

Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes

Alex Prat · Olga Karginova · Joel S. Parker · Cheng Fan · Xiaping He ·
Lisa Bixby · J. Chuck Harrell · Erick Roman · Barbara Adamo ·
Melissa Troester · Charles M. Perou

Received: 8 September 2013 / Accepted: 19 October 2013 / Published online: 27 October 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract Five molecular subtypes (luminal A, luminal B, HER2-enriched, basal-like, and claudin-low) with clinical implications exist in breast cancer. Here, we evaluated the molecular and phenotypic relationships of (1) a large in vitro panel of human breast cancer cell lines (BCCLs), human mammary fibroblasts (HMFs), and human mammary epithelial cells (HMECs); (2) in vivo breast tumors; (3) normal breast cell subpopulations; (4) human embryonic stem cells (hESCs); and (5) bone marrow-derived mesenchymal stem cells (hMSC). First, by integrating genomic data of 337 breast tumor samples with 93 cell lines we were able to identify all the intrinsic tumor subtypes in the cell lines, except for luminal A. Secondly, we observed that the cell lines recapitulate the differentiation

hierarchy detected in the normal mammary gland, with claudin-low BCCLs and HMFs cells showing a stromal phenotype, HMECs showing a mammary stem cell/bipotent progenitor phenotype, basal-like cells showing a luminal progenitor phenotype, and luminal B cell lines showing a mature luminal phenotype. Thirdly, we identified basal-like and highly migratory claudin-low subpopulations of cells within a subset of triple-negative BCCLs (SUM149PT, HCC1143, and HCC38). Interestingly, both subpopulations within SUM149PT were enriched for tumor-initiating cells, but the basal-like subpopulation grew tumors faster than the claudin-low subpopulation. Finally, claudin-low BCCLs resembled the phenotype of hMSCs, whereas hESCs cells showed an epithelial phenotype without basal or luminal differentiation. The results presented here help to improve our understanding of the wide range of breast cancer cell line models through the appropriate pairing of cell lines with relevant in vivo tumor and normal cell counterparts.

Alex Prat and Olga Karginova have contributed equally to this work.

The accession numbers for the microarray data are GSE18229 (UNC337) and GSE50470 (UNC105).

Electronic supplementary material The online version of this article (doi:10.1007/s10549-013-2743-3) contains supplementary material, which is available to authorized users.

A. Prat · O. Karginova · J. S. Parker · C. Fan · X. He ·
J. C. Harrell · B. Adamo · C. M. Perou (✉)
Lineberger Comprehensive Cancer Center, University of North
Carolina at Chapel Hill, 450 West Drive, CB7295, Chapel Hill,
NC 27599, USA
e-mail: cperou@med.unc.edu

A. Prat · O. Karginova · J. S. Parker · C. M. Perou
Department of Genetics, University of North Carolina at Chapel
Hill, Chapel Hill, NC 2759, USA

A. Prat · B. Adamo
Translational Genomics Group, Vall d'Hebron Institute of
Oncology (VHIO), 08035 Barcelona, Spain

Keywords Breast cancer · Cell lines · Intrinsic subtype · Stem cell · Tumor-initiating cell

L. Bixby
Flow Cytometry Core Facility, University of North Carolina at
Chapel Hill, Chapel Hill, NC 2759, USA

E. Roman · M. Troester
Public Health and Epidemiology, University of North Carolina at
Chapel Hill, Chapel Hill, NC 2759, USA

C. M. Perou
Department of Pathology & Laboratory Medicine, University of
North Carolina at Chapel Hill, Chapel Hill, NC 2759, USA

Introduction

Global messenger RNA expression analyses of human breast cancers have established five “intrinsic” molecular subtypes: luminal A, luminal B, basal-like, HER2-enriched, and the recently characterized claudin-low group [1]. These molecular entities have shown significant differences in incidence, survival, and responsiveness to therapies [1–4], and their information complements and expands the information provided by the classical clinical–pathological markers [5–8]. Importantly, studies focused on intrinsic molecular subtyping are improving our understanding of the biologic heterogeneity of breast cancer and its developmental cell(s) of origin [1, 9–11].

Although, the ideal preclinical study should be performed with human tumor samples that represent the complete spectrum of the disease, this type of research is being hampered, in part, by the lack of appropriate *in vivo* assays. Complementary to this approach are *in vitro* studies focused on tumor- or normal tissue-derived cell lines, all of which are being extensively used by the breast cancer research community [12]. Many of these cell lines have served as model systems to either dissect the biology of breast cancer and/or develop novel treatment strategies that are further tested in patients. In some cases, these studies have led to improvements for cancer patients. For example, the estrogen receptor (ER)-positive MCF-7 cell line has been useful for the study of the estrogen pathway and the development of efficacious anti-hormonal therapies such as tamoxifen [13, 14], while HER2-amplified SKBR3 and BT474 cell lines have helped to elucidate various mechanisms of resistance to anti-HER2 therapies [15, 16]. However, these preclinical studies have had limited impact in the management of breast cancer patients [17, 18], partly due to the incomplete understanding of the similarities and differences between these *in vitro* model systems and their relevant *in vivo* tumor counterparts.

Previous work has shown that the main genetic and transcriptional features of breast tumors are present in cell lines [19–22]. In 2006, Neve et al. [19] identified two major groups (basal and luminal) in a panel of ~50 breast cancer cell lines by independently comparing the global expression profiles of cell lines and primary breast tumors. Interestingly, the basal cluster was further subdivided into two subgroups: basal-A, which resembled the basal-like signature in primary breast tumors [2, 3] and basal-B, which exhibited a mesenchymal and a cancer stem cell (CSC)-like profile that was less similar to primary basal-like tumors. The identification of the basal-B group has been confirmed by three other groups [21–23], with one group calling them normal-like [23]. More recently, we and others have shown that a subgroup of 9 (MDA-MB231, SUM159PT, MDA-MB157, BT549, SUM1315MO2, MDA-MB436, MDA-MB435, HBL100 and Hs578T) of

the 12 basal-B breast cancer cell lines best resemble the recently characterized claudin-low tumor subtype [1, 24]. However, it is still unknown if all the intrinsic subtypes are represented in cell lines.

In the human mammary gland, four subpopulations of cells have been identified and functionally characterized [9]. By utilizing a combination of fluorescence-activated cell sorting (FACS) with EpCAM and CD49f cell surface markers and a series of *in vitro* and *in vivo* experiments, Lim et al. [9] observed that the normal breast tissues have at least four subpopulations enriched with mammary stem cells/bipotent progenitors (MaSC/BiPs), luminal progenitors (pLs), mature luminal cells (mLs), and stromal cells (after excluding lineage positive cells, i.e., lymphocytes, red blood and endothelial cells). Using Lim et al.’s [9] gene expression data, we subsequently reported a differentiation model that tracks the epithelial differentiation hierarchy (MaSC/BiP → pL → mL) and is prognostically relevant. More importantly, we showed that the tumor intrinsic subtypes recapitulate the normal breast epithelial differentiation hierarchy, where claudin-low tumors and cell lines are the most similar to the MaSC/BiPs [1, 10]. These and other findings have led to new hypotheses regarding the potential cell of origin and/or transformation of the different breast cancer subtypes [10, 25, 26]. However, it is unknown where other cell lines, including normal human mammary epithelial cells (HMECs), fall into this hierarchy. Still less is known about the relationship of adult human mesenchymal stem cells (hMSCs) and embryonic stem cells (hESCs) to different breast tumor subtypes and cell lines.

In this report, we evaluated a large *in vitro* panel of breast cell lines and compared their features with (1) tumors, (2) four cell subpopulations of the normal breast, and (3) hMSC and hESC. Specifically, we show that all of the tumor subtypes except the luminal A and normal breast-like are well represented in cell lines. In addition, we observed that the cell lines recapitulate many of the features of each normal breast cell subpopulation identified using FACS.

Results

Molecular comparison between cell lines and breast samples

To compare the global gene expression profiles of cell lines with *in vivo* breast samples, we profiled a panel of *in vitro* breast cancer cell lines (BCCLs), human mammary fibroblasts (HMFs), and HMECs, and combined the microarray data (UNC105, GSE50470) with our previously published

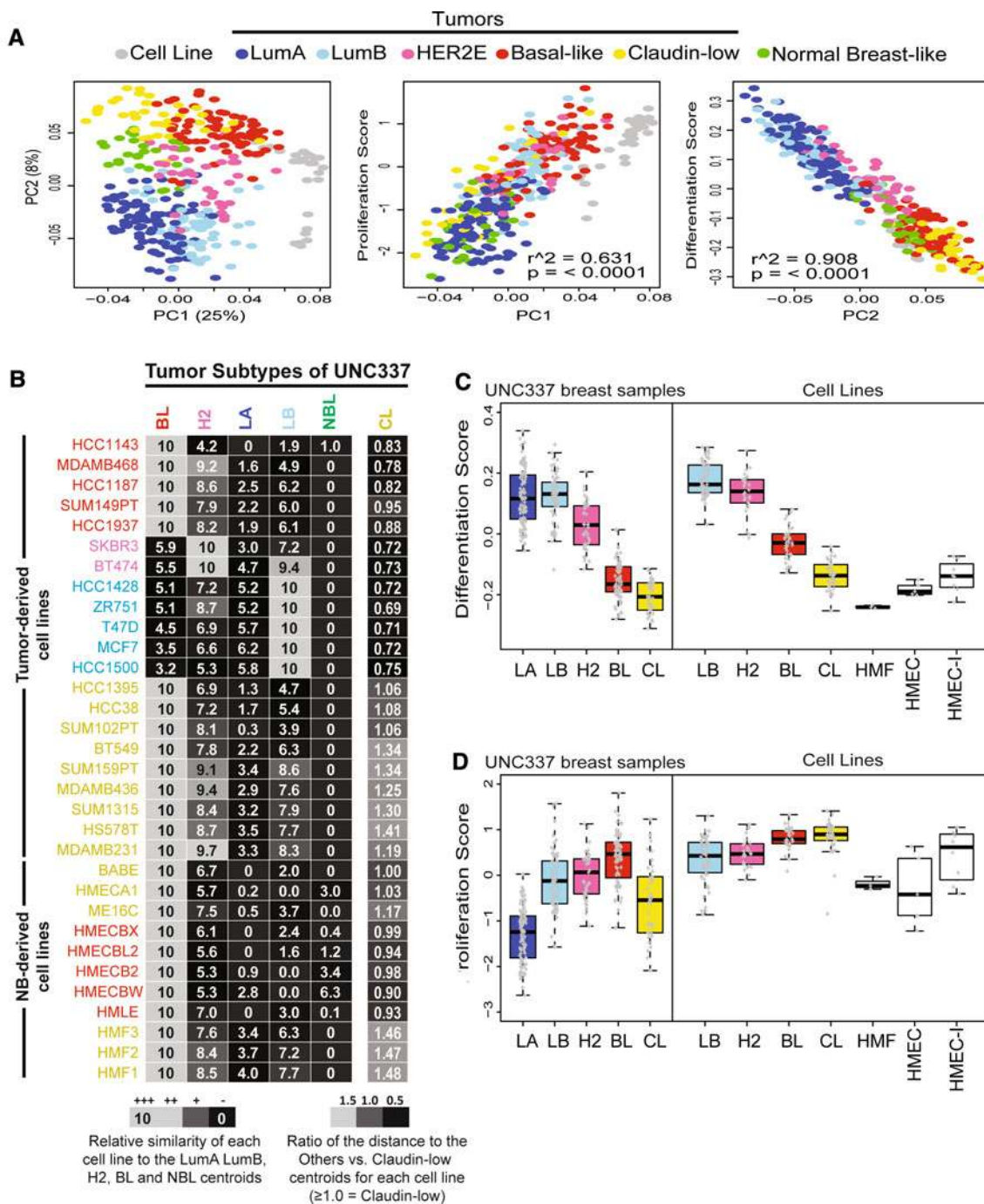


Fig. 1 Combined gene expression data of cell lines, breast tumors and normal breast samples. **a** Principal component (PC) 1 and 2 loading plots and their correlation with proliferation and differentiation scores, respectively. Samples included here were the entire UNC337 data set and all the UNC breast cancer cell lines (UNC105). **b** Euclidean distance of selected UNC cell lines to the UNC337 tumor intrinsic subtypes, including the claudin-low tumor type. **c** Differentiation and **d** proliferation scores of the tumor (*left*) and cell lines

