frontiers
in Genetics

Check for updates

Identification of Breast Cancer Stem Cell Related Genes Using Functional Cellular Assays Combined With Single-Cell RNA Sequencing in MDA-MB-231 Cells

Emma Jonasson[1], Salim Ghannoum[1], Emma Persson[1], Joakim Karlsson[2], Thomas Kroneis[1,3], Erik Larsson[2], Göran Landberg[1,4] and Anders Ståhlberg[1,5,6*]

[1] Department of Pathology and Genetics, Institute of Biomedicine, Sahlgrenska Cancer Center, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden, [2] Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden, [3] Department of Cell Biology, Histology and Embryology, Gottfried Schatz Research Center, Medical University of Graz, Graz, Austria, [4] Department of Clinical Pathology, Sahlgrenska University Hospital, Gothenburg, Sweden, [5] Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Sweden, [6] Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden

OPEN ACCESS

Edited by:
Barbara Di Camillo,
University of Padova, Italy

Reviewed by:
Alexander V. Favorov,
Johns Hopkins University,
United States
Bianca Maria Veneziani,
University of Naples Federico II, Italy

*Correspondence:
Anders Ståhlberg
anders.stahlberg@gu.se

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Breast cancer tumors display different cellular phenotypes. A growing body of evidence points toward a population of cancer stem cells (CSCs) that is important for metastasis and treatment resistance, although the characteristics of these cells are incomplete. We used mammosphere formation assay and label-retention assay as functional cellular approaches to enrich for cells with different degree of CSC properties in the breast cancer cell line MDA-MB-231 and performed single-cell RNA sequencing. We clustered the cells based on their gene expression profiles and identified three subpopulations, including a CSC-like population. The cell clustering into these subpopulations overlapped with the cellular enrichment approach applied. To molecularly define these groups, we identified genes differentially expressed between the three subpopulations which could be matched to enriched gene sets. We also investigated the transition process from CSC-like cells into more differentiated cell states. In the CSC population we found 14 significantly upregulated genes. Some of these potential breast CSC markers are associated to reported stem cell properties and clinical survival data, but further experimental validation is needed to confirm their cellular functions. Detailed characterization of CSCs improve our understanding of mechanisms for tumor progression and contribute to the identification of new treatment targets.

Keywords: breast cancer, cancer stem cell, cell proliferation assay, mammosphere assay, single-cell analysis, single-cell RNA sequencing

INTRODUCTION

Breast cancer is the most common cancer type affecting women worldwide and is one of the main causes of cancer-related deaths in women (Torre et al., 2015). It is a complex disease with many subtypes differing in prognosis and treatment options. Currently, breast cancer can be divided into four intrinsic subtypes; Luminal A, Luminal B, HER2-enriched and basal-like. The subtypes

are further subgrouped based on their expression pattern of proliferation marker Ki67, HER2, and the hormone receptors ER (estrogen receptor) and PgR (progesterone receptor) (Reis-Filho and Pusztai, 2011; Senkus et al., 2013). Breast cancer displays both large inter- and intra-tumor heterogeneity, where each tumor contains small subpopulation of cancer stem cells (CSCs) that drive tumor progression. Similarly to normal stem cells, CSCs are able to self-renew and can give rise to progenitor cells as well as more differentiated cells (Reya et al., 2001; Beck and Blanpain, 2013; Kreso and Dick, 2014). It is also known that CSCs are important for metastasis and therapy resistance in breast cancer (Al-Hajj et al., 2003; Bhola et al., 2013; Lawson et al., 2015). Hence, detailed understanding of CSCs is important when developing novel treatment strategies.

Breast CSCs display several cellular features that can be used to enrich for this specific subpopulation. CSCs, in contrast to other tumor cells, display the ability to initiate tumor formation in immunodeficient mice (Al-Hajj et al., 2003). CSCs are also connected to therapy-resistance (Fillmore and Kuperwasser, 2008). Another property of CSCs is their ability to proliferate and differentiate under non-adherent cell culture conditions, which is used in the mammosphere formation assay (Dontu et al., 2003). Studies have also shown that CSCs, similarly to normal stem cells, often have a slow-dividing, and sometimes quiescent, phenotype (Dembinski and Krauss, 2009; Pece et al., 2010; Lyle and Moore, 2011). This cellular feature can be assessed with label-retention assays, where the ability of cells to maintain high amounts of an incorporated dye indicates few cell divisions (Lyle and Moore, 2011). Specific expression of cell surface markers, like CD44$^+$/CD24$^{-/low}$ (Al-Hajj et al., 2003), and activity of aldehyde dehydrogenase (Ginestier et al., 2007) are also useful when enriching for CSCs. However, all these features are not unique to CSCs but shared by other cells too, limiting their use.

