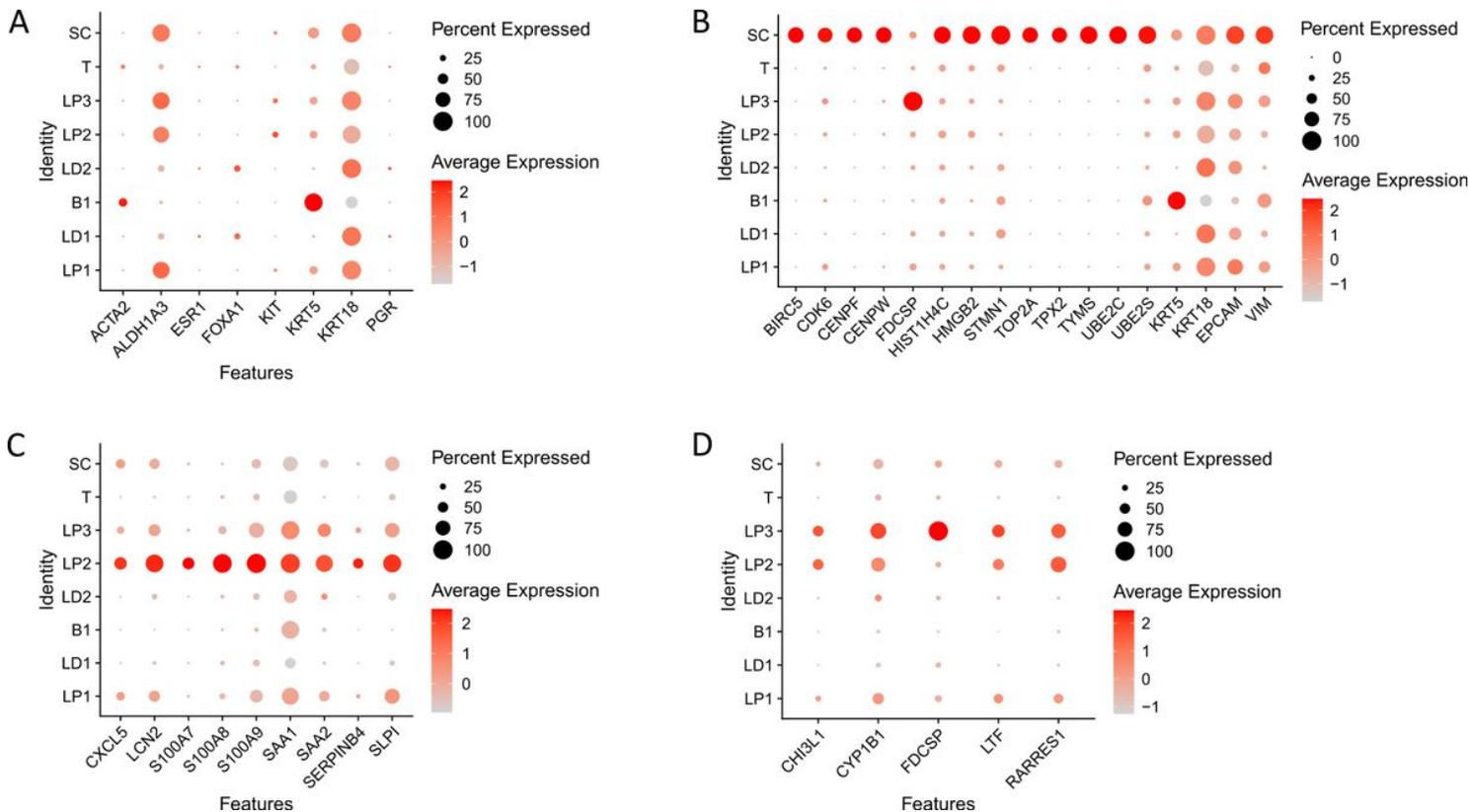
## Figures



**Figure 1**

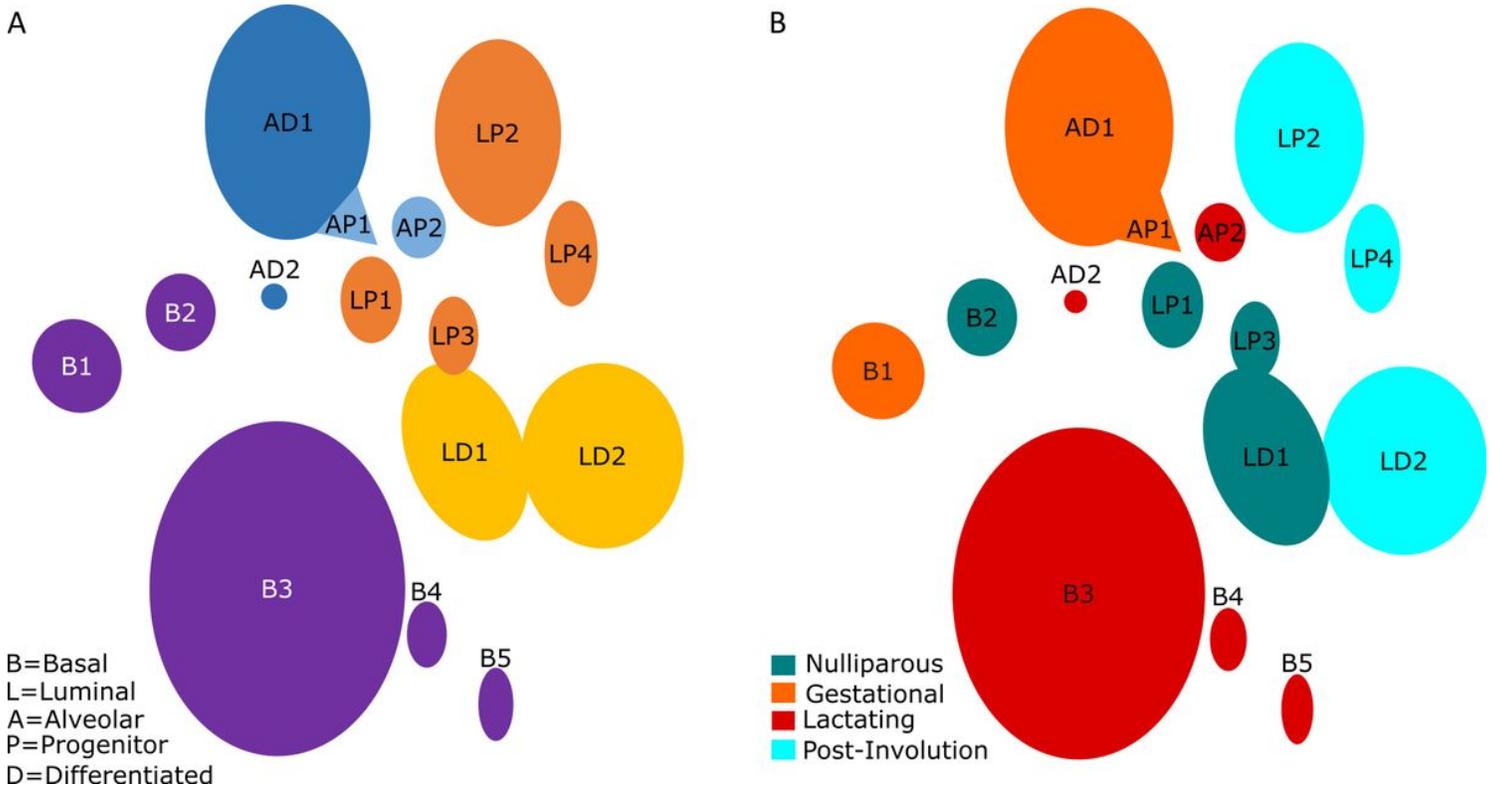
**Human adult mammary cell populations.** **A** t-SNE plot of seven normal breast scRNA-seq datasets. Individual dots correspond to cells and dot color indicates tissue sample. **B** Four cell clusters, luminal progenitor (LP), luminal differentiated (LD), basal (B), and stromal (Str) are identified at 0.05 cluster resolution in a t-SNE plot. **C** Eight cell clusters, three luminal progenitors (LP1-3), two luminal differentiated (LD1-2), a basal (B1), a stem cell like population (SC), a transition (T), and a stromal (Str)

cell cluster are identified at 0.14 cluster resolution in a t-SNE plot. **D** Expression of LP, LD, and B marker genes was analyzed in t-SNE plots. Grey indicates low or no gene expression while red indicates high expression. **E** Heatmaps of the top 20 and 10 differentially expressed genes at the 0.05 and 0.14 cluster resolution, respectively. Red indicates high expression while blue indicates low expression