(*right*) grouped by their molecular subtype. Cells derived from normal breast tissue are shown in the following categories: HMF, HMECs and immortalized HMECs (I-HMECs). Subtype calls: luminal A (LA) *dark blue*; luminal B (LB) *light blue*; HER2-enriched (H2) *pink*; basal-like (BL) *red*; claudin-low (CL) *yellowish*; normal breast-like (NBL) *green*; cell line *gray*. Replicate arrays done in the same cell line have been omitted, prioritizing first the UNC samples followed by the Neve et al. samples when appropriate

UNC337 data set that is composed of 320 breast tumors and 17 normal breast samples (GSE18229) [1]. In this combined data set (Fig. 1a), the gene expression loading plot of the

principal components 1 (PC1) and 2 (PC2) discriminate the tumor intrinsic subtypes with the vast majority of cell lines being displayed toward high PC1. Interestingly, we found

PC1 to be correlated with expression of proliferation-related genes ($r^2 = 0.634$, $p < 0.0001$), while PC2 was highly correlated to our previously described differentiation genomic model (MaSC/BiP \rightarrow pL \rightarrow mL axis) [1] ($r^2 = 0.908$, $p < 0.0001$). Thus, this data suggests that the majority of gene expression variation occurring in breast tumors and normal samples is explained by two variables (1) differentiation status along a MaSC/BiP \rightarrow pL \rightarrow mL axis and (2) proliferation.

To determine which intrinsic molecular profile each cell line best resembles, we calculated the distances of each cell line to the luminal A, luminal B, basal-like, HER2-enriched tumor centroids, and the normal breast-like group in the combined tumor-cell line data set; we also applied the 9-cell line claudin-low predictor to identify claudin-low cell lines. As shown in Fig. 1b and Supplemental material, all the intrinsic molecular subtypes were identified in tumor- and normal breast-derived cell lines except for the luminal A subtype and the normal breast-like group. Interestingly, all HMFs were identified as claudin-low, whereas HMECs were called either basal-like (i.e., HMLE, HMECBX) or claudin-low (i.e., ME16C, HMECA1). However, all HMECs showed borderline significance (i.e. ratios close to 1.0) for being identified as claudin-low, thus suggesting that these normal cell-type lines have mixed claudin-low/basal-like characteristics.

We then combined our cell line-tumor data set with three previously published cell line data sets [19, 21, 22] for a total of 93 different microarrays of cell lines. As expected, we observed the same findings; namely, no cell line was identified as luminal A or normal breast-like. However, among the 52 cell lines evaluated with at least one replicate sample among the four data sets, seven (13 %) discrepancies in the subtype calls were observed. Nonetheless, these data highlight that in virtually all cases, breast cancer cell lines maintain their overall genomic profile even if cultured in different laboratories and assayed on different platforms.

We have previously shown that BCCLs can be discriminated by their differentiation status along a MaSC/BiP \rightarrow pL \rightarrow mL axis [1]. To determine if this is also valid when a larger panel of cell lines is evaluated together with in vivo breast samples, we determined the differentiation scores of the combined tumor and cell line data set. Interestingly, we observed that the cell lines recapitulate all the differentiation states observed in tumors (Fig. 1c), which is consistent with the fact that we observed all of the molecular subtypes in cell lines (except for the luminal A and the normal breast-like). As previously reported and as shown in Fig. 1c, luminal A and B tumors are indistinguishable based on their differentiation score status, whereas proliferation (i.e., PC2) is one of the main differences between them (Fig. 1d). Evaluation of the

proliferation status revealed that BCCLs show higher expression of proliferation-related genes compared to their in vivo tumor subtype counterpart (Fig. 1d, $p < 0.0001$, Student *t* test). As expected, primary HMECs and HMFs showed lower expression of proliferation-related genes compared to BCCLs, although still higher than luminal A tumors.

Finally, to further understand the differences between tumors and cell lines, we identified those genes that are either up- or down-regulated in all cell lines when compared to their respective tumor counterpart (Supplemental material). As expected, the down-regulated cell line specific genes ($n = 244$) were mainly involved in immune response (i.e., lymphocyte antigen 96 and adenosine deaminase) and extracellular matrix (i.e., adrenomedullin), concordant with a lack of lymphocytic and stromal components in these in vitro cultures. Interestingly, removing these potential cell-type or heterotypic interaction specific genes from the tumor and cell line combined data set did not substantially affect the differentiation status of the tumors (Supplemental material), suggesting that the differentiation predictor is mainly tracking the in vivo biology of the tumor epithelial cells themselves.

Characterization of stromal and epithelial FACS subpopulations of the normal breast

To understand the relationship between cell lines and the different subpopulations of the normal breast, we FACS-sorted and expression-profiled stromal, MaSC/BiP, pL, and mL subpopulations from at least three independent reduction mammaplasties as previously described by Lim et al. [9] (Fig. 2a). As shown in Fig. 2b and Supplemental material, ~ 35 and ~ 25 % of lineage-negative cells (i.e., CD45-negative and/or CD31-negative) fall in the stromal- and MaSC/BiP-enriched groups, respectively, whereas only ~ 10 % of the cells are identified in the pL and mL-enriched groups. To further complement these results, we also analyzed six independent reduction mammaplasties using CD24 and CD44 cell surface markers. As shown in Fig. 2c, 50, 22, and 4.6 % of the sorted cells were CD24⁻/CD44⁺, CD24⁻/CD44⁻, and CD24⁺/CD44⁺ in one sample. However, the CD24⁺/CD44⁺ subpopulation was not always identified (data not shown).

Our previously described differentiation score predictor was based on genomic data of the MaSC/BiP-, pL- and mL-enriched subpopulations of Lim's et al. [9] (i.e., training set). To validate this differentiation model, we estimated the differentiation status of our newly sorted cell subpopulations (Fig. 2d). As expected, the MaSC/BiPs showed the lowest scores (mean -0.350 ± 0.047), the mLs showed the highest scores (mean 0.625 ± 0.0328), and the pLs showed a slightly higher differentiation status

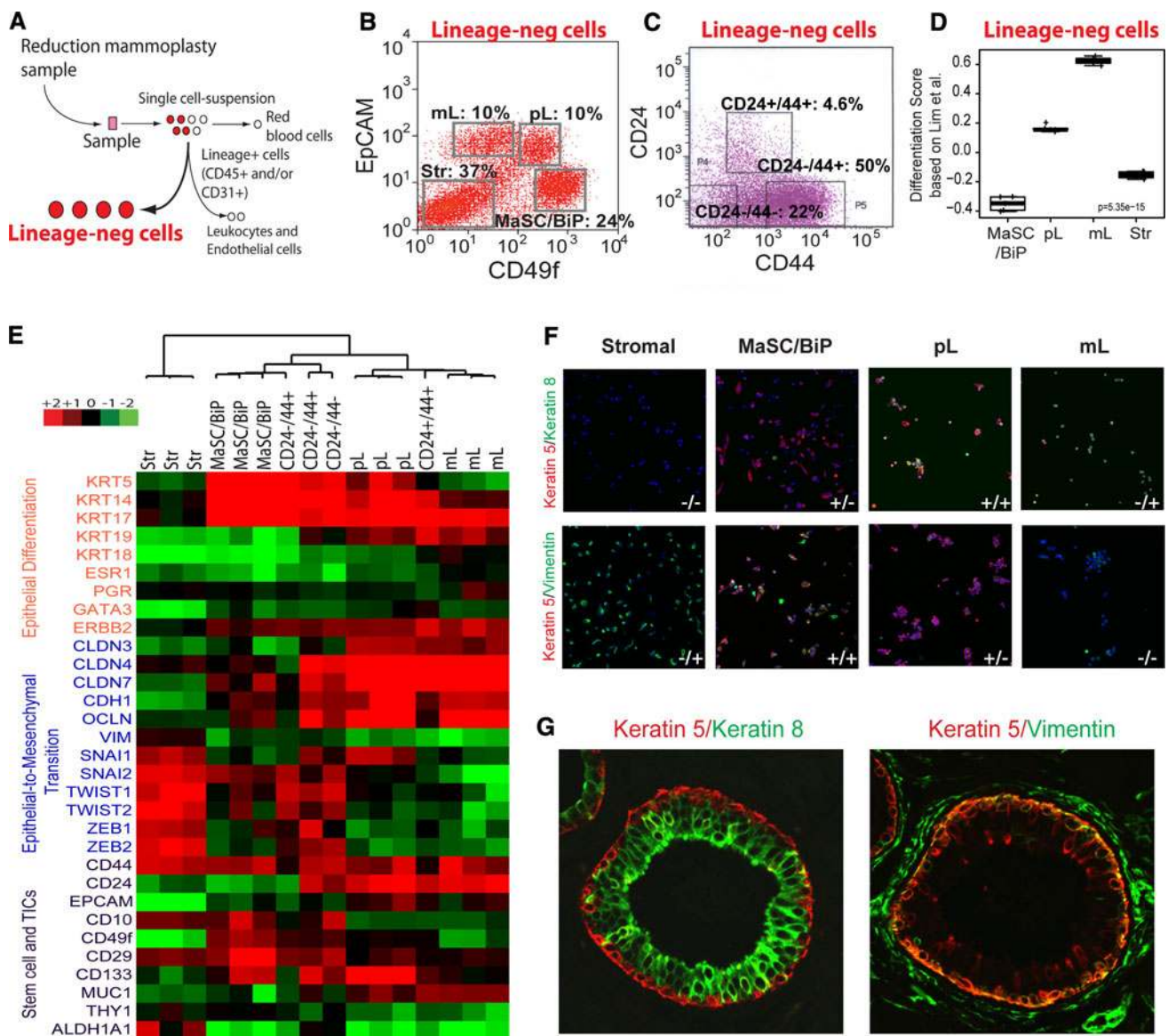


Fig. 2 Characterization of mesenchymal and epithelial subpopulations of the normal breast. **a** Diagram summarizing the processing and steps taken for FACS of normal breast tissue. **b, c** Distribution of the lineage-negative cell subpopulations in one representative reduction mammoplasty sample using EpCAM/CD49f and CD24/CD44 surface markers. **d** Differentiation score of the four-sorted cell subpopulations of at least three reduction mammoplasties samples. *p* values shown here have been calculated by comparing gene expression means across all subpopulations. **e** Supervised hierarchical clustering of the sorted subpopulations based on the expression of a panel of markers of basal and luminal differentiation, EMT and CSCs markers. Each colored square of the heatmap represents the relative transcript abundance (in log₂ space) for each cell fraction with highest

expression being red, average expression being black, and lowest expression being green. Keratins 5 [KRT5], 14 [KRT14] and 17 [KRT17], 18 [KRT18] and 19 [KRT19]; ER (ESR1); progesterone receptor (PGR); HER2 (ERBB2); vimentin [VIM]; snail-1 [SNAI1]; snail-2 [SNAI2]; Zinc finger E-box homeobox 1 and 2 [ZEB1 and ZEB2]; E-cadherin [CDH1]; Claudins -3 [CLDN3], -4 [CLDN4] and -7 [CLDN7]; prominin 1 [CD133]; epithelial cell-adhesion molecule [EpCAM]; mucin 1 [MUC1]; integrin alpha 6 [CD49f]; integrin beta 1 [CD29]; membrane metallo-endopeptidase [CD10]; aldehyde dehydrogenase family 1, subfamily A1 [ALDH1A1]. **f, g** Immunofluorescent staining of the four breast cell subpopulations and normal breast ducts using antibodies against vimentin (green), keratin 5 (red) and keratin 8 (green)

(0.16 ± 0.0237) than Lim's pL (trained to be zero). Overall, high correlation coefficients (>0.90 , $p < 0.0001$) between EpCAM and CD49f gene and protein expression was observed in the four normal-sorted fractions (Supplemental material).

To further characterize the four normal breast FACS subpopulations, including the three CD24/CD44-sorted fractions, we evaluated the expression of markers associated with epithelial and luminal differentiation, epithelial-to-mesenchymal transition and stem cell and/or tumor-

initiating cells (TIC) as described in Prat et al. [1]. As observed (Fig. 2e) hierarchical clustering of the FAC-sorted samples with these selected gene expression values revealed three main groups: stromal, MaSC/BiP, and pL/mL. Interestingly, the MaSC/BiP-enriched subpopulation is characterized by (1) the highest expression of basal keratins (KRT17, KRT14 and KRT5) and CD49f compared to the other subpopulations ($p < 0.0001$, Student's t test), and (2) a lower expression of luminal keratins and epithelial cell-adhesion genes such as CLDN7 and CDH1 compared to the pL- and mL-enriched subpopulations ($p < 0.0001$, Student's t test). Interestingly, the CD24⁻/CD44⁺ and CD24⁻/CD44⁻ cell fractions clustered with the MaSC/BiP-enriched group, while the CD24⁺/CD44⁺ subpopulation clustered with the pL/mL subpopulations, suggesting that a greater homogeneity can be obtained with CD49f/EpCAM combination of markers. Finally, the stromal-enriched subpopulation showed a lack of expression of epithelial markers and cell–cell-adhesion genes, with high expression of stromal markers (i.e., vimentin) and transcription factors such as ZEB1 and SNAI2.

Further analyses of four normal breast FACS subpopulations by immunofluorescent (IF) staining with antibodies against basal (KRT5), luminal (KRT8), and stromal (VIM) markers confirmed these findings (Fig. 2f), although certain heterogeneity within each sorted subpopulation was also observed. Finally, IF imaging of normal breast ducts revealed that the majority of cells within the stromal-enriched group (VIM+/KRT5–/KRT8–) are found in the stroma, the MaSC/BiP-enriched cells (VIM+/KRT5+/KRT8–) are found in the basal/myoepithelial layer and, finally, the pL (VIM–/KRT5+/KRT8+) and mL (VIM–/KRT5–/KRT8+) cells are found in the luminal layer of the duct (Fig. 2g).

Cell lines recapitulate the differentiation hierarchy of the normal breast

To determine the transcriptomic similarities between the normal breast subpopulations (stromal, MaSC/BiP, pL, and mL) and cell lines in 2D culture, including HMFs and HMECs, we first calculated a signature enrichment score for each of the centroids using the Lim et al. [9] microarray data, and we included in this analysis our normal breast-sorted fractions as controls. As shown in Fig. 3a, HMFs showed the highest enrichment for the stromal signature suggesting that the vast majority of sorted EpCAM⁻CD49f^{-low} cells from normal breast tissue are indeed fibroblasts. On the other hand, the MaSC/BiP signature was found almost uniquely enriched in HMECs. We identified 1,530 genes that are similarly expressed between HMECs and MaSC/BiPs [significance analyses of microarrays (SAM) one-class, FDR = 0 %; Supplemental

material]. Among the up-regulated genes, we observed basal keratins 5/14/17, p63, CD49f, and CD44. Conversely, claudin-low cell lines showed an intermediate differentiation status between the MaSC/BiP and the stromal state. This is consistent with our previous report showing that, compared to basal-like and luminal BCCLs, claudin-low BCCLs are closer to the MaSC/BiP centroid than to the pL or mL centroids [1]. However, here we show that the claudin-low cells are uniquely enriched with stromal-like biological processes similar to HMFs. Both the stromal fraction of the normal breast and claudin-low BCCLs were found to share similar expression of 1,334 genes (SAM one-class, FDR = 0 %; Supplemental material). Among the 1,097 down-regulated genes, we observed genes associated with cell–cell adhesion (i.e., Claudin -3, E-cadherin, and desmoplakin), epithelial markers (EPCAM and CD24), and markers of luminal differentiation (GATA3, FOXA1 and keratins -8, -18, and -19). Finally, basal-like and luminal B BCCL genomic profiles showed to be the closest to the pL and mL centroids, respectively, whereas the differentiation status of the HER2-enriched cell lines was found to be between the pL and mL state.

To characterize cell lines from a phenotypic perspective, we performed FACS on a panel of 30 cell lines, representing all the molecular subtypes using established stem and/or CSC surface markers including EpCAM, CD49f, CD24, and CD44. As shown in Fig. 3b and in Supplemental data, claudin-low cell lines and HMFs express a stromal/MaSC phenotype with the majority of cells within these cell lines being EpCAM^{-low}/CD49f⁺ and CD24^{-low}/CD44⁺. Conversely, all of the evaluated HMECs, regardless if they were immortalized or not, showed a MaSC/BiP phenotypic profile (EpCAM^{-low}/CD49f^{high}), which is consistent with the gene expression data. In terms of proposed CSC markers, HMECs showed a non-CSC phenotype (CD24⁺/CD44⁺); except for cells from the immortalized ME16C cell line that showed a CD24⁻/CD44⁺ phenotype. Moreover, the majority of cells within most basal-like cell lines showed a pL/mL and a non-CSC phenotypes (EpCAM⁺/CD49f⁺ and CD24⁺/CD44⁺), and the vast majority of luminal B and HER2-enriched cell lines also showed a pL–mL and a non-CSC phenotypes. However, it is important to note that FACS alone did not fully discriminate basal-like from luminal/HER2-enriched cell lines, since CD49f levels were variable.

Finally, we performed IF staining analysis of cell lines using the same stromal and epithelial markers as previously done for the normal breast. As shown in Fig. 3c and Supplemental material, all claudin-low cell lines showed strong membrane staining for vimentin and little to no keratin 5 and 8 positivity (i.e., MDA-MB231, SUM159PT) similar to HMFs, or keratin 5-positivity only (i.e., HCC38, HCC1395). The majority of basal-like cell lines, however, contained a

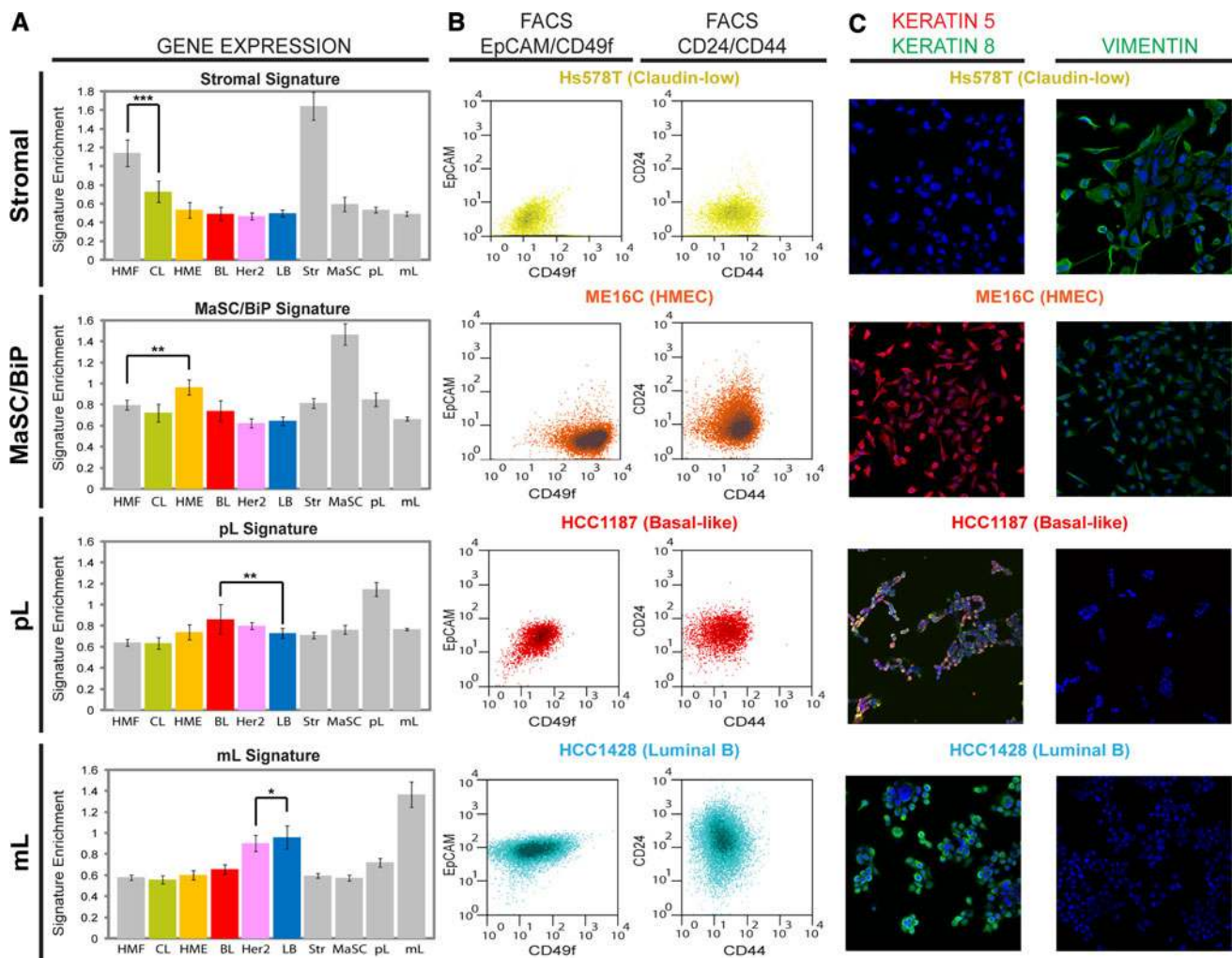


Fig. 3 Cell lines recapitulate the differentiation hierarchy of the normal breast. **a** Enrichment scores of Lim’s stromal, MaSC/BiP, pL, and mL gene signatures in BCCLs grouped by their molecular subtypes, HMFs, HMECs, and our normal breast FAC-sorted subpopulations. **p* value < 0.05, ***p* value < 0.001. Replicate arrays done in the same cell line have been omitted, prioritizing first the UNC samples followed by the Neve et al. samples when appropriate. Signature enrichment scores for each Lim et al. fraction (stromal, MaSC/BiP, pL and mL) has been obtained by calculating the distance of each cell line to two Lim et al. centroids: “others” versus “each