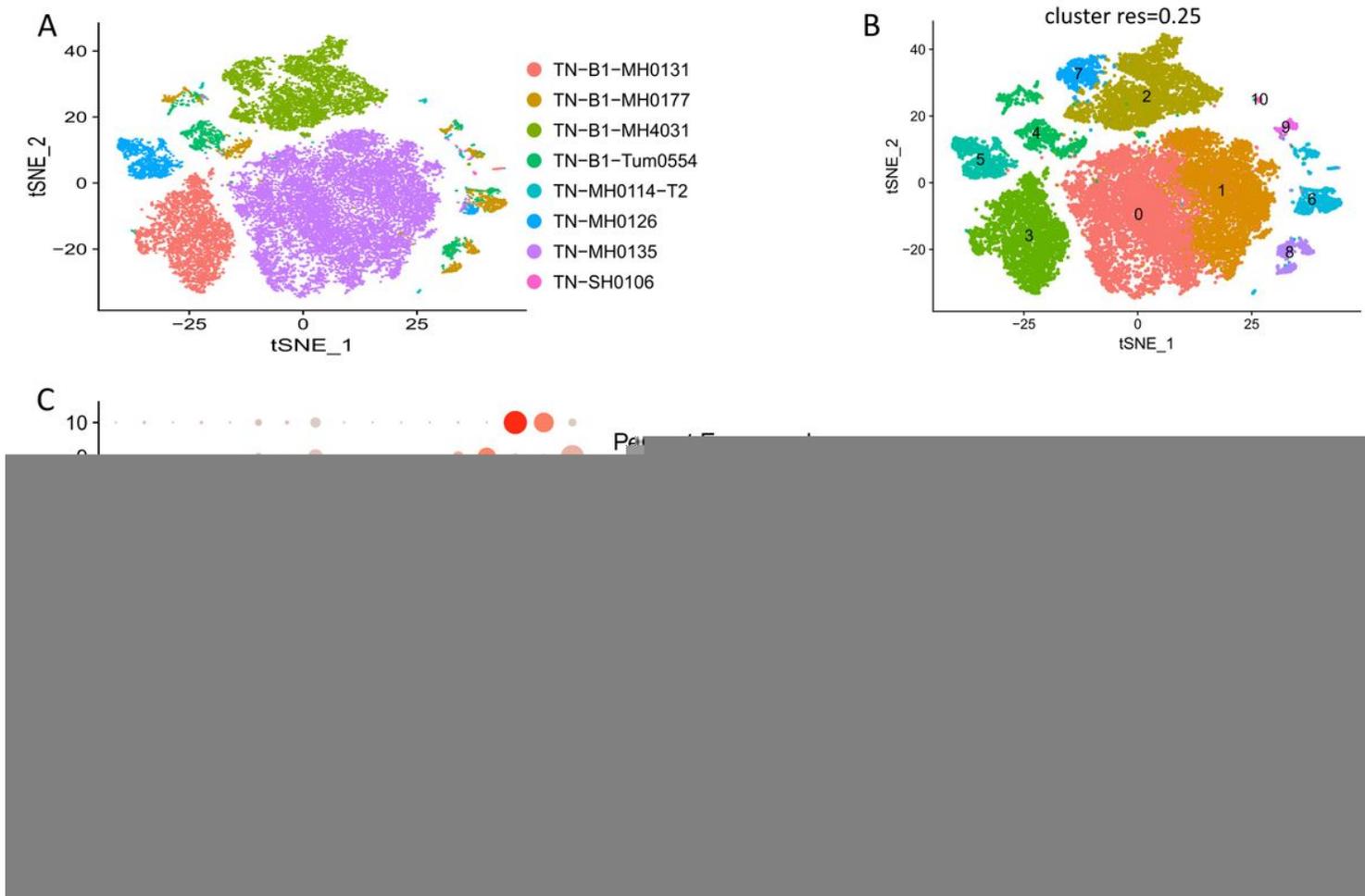
**Figure 2**

**Gene expression in human adult mammary cell populations.** Dot plots show the expression of population specific genes in the cell populations at 0.14 cluster resolution. **A** Mammary luminal progenitor, luminal differentiated, and basal cell marker expression in human adult mammary cell populations. **B** SC gene expression in human mammary cells, as well as a basal (KRT5), luminal (KRT18), epithelial (EPCAM), and mesenchymal (VIM) marker expression. **C** LP2 and **D** LP3 marker gene expression in human mammary cells



**Figure 3**

**Mammary cell populations in the adult mouse in nulliparous, gestational, lactating, and post-involution stages.** **A** 15 cell populations, basal (purple), differentiated alveolar (blue), alveolar progenitor (light blue), luminal progenitor (orange), and luminal differentiated (yellow) cell populations are shown. **B** 15 cell populations, nulliparous (teal), gestational (orange), lactating (red), and post-involution (turquoise) cell populations are shown. Based on the work of Bach et al [9]



**Figure 4**

**SC gene expression in eight human triple-negative breast cancer cell datasets.** **A** t-SNE plot of eight triple-negative breast scRNA-seq datasets. Individual dots correspond to cells and dot color indicates the dataset. **B** Ten cancer cell populations were identified at 0.25 cluster resolution. **C** Dot plot showing the expression of SC genes in the breast cancer cell populations, as well as a basal (KRT5), luminal (KRT18), epithelial (EPCAM), and mesenchymal (VIM) marker

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplInformationNewlyIdentifiedDiscoverOncolMcMullenandSoto031422.pdf](#)