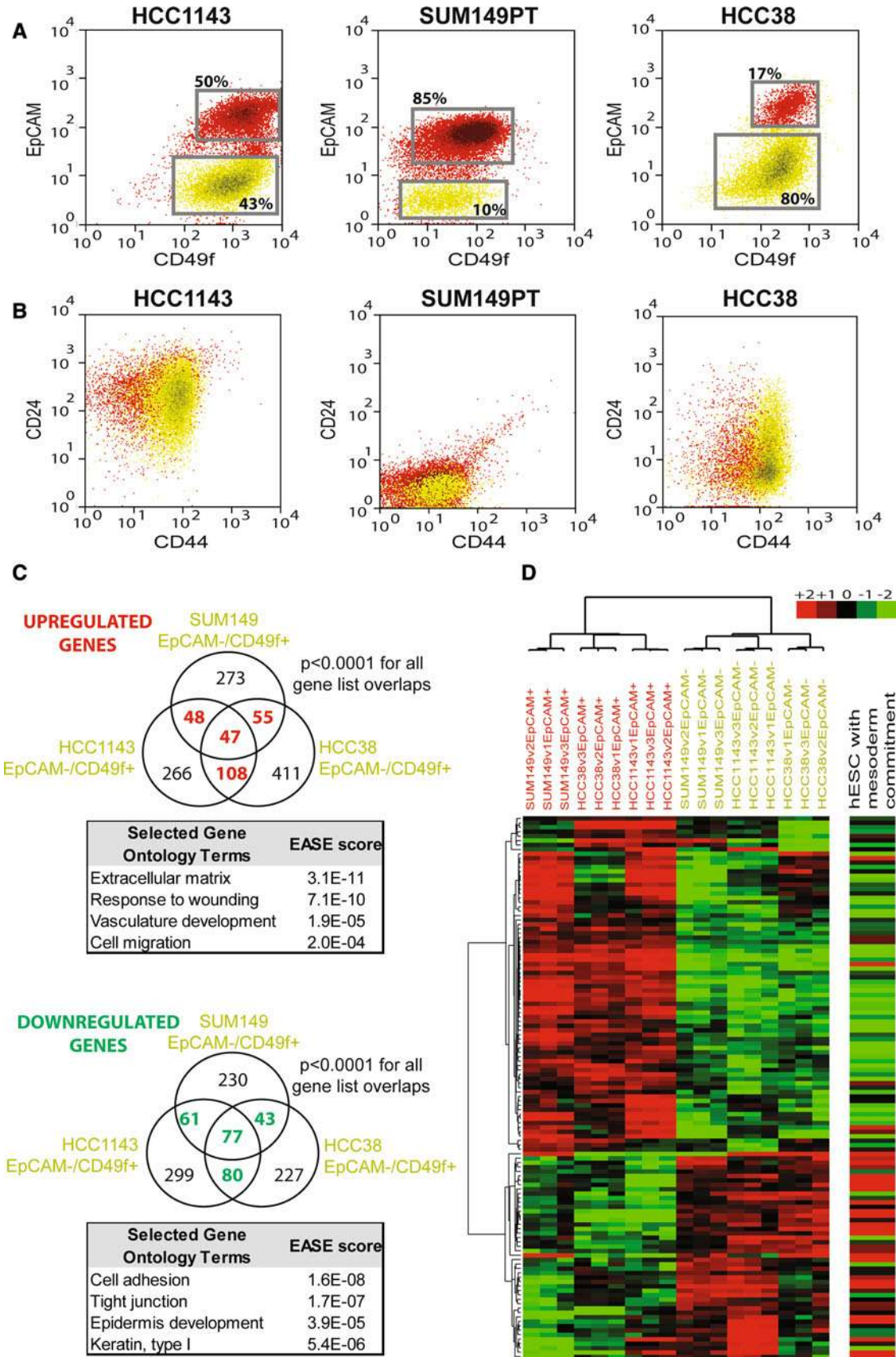
mixed population of cells where some had only basal keratin positivity and others had dual basal/luminal keratin positivity. However, none of these basal-like cells showed strong vimentin positivity, concordant with their epithelial state. Furthermore, luminal B and HER2-enriched cell lines showed strong positivity for luminal keratin 8 and the majority had no vimentin or keratin 5 staining; exceptions include the HER2-enriched cell lines BT474 and SKBR3 and luminal B cell line HCC1500 that showed some expression of keratin 5. Finally, we observed that the primary and immortalized HMECs showed strong positivity for keratin 5, positivity for vimentin, and varying degrees of

Lim et al. fraction”. In the plot, the ratio of the “others” distance versus “each Lim et al. fraction” distance is shown. **b** EpCAM/CD49f and CD24/CD44 FACS of claudin-low (Hs578T), HMEC (HMECBL), basal-like cell line (HCC1187) and luminal B (HCC1428) cell lines. **c** IF staining of Hs578T, HMECBL, HCC1187 and HCC1428 cell lines using antibodies against vimentin (green), keratin 5 (red) and keratin 8 (green). The complete FACS and IF staining data of all cell lines evaluated can be obtained in Supplemental material

keratin 8 protein expression, concordant with a partial epithelial-to-mesenchymal state in these cells similar to the one observed in the MaSC/BiP (EpCAM⁺/CD49f⁺) FACS subpopulation.

Distinct subpopulations of cells within Claudin-low and basal-like cell lines

Among the cell lines evaluated, SUM149PT (basal-like), HCC1143 (basal-like), and HCC38 (claudin-low) showed two clear separate cell subpopulations based on the levels of EpCAM and CD49f surface markers (Fig. 4a). As we



◀ **Fig. 4** Genomic analyses of distinct cell subpopulations within basal-like and claudin-low cell lines. **a** Expression of EpCAM/CD49f in HCC1143 (basal-like), SUM149PT (basal-like) and HCC38 (claudin-low) cell lines. The gates shown in each cell line (*gray squares*) represent the different sorted subpopulations that were further evaluated. **b** Expression of CD24/CD44 in the three cell lines. The *colors* represent the distribution of the sorted fractions in **(a)**. **c** Overlap of genes differentially expressed between EpCAM^{-low}/CD49f⁺ and EpCAM⁺/CD49f⁺ cells across HCC1143, SUM149PT and HCC38. *p* values denote the probability of the overlap being by chance. Below each Venn diagram, the up- and down-regulated gene ontology (GO) terms are shown. Each list included the genes that overlapped between at least two cell fractions (*red* or *green* genes). **d** Supervised hierarchical clustering of the sorted subpopulations based on the expression of the differentially expressed genes between EpCAM^{-low}/CD49f⁺ and EpCAM⁺/CD49f⁺ cells across HCC1143, SUM149PT and HCC38. On the right, relative expression data in hESC cells that have acquired a mesodermal state. All the gene lists and the clustering can be obtained in Supplemental material

previously reported for the SUM149PT cell line [1], molecular subtyping of the EpCAM^{-low}/CD49f⁺ and EpCAM⁺/CD49f⁺-sorted fractions of HCC38 and HCC1143 cell lines confirmed that they are claudin-low and basal-like, respectively. Intriguingly, we could not identify these subpopulations of cells using CD24/CD44 surface markers. All HCC1143 cells showed high levels of the luminal marker CD24 when compared to SUM149PT and HCC38, where the majority of cells showed a CD24^{-low}/CD44⁺ phenotype (Fig. 4b).

To identify potential genes and/or biological processes responsible for the two states of differentiation in SUM149PT, HCC1143, and HCC38 cell lines, we performed a paired two-class SAM analysis between the EpCAM^{-low}/CD49f⁺ and EpCAM⁺/CD49f⁺-sorted fractions within each cell line, and identified the overlapping genes across all three cell lines. As shown in Fig. 4c, the numbers of up- and down-regulated genes appearing in all comparisons exceeded the chance of overlap expected ($P < 0.0001$, exact hypergeometric probability). Among them, we observed high expression in the EpCAM^{-low}/CD49f⁺ fraction of genes involved in “response to wounding” (adrenomedullin and chemokine [C-X-C motif] ligand 1), “extracellular matrix” (collagen, type V, alpha 1, and laminin, gamma 1), and “migration” (interleukin 6 and fibronectin 1). Conversely, we observed low expression in the EpCAM^{-low}/CD49f⁺ fraction of genes involved in tight junction (Claudin 3, 4, and 7 and par-6 partitioning defective 6 homolog alpha) and keratinization (keratins 14, 15, 16, 19).

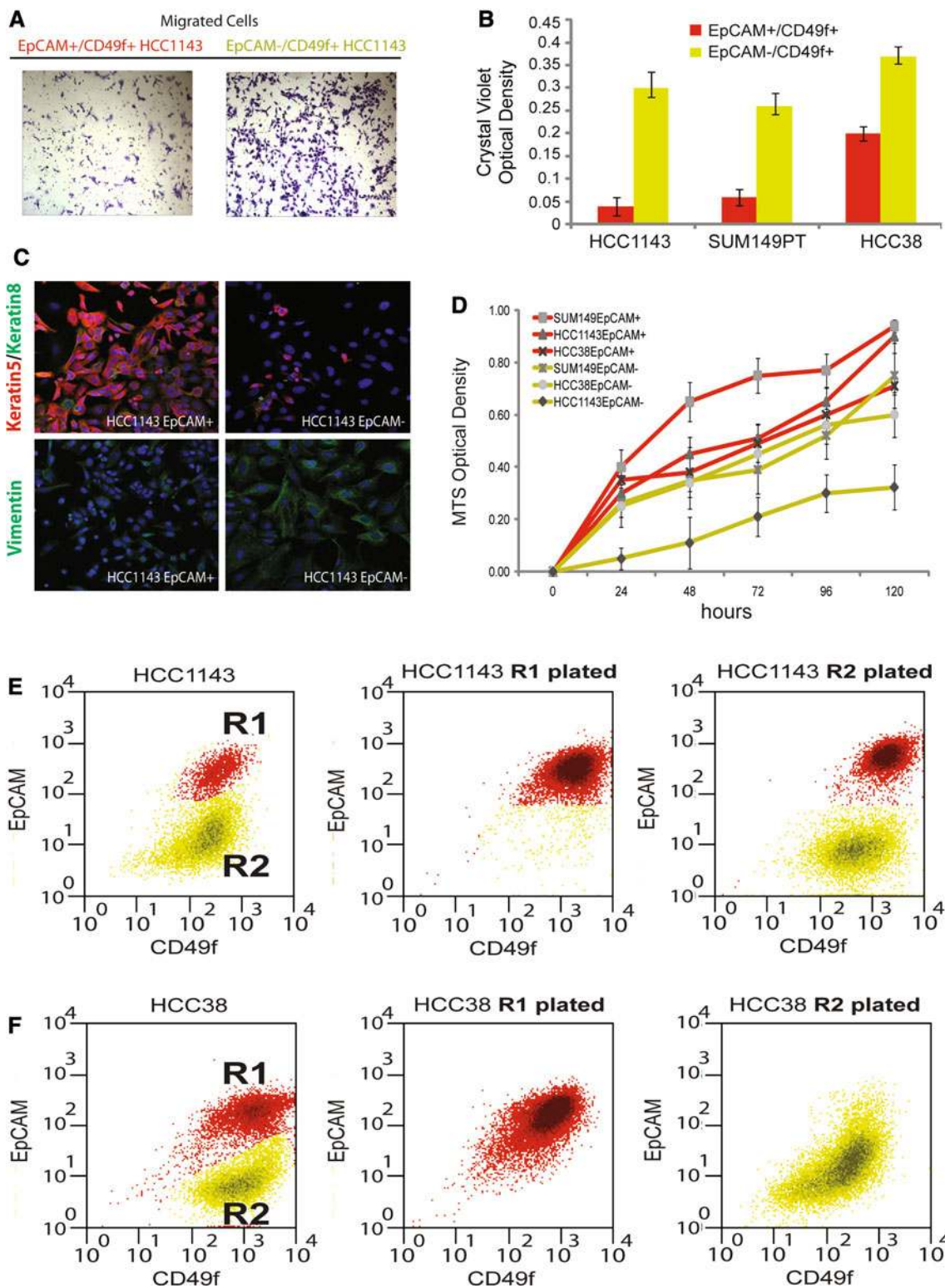
Consistent with these findings, EpCAM^{-low}/CD49f⁺ fractions from the three cell lines showed increased in vitro transwell migration, and HCC1143 and HCC38 EpCAM^{-low}/CD49f⁺ showed lower protein expression of keratins and higher vimentin expression compared to their respective EpCAM⁺/CD49f⁺ cells (Fig. 5a–c and Supplemental material). Interestingly, decreased proliferation during

5 days of growth was observed in the EpCAM^{-low} cells compared to their respective EpCAM⁺ fractions (Fig. 5d) and this is similar to comparisons of human claudin-low tumors versus basal-like [1].

Finally, when EpCAM^{-low}/CD49f⁺ and EpCAM⁺/CD49f⁺ subpopulations within HCC1143 cell lines were sorted and plated separately, we observed that ~25 % of the EpCAM^{-low}/CD49f⁺ cells differentiated into EpCAM⁺/CD49f⁺ basal-like cells over a 14–18 day period, whereas the EpCAM⁺/CD49f⁺ basal-like cells seem to maintain their differentiated status during in vitro culture (Fig. 5e), similar to our previous findings in SUM149PT [1]. Interestingly, the majority of HCC38 EpCAM^{-low}/CD49f⁺ cells remained HCC38 EpCAM^{-low}/CD49f⁺ during the same period of time, concordant with decrease in asymmetric division and an overall claudin-low profile identified in this cell line.

Tumor initiating cell ability of the two cell subpopulations within SUM149PT cell line and basal-like xenograft WashU-WHIM2

To explore the TIC ability of the two subpopulation of cells in the SUM149PT cell line, we FAC-sorted EpCAM^{-low}/CD49f⁺ and EpCAM⁺/CD49f⁺ cells (Fig. 6a) and injected different numbers of cells (100, 1,000, and 10,000 cells) of each subpopulation into nude immunocompromised mice. Interestingly, the distinct fractions could not be determined by CD24/CD44 staining (Fig. 6b), despite the fact that the EpCAM^{-low}/CD49f⁺ fraction showed the expected stromal phenotype (Fig. 6c). The in vivo experiment revealed that over long-term growth conditions (>100 days), both fractions were enriched with TICs. However, the TICs in the EpCAM⁺/CD49f⁺ fraction grew tumors significantly faster than the TICs in the claudin-low fraction ($p < 0.0001$ at the 60 days' time-point, extreme limiting dilution analysis (ELDA) [27]; Fig. 6d, e). Thus, if a 1 month time-point was used, only the EpCAM⁺/CD49f⁺ fraction would be considered a TIC-containing fraction. Nearly identical results were obtained for EpCAM/CD49f FAC sort of the human triple-negative WashU-WHIM2 xenograft, which has been whole genome sequenced [28]. One notable difference between SUM149PT and the WashU-WHIM2 FACS profile was that in the latter no distinct cell subpopulations were found, instead a more homogeneous oval-shaped profile was seen (Fig. 6f). Similar findings were obtained when the two populations of WashU-WHIM2 FACS were compared; namely that the EpCAM⁺/CD49f^{+/high} population was the TIC-enriched fraction ($p = 0.0136$, ELDA; Fig. 6g) and showed stronger basal gene expression characteristics (data not shown). Thus, overall, both in a cell line and in a xenograft tumor, the TIC component was the highly proliferative and more basal-like EpCAM⁺/CD49f⁺ cells.



HMECs as model systems for the study of the Claudin-low transformation

Human mammary epithelial cells have been used as model systems to study the CSC transformation in breast cancer

[29–32] and our FACS and IF analyses presented here point to a MaSC/BiP-state and a non-CSC phenotype in these cells. To better understand the features of the transformation of HMECs, we first expression profiled the previously published HMLE cell line (a non-transfected

◀ **Fig. 5** Functional analyses of distinct cell subpopulations within basal-like and claudin-low cell lines. **a, b** Trans-well migration capability of the EpCAM^{low}/CD49⁺ and EpCAM⁺/CD49⁺ cell fractions within each cell line. Microscopic image ($\times 20$) of migrated cells (underside of the membrane) within HCC1143-sorted fractions stained with 0.2 % crystal violet. Migration was quantified by measuring the optical density of the eluted crystal violet solubilized with 100 μ l of methanol. **c** Dual keratin 5/keratin 8 and vimentin IF imaging of HCC1143-sorted fractions. **d** Proliferation status of the different sorted fractions during cell culture after FACS. Proliferation was estimated by recording the absorbance at 490 nm of the MTS-PES compound in each time point. **e, f** In vitro differentiation of EpCAM^{low}/CD49⁺ HCC1143 and HCC38 cells. The two-sorted cell subpopulations from each cell line were grown in vitro under the same conditions as before FACS. After 14–18 days in culture, expression of CD49 and EpCAM was reanalyzed in both subpopulations using FACS

immortalized HMEC) and its two stem cell-like HMLE counterparts (coming from the same source) where TWIST1 and SNAI1 had been artificially induced, and are tumorigenic in nude mice by introduction of an activated form of the HER2/neu oncogene [30]. While the HMLE cell line showed the expected antigenic and gene expression MaSC/BiP phenotype, induction of a stem cell phenotype in HMLE cell line by expression of TWIST1 or SNAI1 resembled a MaSC/BiP \rightarrow stromal differentiation direction with acquisition of a claudin-low profile and phenotype (EpCAM⁻/CD49⁺) together with an increased expression of vimentin and decreased expression of keratin 8 (Fig. 7 and Supplemental material).

Claudin-low/stromal cells resemble the hMSC phenotype

Recent reports have implicated hMSCs and/or hESCs and their biological processes in the development and/or progression of breast cancer [33–36]. To gain further insight into the relationship of hMSC and hESC with breast-derived cell lines, we characterized and expression profiled a panel of three hMSC and two hESC cell lines. In terms of antigenic phenotype, all three hMSC cell lines showed a stromal (EpCAM⁻/CD49^{low}) and a CSC (CD24⁻/CD44⁺) phenotype (Supplemental material), while the two hESC lines showed a luminal progenitor (EpCAM⁺/CD49⁺) and a non-stem cell phenotypes of CD24⁺/CD44⁺. Concordant with the FACS data, hMSCs were vimentin-positive and keratin –5 and –8 negative, whereas hESCs were found negative for all three markers. More importantly, gene expression analyses revealed that the hMSCs cluster with claudin-low BCCLs and HMFs, while the two hESCs cluster as its own group, but closest to the epithelial basal-like and luminal/HER2-enriched subtypes (Fig. 7). Intrinsic molecular subtyping revealed that the hMSCs and hESC are identified as claudin-low and basal-like,

respectively. Finally, we observed that the core gene expression signature obtained by comparing the EpCAM⁺ (basal-like) and EpCAM⁻ (claudin-low) cell fractions of the HCC38, SUM149PT, and HCC1143 cell lines resembles a previously published gene expression pattern of hESC cells after differentiating into the mesodermal lineage (i.e., hMSCs) [37] (Fig. 4d).

Summary of the characterization of cell lines

A summary of the phenotypic and molecular data of all cell lines derived from breast cancers and normal mammary tissues evaluated in this study is provided (Fig. 8).

Discussion

In this report, we have characterized the phenotypic and molecular features of a large panel of cell lines derived from breast cancers and normal mammary tissues, and we have linked these features with the intrinsic subtypes of breast tumors, FACS enriched cell subpopulations of the normal mammary gland, and two types of true stem cells. Specifically, we made the following observations: (1) BCCLs in general resemble all the intrinsic subtypes of breast cancer except for luminal A, (2) BCCLs recapitulate all the differentiation statuses observed in the normal breast with HMECs best resembling the MaSC/BiP-enriched subpopulation, (3) subpopulations of cells with claudin-low and basal-like features are typically found within the subset of triple-negative cancer cell lines with overall basal-like features, and (4) within these mixed basal-like cell lines (or primary tumor xenografts WashU-WHIM2) the EpCAM⁺/CD49⁺ cells are more proliferative and more tumorigenic than the Claudin-low-like EpCAM⁻/CD49⁺ fraction, which is more motile.

Established in vitro breast cancer cell lines are being extensively used by the research community to address various aspects of cancer biology [12, 38–40]. Our data indicate that cell lines do recapitulate most of the differentiation states observed in breast cancer; however, we did not identify cell lines that resemble the good prognosis luminal A tumor type, which is the most frequent subtype identified in breast cancer [1–4]. One potential explanation for this finding is that the vast majority of luminal cell lines have been derived from metastatic tumor samples, such as pleural effusions (i.e., MCF7, T47D) or ascites (i.e., ZR75-1), therefore, introducing a selection bias toward more aggressive subtypes, such as those observed in the poor prognostic luminal B subtype. In addition, 2D in vitro assay itself and/or the media conditions used for cell culture might be a harsh environment for luminal A-like cells which is also reflected by the fact that despite the

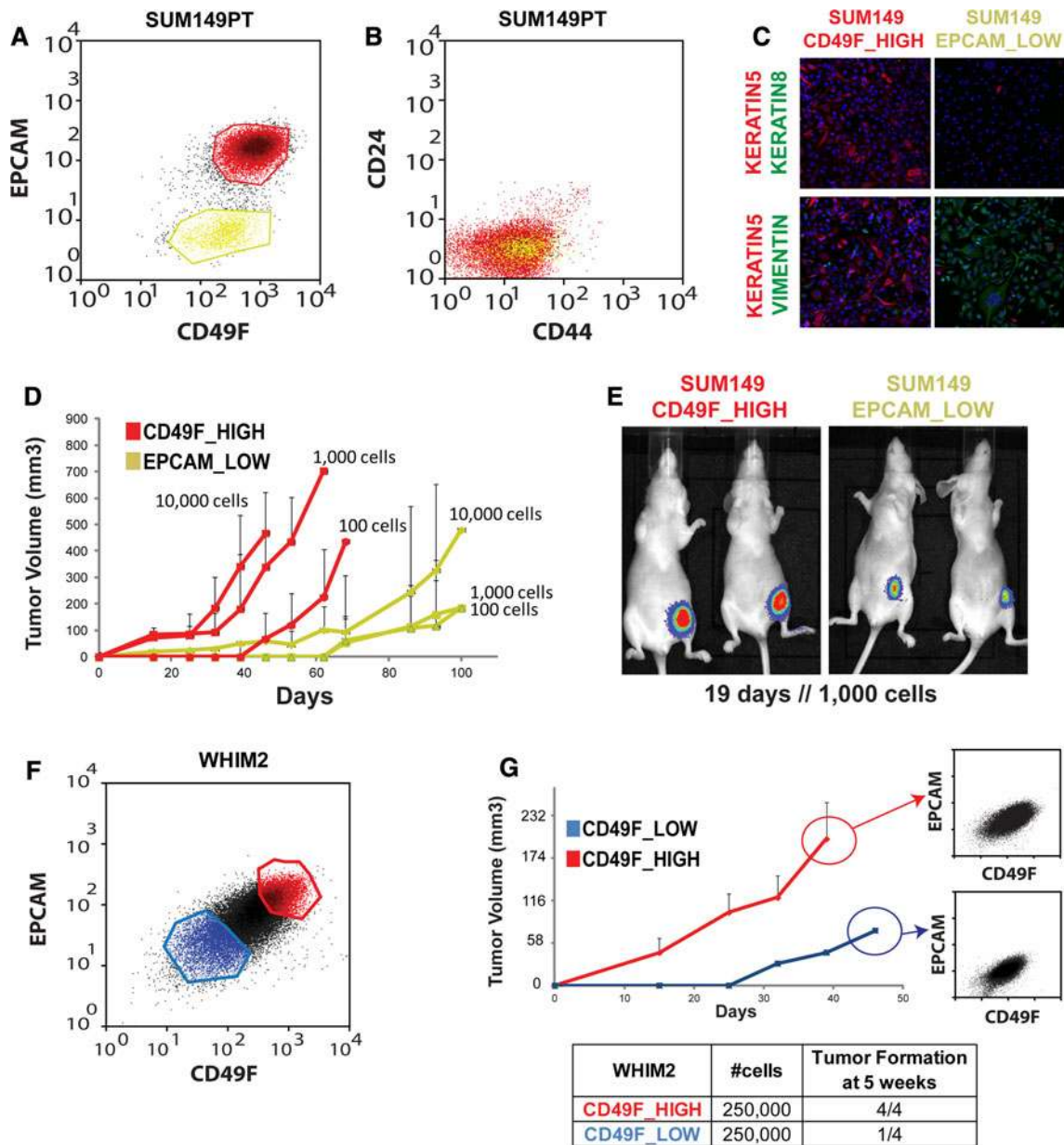


Fig. 6 TIC experiments in SUM149PT cell line and WashU-WHIM2. **a** FACS-sorted plot based on EpCAM and CD49f expression. Red color: basal-like; yellowish color: Claudin-low. **b** FACS-sorted plot based on CD24 and CD44 expression. Red and yellowish colors identify the population of cells identified in the EpCAM/CD49f plot. **c** Dual keratin 5/keratin 8 and keratin 5/vimentin IF imaging of SUM149PT-sorted fractions. **d** TIC experiment for both sorted

fractions. **e** Luciferase imaging of tumors in nude mice ($n = 2$ for each fraction) 19 days after injecting 1,000 cells. **f** Expression of EpCAM/CD49f and identification of two different cell fractions within the WashU WHIM2 model: CD49Flow and CD49Fhigh. **g** Tumor-initiating ability of the two-sorted fractions in (f), and expression of EpCAM/CD49f of the resulting tumors

observation that 10 % of lineage-negative cells in the normal breast FACS experiments are mL or pL. However, none of these cells could be readily identified in our 2D cultures of primary HMECs. In fact, the percentage of success of obtaining a cell line from ER+ primary tumors has been reported to be <10 % [41, 42]. This suggests that only cells with low adherence, high proliferation, and migration capabilities are more likely to be selected for

further passage, thus precluding the establishment of low proliferative and highly adherent luminal A/mL cells. This hypothesis could explain why among the 65 BCCLs evaluated, 66 % (43/65) are ER-negative, which is clearly not representative of the subtype incidence in patients.

The overall gene expression profiles of the cell lines that technically overlapped ($n = 52$) across four independent cell line data sets were highly similar. However, seven

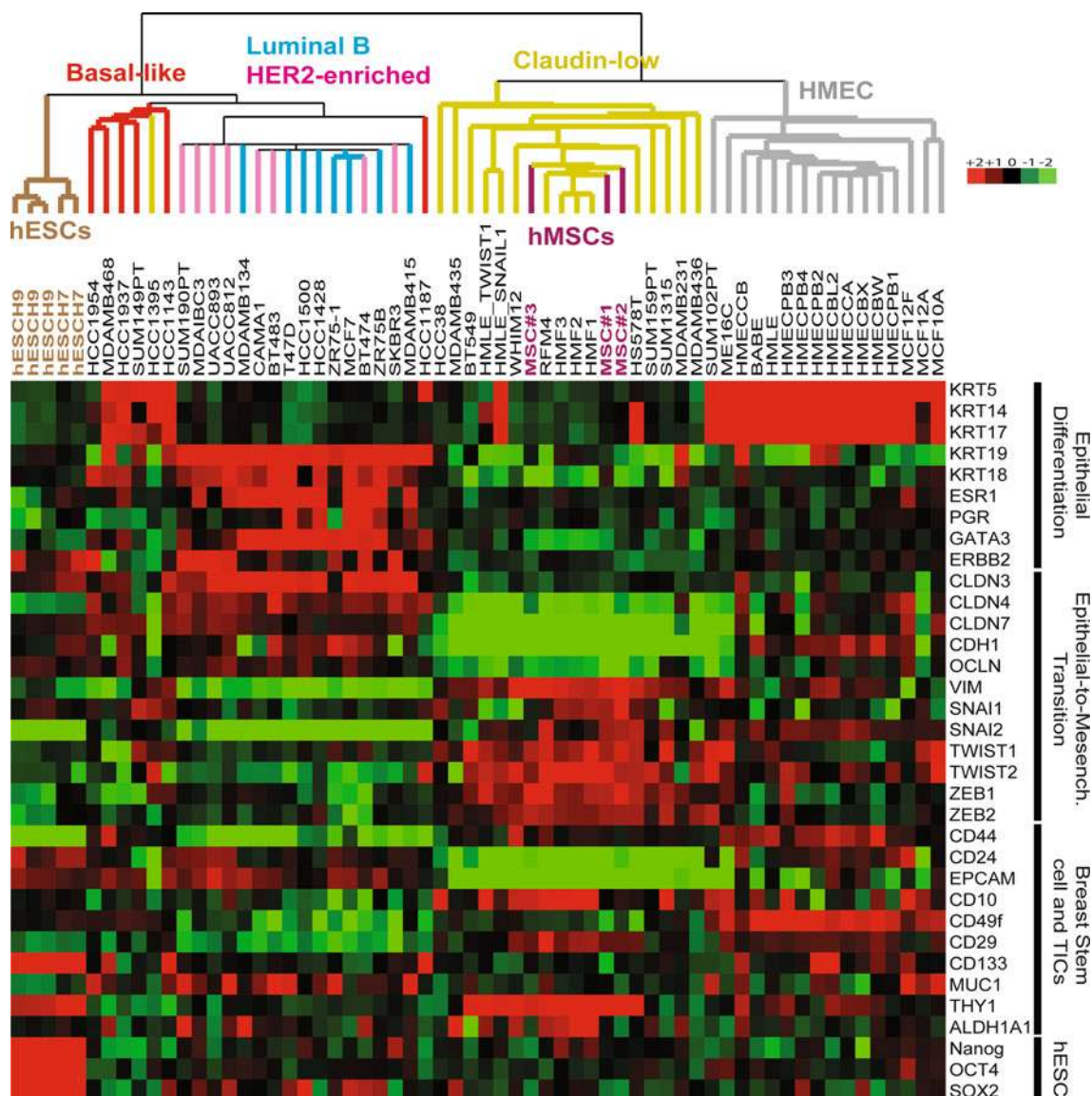


Fig. 7 Characterization of hMSC and hESC. Unsupervised hierarchical clustering of cell lines using the most variable genes ($n = 17,824$). Expression of selected genes is shown in the heatmap. Each colored square of the heatmap represents the relative transcript

abundance (in log₂ space) for each cell fraction with highest expression being red, average expression being black, and lowest expression being green

(13 %) discrepancies were noted. Most of these discrepancies occurred in cell lines whose gene expression profiles were found borderline between two subtypes, except for HCC1500 cell line (Supplemental material). For example, ER-negative/HER2-negative MDA-MB468 cell line is basal-like in two data sets (Hollestelle et al. [21] and UNC105), and shows borderline significance for HER2-enriched in the other two data sets, while ER-positive/HER2-amplified BT474 is called HER2-enriched in three data sets and luminal B in Kao et al. [22]. This finding could be explained by the specific genotypic/phenotypic features of these cell lines that are also observed in the two subtypes. For example, BT474 is a known ER+/HER2-

amplified cell line [43, 44]; while MDA-MB468 is a ER-negative/HER2-negative cell line with EGFR amplification [45], which might activate, in part, the HER2 pathway as in a HER2-amplified tumor.

The cell line data presented here also support our previously reported relationship between the basal-like and the claudin-low phenotypes [1]. Namely, we observed that the three ER-negative/HER2-negative cell lines classified as basal-like (HCC1143, SUM149PT) or claudin-low (HCC38) have basal-like and claudin-low subpopulations of cells within them, albeit with different proportions. Besides, similar to EpCAM^{-low}/CD49f⁺ cells in SUM149PT [1], claudin-low EpCAM^{-low}/CD49f⁺ cells from HCC1143 cell

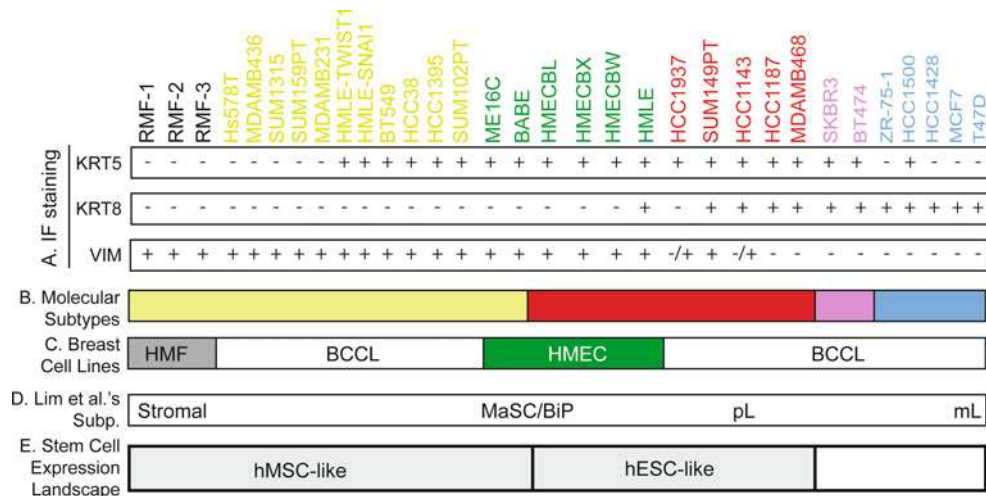


Fig. 8 Summary of the characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic subtypes. **a** IF staining for vimentin, keratin 5 and keratin 8 proteins. **b** Tumor molecular subtypes that each cell line best resemble. **c** Cell-type of each cell line. **d** Approximate localization of

Lim et al.'s gene expression profiles of each normal breast subpopulations. **e** Approximate genomic expression landscape of human embryonic and mesenchymal stem cell profiles. **f** Approximate genomic expression landscape of the mesenchymal and epithelial profiles

lines can differentiate and give rise to basal-like $\text{EpCAM}^+/\text{CD49f}^+$ cells. In vivo, tumors obtained from the $\text{EpCAM}^-/\text{low}$ and $\text{EpCAM}^+/\text{high}$ fractions show a FACS profile similar to the starting cell line (or tumor for WashU-WHIM2). Thus, even when only $\text{EpCAM}^-/\text{low}$ claudin-low-like cells are used, the natural state and balance are re-established both in vitro and in vivo.

Furthermore, we have shown that despite expressing different levels of surface markers CD44 and CD24, the gene expression differences between $\text{EpCAM}^-/\text{low}/\text{CD49f}^+$ versus $\text{EpCAM}^+/\text{CD49f}^+$ cells within each cell line are highly similar across all the three cell lines, suggesting that the similar biological events (e.g., migration capability) are occurring between these two fractions. However, it is important to note that we did not evaluate other stem cell or TIC markers such as ALDH1 [46], and that the Matrigel used during the xenotransplantation assay can influence the properties of stem cells and TICs [47, 48]. In any case, recent RNAi knockdown experiments in the SUM149PT cell line have identified Smarcd3/Baf60c, and thus the SWI/SNF chromatin-remodeling complex, as a key mediator of this EMT by activating WNT signaling pathways [49].

Human epithelial cell lines derived from normal breast tissue are being extensively used by the research community either as primary cells or after immortalization by exogenous hTERT transduction [30, 50, 51]. Although, speculation of their basal origin and MaSC/BiP capacity has been previously suggested by others [51], no study to the best of our knowledge has specifically addressed to which epithelial cell-type these cell lines best resemble. Using a genomic, FACS and IF staining analyses with

luminal, basal and mesenchymal markers, we observed that both immortalized and primary HMECs in the pre-stasis stage [52] resemble a phenotype similar to the MaSC/BiP-enriched subpopulation as defined by Lim et al. [9]. Indeed, we observed that the vast majority of cells within HMECs express high levels of basal keratin 5 and are vimentin-positive. This is concordant with our data and Lim et al.'s [9] data showing that the highest percentage of keratin 5 and vimentin positivity is observed in the MaSC/BiP subpopulation. On the other hand, when compared to tumors, HMECs showed a differentiation state between the claudin-low and the basal-like tumor subtype, concordant overall with a simultaneous mesenchymal and basal state within these cells.

We and others have previously shown that the claudin-low tumors and cell lines are enriched for CSC biological processes [1, 53–56]. In this report, we have observed that although this subtype is more similar to the MaSC/BiP-enriched subpopulation than the other breast cancer subtypes, claudin-low cell lines show a loss of epithelial markers with acquisition of a stromal state that also resembles the stromal-enriched subpopulation (i.e., fibroblasts) as defined by Lim et al. [9]. This is concordant with the seminal article by Mani et al. [30] showing that the acquisition of a full epithelial-to-mesenchymal transition after transfecting EMT-inducing transcription factors TWIST1 or SNAI1 into an immortalized HMEC increases the self-renewal capacity (a feature of stemness [51]) of the cells, and when transformed with KRAS oncogene allows to form tumors more efficiently in nude mice. In this report, using the same cell line variants developed by Mani et al. [30], together with a combination of genomics and EpCAM

and CD49f surface markers, we have shown that this mesenchymal transformation actually resembles a MaSC/BiP → stromal direction. Nonetheless, Battula et al. [34] have further characterized these EMT-derived HMECs and have shown that these cells are similar to bone marrow-derived mesenchymal stem cells with the capacity to differentiate into multiple tissue lineages such as osteoblasts, chondrocytes, and adipocytes. Intriguingly, transformation into tissue types other than the ones found in the mammary gland, such as, bone or cartilage is also observed in metaplastic tumors [57, 58], a rare histological type of breast cancer associated with poor prognosis and enriched for CSC/claudin-low profiles [56, 59]. Overall, these data suggest that the acquisition of a full mesenchymal state induces a multi-potent state more similar to mesenchymal stem cells than the more restricted MaSC/BiP, which seem to be in a partial mesenchymal and basal state. Thus, claudin-low tumors and cell lines might have an origin in a yet unidentified cell-type that is less differentiated than the MaSC/BiP-enriched subpopulation as defined in Lim et al. [9]. Conversely, the cell of origin of claudin-low and basal-like tumors could still be a MaSC/BiP phenotype, featuring various degrees of the EMT induction with claudin-low cells going to the full EMT state. Alternatively, the cell of origin of claudin-low tumors could be a highly undifferentiated normal cell that already expresses these stromal features, thus without the need for an EMT transition. Further studies that combine molecular profiling and lineage tracing experiments are needed to determine the cell of origin of each subtype.

To conclude, the integration of global gene expression data of cell lines with tumors and normal cell subpopulations is a novel strategy and could be used in other tumor types since it allows determining objectively which tumor or cell-type each cell line best resembles. The results presented here should also help to improve our understanding of the widely used encyclopedia of breast cell line models, and provide more precise tools for the study of breast cancers.

Materials and methods

UNC human breast tumor and cell line microarray data sets

For human tumor and normal tissue samples, we used all the microarrays and clinical data from Prat et al. (UNC337, GSE18229) [1]. For cell lines and sorted tissue, RNA was purified using RNeasy Mini kit and profiled as described previously using oligo microarrays (Agilent Technologies, USA) [60]. All microarray cell line data has been deposited in the Gene Expression Omnibus under the accession

number GSE50470 (referred to here as UNC105). The probes or genes of the combined UNC337 and UNC105 data set for all analyses were filtered by requiring the lowest normalized intensity values in both sample and control to be >10. The normalized log₂ ratios (Cy5 sample/Cy3 control) of probes mapping to the same gene (Entrez ID as defined by the manufacturer) were averaged to generate independent expression estimates.

Integration of three independent cell line data sets to the UNC337-UNC105 set

We used our cohort of cell lines (UNC105) and three publicly available microarray cell line data from the following data sets: Neve et al. (<http://icbp.lbl.gov/ccc/index.php>) [19], Hollestelle et al. (GSE16795) [21], and Kao et al. (<http://smd.stanford.edu/>) [22]. For all publicly data sets, raw data was normalized using the robust multi-array analysis normalization approach. To integrate all the datasets, we assumed that the five matched cell lines that are common to all four cohorts were the same and thus used them as controls. In supplemental material, a diagram summarizes the different microarray data sets analyzed in the different figures and the combination strategy for molecular subtyping each cell line.

Intrinsic subtype classification of cell lines

For the basal-like, HER2-enriched, luminal A, luminal B, and normal breast-like intrinsic subtype classification, we calculated the distance of each cell line to each of the tumor subtype centroids, and assigned a subtype call where the lowest distance was identified. Next, claudin-low cell lines were identified using the previously reported 9-cell line claudin-low predictor [1]. Samples identified as claudin-low were called claudin-low regardless of the previous subtype call. Euclidian distances and subtype calls for all cell lines are provided in Supplemental data.

Breast cancer cell lines, and immortalized HMEC/HMFs

SUM159PT (Asterand) and SUM1315O2 cells (Asterand) were maintained in Ham's F12 with 5 % fetal bovine serum (FBS), insulin (5 µg/ml), hydrocortisone (1 µg/ml, SUM159PT-only), and EGF (10 ng/ml, SUM1315O2-only). MCF-7, BT474, SKBR3, HCC1428, HCC1187, HCC1143, BT549, HCC1395, HCC38, UAC893, ZR75-1, HCC1500, T47D, and HCC1937 were cultured in RPMI with 10 % FBS [61]. SUM149PT was maintained in HMEC media with supplements (Gibco) with 5 % FBS [62]. MDA-MB231, Hs578T, and MDA-MB436 were cultured in DMEM (high glucose) with 10 % FBS. HME-CC

(BABE) [61], SUM102PT, HMLE, HMLE-SNAI1, HMLE-TWIST1, and HME31-hTERT no. 16C (ME16C) [61] were cultured in HuMEC media with supplements (Gibco). MDA-MB468 was cultured in Leibovitz's L-15 medium with 10 % FBS. HMLE, HMLE-SNAI1, and HMLE-TWIST1 cell lines were a kind gift of Sendurai A. Mani (University of Texas M.D. Anderson Cancer Center). An immortalized human mammary fibroblast cell line (called here HMF4) was a kind gift of Charlotte Kuperwasser (Tufts University School of Medicine). All cell lines were grown at 37 °C and 5 % carbon dioxide, and were obtained from the American Type Culture Collection unless otherwise specified. We also obtained total RNA from the following collaborators: Jeffrey M. Rosen and Rachel Schiff (Baylor College of Medicine; MCF10A, MDAMB415, MDAMB435, MDAMB134; BT483, CAMA1, UACC812, ZR75B); Ned Sharpless (UNC; UACC893); Sendurai A. Mani and Wendy Woodward (University of Texas M.D. Anderson Cancer Center; MCF12A, MCF12F, MDAIBC3, SUM190PT).

Mammary tissue and xenograft tumor tissue preparations

Fresh human normal breast tissues from five reduction mammoplasties were obtained using Institutional Review Board approved protocols. Unless otherwise stated, all reagents were from Stem Cell Technologies. Samples were minced and digested at 37 °C for 16 h in DMEM/F12 (GIBCO #11330) containing 0.5 µg/ml hydrocortisone, 5 µg/ml insulin, and 1× collagenase/hyaluronidase (#07912). Xenograft tumor tissues were dissociated for 2 h. The pellet from digested tissue was resuspended by pipetting for 5 min in warm 0.05 % trypsin–EDTA (GIBCO # 25300054) followed by addition of 1:10 mixture of DNase I (#07900), and Dispase (#07923). Red blood cells were removed by lysis in 1:4 mixture of cold Hanks' balanced salt solution (#37150) containing 2 % FBS (HF) and 0.8 % ammonium chloride solution (#07850). Cells were resuspended in HF and filtered through a 40 µm cell strainer (BD Falcon #352340) to obtain single cell suspensions.

Isolation of primary HMECs

Tissue obtained from four reduction mammoplasties were processed to obtain organoids. For this purpose dissociated tissue, as described above, were passed through 40 µm cell strainers. Organoids were collected from the top of the strainers using HMEC culture media, plated in 2D cultures and maintained in HuMEC media with supplements (Gibco). RNA was purified from all primary HMECs before passage 3 (pre-stasis stage) [52]. We also obtained total

RNA of four primary HMECs isolated by Pilar Blancafort (UNC; HMECPB1, HMECPB2, HMECPB3, HMECPB4) [63].

Isolation of primary HMFs

Single cells suspensions obtained from dissociation of three independent reduction mammoplasties as described above were cultured in DMEM/F-12 medium with 10 % FBS.

hESC and hMSCs

Two independent NIH hESC cell lines (H9 and H7) were obtained from the University of North Carolina Embryonic Stem Cell Core directed by B. Matthew Fagan. Commercially available hMSCs were purchased from Millipore, PromoCell, and Lonza.

Flow cytometry

Cells obtained from dissociated normal or tumor tissue, or trypsinized cell lines were counted, washed with HF, and stained for 30 min at 4 °C with antibodies specific for human cell surface markers from BD Pharmingen, except otherwise noted: EpCAM-FITC (Stem Cell Technologies, #10109), CD49f-PE-Cy5, (#551129), CD24-PE (#555428), CD44-APC (#559942), CD31-FITC (#555445), and CD45-FITC(#555482). Cells were washed from unbound antibodies and immediately analyzed using Beckman-Coulter (Dako) CyAn ADP or sorted using iCyt Reflection instrument. Cell viability was determined by using either blue-fluorescent reactive dye (Molecular Probes #L23105) or 7AAD (Molecular Probes #A1310). Dead cells and cells positive for lineage markers CD31 and CD45 were removed during sorting experiments. RNA was purified from sorted cells using RNeasy Mini kit (Qiagen).

Cell proliferation assay

Thousand cells from each sorted fraction were plated in 36 wells of a 96-well plate. At each time point, 20 µl of MTS-PES reagent was added in each well as provided in the CellTiter 96® AQueous One Solution Cell Proliferation Assay (Promega, USA), and we recorded its absorbance at 490 nm after 1 h of incubation. Three replicates for each time point and cell line were measured.

Immunofluorescence

Cell lines and normal breasts were processed using standard immunofluorescence staining methods as previously described [4]. The primary antibodies and their dilution were anti-vimentin (mouse anti-human IgG1-Kappa, dilution

1:100; Invitrogen/Zymed), anti-cytokeratin 5 (rabbit anti-human/mouse, dilution 1:50; Abcam, #ab24647), anti-cytokeratin 8 (CAM 5.2, mouse anti-human, dilution 1:2; Becton–Dickinson, #349205 and Zymed 18-0213, monoclonal, dilution 1:50).

TIC experiments

Luciferase stable SUM149PT cell line and tumors obtained from WashU-WHIM2 xenograft model were FAC sorted into subpopulations based on EpCAM and CD49F expression as described earlier. FAC-sorted cell fractions were placed on HuMEC media with supplements, 5 % FBS and 5 % MatrigelTM. For SUM149PT cell line, three different aliquots containing 100, 1,000, and 10,000 cells were injected into five nude mice each. Tumor volume was measured every 5–7 days by caliper in two dimensions. Experiments were done in triplicate. For the WashU-WHIM2 model, 250,000 cells of each fraction were injected in 4 NOD scid gamma mice.

Statistical analyses

Biologic analysis of microarray data was performed with DAVID annotation tool (<http://david.abcc.ncifcrf.gov/>) [64. SAM was performed in Excel as previously described] [1]. ANOVA, Student's *t* tests, and exact hypergeometric probability for gene expression data and Pearson correlation for protein–gene expression were performed using R (<http://cran.r-project.org>). Reported *p* are two-sided.

Acknowledgments We thank Jeffrey M. Rosen for critical review of the manuscript. C.M.P. was supported by grants from the NCI Breast SPORE program Grant P50-CA58223-09A1, RO1-CA-138255, RO1-CA148761, NCI Work Assignment HHSN-261200433008C, N01-CN43308, and the Breast Cancer Research Foundation.

Conflict of interest C.M.P. is an equity stock holder, and Board of Director Member, of BioClassifier LLC and University Genomics. C.M.P. is also listed and an inventor on a patent application on the PAM50 molecular assay.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Prat A, Parker J, Karginova O, Fan C, Livasy C, Herschkowitz J (2010) Phenotypic and molecular characterization of the Claudin-low Intrinsic subtype of breast cancer. *Breast Cancer Res* 12:R68
- Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA et al (2000) Molecular portraits of human breast tumours. *Nature* 406:747–752
- Sorlie T, Perou C, Tibshirani R, Aas T, Geisler S, Johnsen H et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 98:10869–10874
- Herschkowitz J, Simin K, Weigman V, Mikaelian I, Usary J, Hu Z et al (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* 8:R76
- Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T et al (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 27:1160–1167
- Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J et al (2009) Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst* 101:736–750
- Charles MP, Joel SP, Aleix P, Matthew JE, Philip SB: Clinical implementation of the intrinsic subtypes of breast cancer. *Lancet Oncol* 11:718–719
- Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T (2010) A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor positive breast cancer. *Clin Cancer Res* 16(21):5222–5232
- Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH et al (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 15:907
- Prat A, Perou CM (2009) Mammary development meets cancer genomics. *Nat Med* 15:842
- Prat A, Perou CM (2011) Deconstructing the molecular portraits of breast cancer. *Mol Oncol* 5:5–23
- Sharma SV, Haber DA, Settleman J (2010) Cell line-based platforms to evaluate the therapeutic efficacy of candidate anti-cancer agents. *Nat Rev Cancer* 10:241–253
- Soule H, Vazquez J, Long A, Albert S, Brennan M (1973) A human cell line from a pleural effusion derived from a breast carcinoma. *J Natl Cancer Inst* 51:1409–1416
- Osborne CK (1998) Tamoxifen in the treatment of breast cancer. *N Engl J Med* 339:1609–1618
- Scaltriti M, Verma C, Guzman M, Jimenez J, Parra JL, Pedersen K et al (2008) Lapatinib, a HER2 tyrosine kinase inhibitor, induces stabilization and accumulation of HER2 and potentiates trastuzumab-dependent cell cytotoxicity. *Oncogene* 28:803–814
- Scaltriti M, Eichhorn PJ, Cortés J, Prudkin L, Aura C, Jiménez J et al (2011) Cyclin E amplification/overexpression is a mechanism of trastuzumab resistance in HER2+ breast cancer patients. *Proc Natl Acad Sci USA* 108:3761–3766
- Huang F, Reeves K, Han X, Fairchild C, Platero S, Wong TW et al (2007) Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection. *Cancer Res* 67:2226–2238
- Abrams TJ, Murray LJ, Pesenti E, Walker Holway V, Colombo T, Lee LB et al (2003) Preclinical evaluation of the tyrosine kinase inhibitor SU11248 as a single agent and in combination with “standard of care” therapeutic agents for the treatment of breast cancer. *Mol Cancer Ther* 2:1011–1021
- Neve R, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T et al (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527
- Chin K, DeVries S, DFridlyand J, Spellman P, Roydasgupta R, Kuo W et al (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 10: 529–541
- Hollestelle A, Nagel J, Smid M, Lam S, Elstrodt F, Wasielewski M (2009) Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Breast Cancer Res Treat* 121(1):53–64

22. Kao J, Salari K, Bocanegra M, Choi Y-L, Girard L, Gandhi J (2009) Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One* 4:e6146
23. Sieuwerts A, Kraan J, Bolt J, Van der Spoel P, Elstrodt F, Schutte M et al (2009) Anti-epithelial cell adhesion molecule antibodies and the detection of circulating normal-like breast tumor cells. *J Natl Cancer Inst* 101:61–66
24. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S et al (2012) Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci USA* 109:2724–2729
25. Li Y, Welm B, Podsypanina K, Huang S, Chamorro M, Zhang X 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
26. Charafe-Jauffret E, Monville F, Ginestier C, Dontu G, Birnbaum D, Wicha MS (2008) Cancer stem cells in breast: current opinion and future challenges. *Pathobiology* 75:75–84
27. Hu Y, Smyth GK (2009) ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J Immunol Methods* 347:70–78
28. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW et al (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464:999–1005
29. Morel A-P, Lièvre M, Thomas Cm, Hinkal G, Ansieau Sp, Puisieux A (2008) Generation of breast cancer stem cells through epithelial–mesenchymal transition. *PLoS One* 3:e2888
30. Mani SA, Guo W, Liao M-J, Eaton EN, Ayyanan A, Zhou AY et al (2008) The epithelial–mesenchymal transition generates cells with properties of stem cells. *Cell* 133:704–715
31. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J et al (2010) Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with Claudin-low and metaplastic breast cancer subtypes. *PNAS* 107:15449–15454
32. Zhao X, Lu L, Pokhriyal N, Ma H, Duan L, Lin S et al (2009) Overexpression of RhoA induces preneoplastic transformation of primary mammary epithelial cells. *Cancer Res* 69:483–491
33. Ben-Porath I, Thomson M, Carey V, Ge R, Bell G, Regev A et al (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* 40:499–507
34. Battula VL, Evans KW, Hollier BG, Shi Y, Marini FC, Ayyanan A (2010) Epithelial–mesenchymal transition-derived cells exhibit multi-lineage differentiation potential similar to mesenchymal stem cells. *Stem Cells* 28(8):1435–1445
35. Liu S, Ginestier C, Ou SJ, Clouthier SG, Patel SH, Monville F et al (2011) Breast cancer stem cells are regulated by mesenchymal stem cells through cytokine networks. *Cancer Res* 71:614–624
36. Klopp AH, Lacerda L, Gupta A, Debeb BG, Solley T, Li L (2010) Mesenchymal stem cells promote mammosphere formation and decrease E-cadherin in normal and malignant breast cells. *PLoS One* 5:e12180
37. Evseenko D, Zhu Y, Schenke-Layland K, Kuo J, Latour B, Ge S et al (2010) Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proc Natl Acad Sci USA* 107:13742–13747
38. Lacroix M, Leclercq G (2004) Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast Cancer Res Treat* 83:249–289
39. Burdall S, Hanby A, Lansdown M, Speirs V (2003) Breast cancer cell lines: friend or foe? *Breast Cancer Res* 5:89–95
40. Forozan F, Veldman R, Ammerman CA, Parsa NZ, Kallioniemi A, Kallioniemi OP et al (1999) Molecular cytogenetic analysis of 11 new breast cancer cell lines. *Br J Cancer* 81:1328–1334
41. Gazdar AF, Kurvari V, Virmani A, Gollahon L, Sakaguchi M, Westerfield M et al (1998) Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int J Cancer* 78:766–774
42. Amadori D, Bertoni L, Flamigni A, Savini S, Giovanni C, Casanova S et al (1993) Establishment and characterization of a new cell line from primary human breast carcinoma. *Breast Cancer Res Treat* 28:251–260
43. Kallioniemi OP, Kallioniemi A, Kurisu W, Thor A, Chen LC, Smith HS et al (1992) ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *Proc Natl Acad Sci USA* 89:5321–5325
44. Lasfargues E, Coutinho W, Redfield E (1978) Isolation of two human tumor epithelial cell lines from solid breast carcinomas. *J Natl Cancer Inst* 61:967–978
45. deFazio A, Chiew Y-E, Sini RL, Janes PW, Sutherland RL (2000) Expression of c- < I>erb </I> B receptors, heregulin and oestrogen receptor in human breast cell lines. *Int J Cancer* 87:487–498
46. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, Brown M et al (2007) ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell* 1:555–567
47. Quintana E, Shackleton M, Sabel MS, Fullen DR, Johnson TM, Morrison SJ (2008) Efficient tumour formation by single human melanoma cells. *Nature* 456:593–598
48. Vaillant F, Lindeman G, Visvader J (2011) Jekyll or Hyde: does Matrigel provide a more or less physiological environment in mammary repopulating assays? *Breast Cancer Res* 13:108
49. Jordan NV, Prat A, Abell AN, Zawistowski JS, Sciaky N, Karginova OA et al (2013) SWI/SNF chromatin-remodeling factor Smarcd3/Baf60c controls epithelial–mesenchymal transition by inducing Wnt5a signaling. *Mol Cell Biol* 33:3011–3025
50. Tian Y, Wang N, Lu Z (2010) Repression of Lim only protein 4-activated transcription inhibits proliferation and induces apoptosis of normal mammary epithelial cells and breast cancer cells. *Clin Exp Metastasis* 27(7):455–463
51. Dontu G, Abdallah WM, Foley JM, Jackson KW, Clarke MF, Kawamura MJ et al (2003) In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev* 17:1253–1270
52. Dimri GP, Lee X, Basile G, Acosta M, Scott G, Roskelley C et al (1995) A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proc Natl Acad Sci USA* 92:9363–9367
53. Charafe-Jauffret E, Ginestier C, Lovino F, Wicinski J, Cervera N, Finetti P et al (2009) Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer Res* 69:1302–1313
54. Fillmore C, Kuperwasser C (2008) Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. *Breast Cancer Res* 10:R25
55. Creighton CJ, Li X, Landis M, Dixon JM, Neumeister VM, Sjolund A et al (2009) Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proc Natl Acad Sci USA* 106:13820–13825
56. Hennessy BT, Gonzalez-Angulo A-M, Stemke-Hale K, Gilcrease MZ, Krishnamurthy S, Lee J-S et al (2009) Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res* 69:4116–4124
57. Lee JH, Kim E-K, Choi S, Nam KJ, Kim DC, Cho SH (2008) Metaplastic breast carcinoma with extensive osseous differentiation: a case report. *Breast* 17:314–316

58. Gwin K, Wheeler DT, Bossuyt V, Tavassoli FA (2010) Breast carcinoma with chondroid differentiation: a clinicopathologic study of 21 triple negative (ER–, PR–, Her2/neu–) cases. *Int J Surg Pathol* 18:27–35
59. Luini A, Aguilar M, Gatti G, Fasani R, Botteri E, Brito J et al (2007) Metaplastic carcinoma of the breast, an unusual disease with worse prognosis: the experience of the European Institute of Oncology and review of the literature. *Breast Cancer Res Treat* 101:349–353
60. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF et al (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96
61. Troester MA, Hoadley KA, Sorlie T, Herbert BS, Borresen-Dale AL, Lonning PE et al (2004) Cell-type-specific responses to chemotherapeutics in breast cancer. *Cancer Res* 64:4218–4226
62. Hoadley K, Weigman V, Fan C, Sawyer L, He X, Troester M et al (2007) EGFR associated expression profiles vary with breast tumor subtype. *BMC Genomics* 8:258
63. Beltran A, Rivenbark A, Richardson B, Yuan X, Quian H, Hunt J et al (2011) Generation of tumor-initiating cells by exogenous delivery of OCT4 transcription factor. *Breast Cancer Res* 13:R94
64. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC et al (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 4:R60