Most performed breast CSC studies are on populations of cells that cannot reveal any information about CSC heterogeneity. Single-cell analysis overcomes this obstacle. In breast cancer, single-cell gene expression analysis have been used to study, for example, metastatic potential (Lawson et al., 2015; Nguyen et al., 2016), drug response (Lee et al., 2014; Savage et al., 2017), intratumoral heterogeneity (Chung et al., 2017), and cell state transitions (Akrap et al., 2016).

In this study, we first enriched for breast CSCs by collecting slow/non-dividing MDA-MB-231 cells, identified by a label-retention assay, that formed mammospheres in non-adherent cell culture conditions. We then performed single-cell RNA sequencing and used single-cell algorithms to define CSC subpopulations. We also studied CSCs transition into more differentiated cells and identified potential breast CSC-specific biomarkers.

MATERIALS AND METHODS

Cell Culture

MDA-MB-231 luciferase cells were cultured in DMEM medium (Lonza) supplied with 10% FBS (Gibco, Thermo Fisher Scientific), 1% penicillin/streptomycin (Gibco, Thermo Fisher Scientific), 1 × MEM Non-essential Amino Acid Solution (Sigma-Aldrich), and 1% L-glutamine (GE healthcare). When passaging, cells were washed with DPBS (Gibco, Thermo Fisher Scientific) and detached using 0.25% trypsin (Gibco, Thermo Fisher Scientific) supplemented with 0.5 mM EDTA (Invitrogen, Thermo Fisher Scientific).

Cell Proliferation Staining and Mammosphere Formation Assay

Cells were stained with PKH26 Red Fluorescent Cell Linker Kit (Sigma-Aldrich), according to the manufacturer's instructions with some modifications. Briefly, cells were trypsinized and washed once with serum-free media. 1×10^6 cells were resuspended in 1 ml Diluent C and then mixed with 1 ml Diluent C containing 2 μM PKH26. After 3 min in room temperature the staining was inactivated by addition of 2 ml FBS (Gibco, Thermo Fisher Scientific) for 1 min in room temperature. Cells were then centrifuged for 10 min at 300 g and washed 3 times with 5 ml complete medium. For each wash, cell suspension was transferred to a clean tube. Finally, cells were resuspended in 1 ml DPBS.

For non-adherent mammosphere formation assay, single cells were generated using a 25G needle (HSW FINE-JECT). 90,000 cells were added to T175 flasks (TPP) pre-coated with 1.2% poly(2-hydroxyethyl methacrylate) (Sigma-Aldrich) in 95% ethanol (Apoteket) using 30 ml phenol-red free DMEM/F-12 medium supplied with 2% B-27 supplement (both Gibco, Thermo Fisher Scientific), 1% penicillin/streptomycin and 20 ng/ml epidermal growth factor (Corning).

After 120 h, spheres were carefully collected with a pipette and centrifuged for 3 min at 10 g to include spheres, but avoid single cells. The supernatant was discarded leaving 2 ml of media in the bottom of the tube. Mammospheres from 6 T175 flasks were pooled and then centrifuged for 5 min at 120 g. The cells were resuspended in 0.25% trypsin followed by 3 min incubation at 37°C. A single-cell suspension was obtained using a 25G needle and trypsin was inactivated using complete medium. Cells were centrifuged for 5 min at 580 g, washed once with DPBS, centrifuged for 5 min at 300 g and resuspended in DPBS supplied with 2% bovine serum albumin (Sigma-Aldrich). Finally, cells were filtered through a 70 μm cell strainer into a FACS tube (both Falcon, Corning) and kept on ice until sorting.

Cell Cycle Staining

To sort cells in the G1 cell-cycle phase, cells were trypsinized and 1×10^6 cells/ml were resuspended in Hanks' balanced salt solution (Gibco, Thermo Fisher Scientific) and stained for nuclear DNA using Vybrant DyeCycle Violet stain (Invitrogen, Thermo Fisher Scientific) using a final concentration of 5 μM for 30 min at 37°C. After staining, cells were resuspended in Hanks' balanced salt solution and filtered through a 70 μm cell strainer into a FACS tube. Cells were kept on ice until sorting.

Single-Cell Sorting

Single cells were sorted using a BD FACSAria II instrument and the FACSDiva software (both BD Biosciences). Single cells with high or low PKH26 intensity or with low intensity of

Vybrant DyeCycle Violet were sorted into 96-well PCR plates (Applied Biosystems, Thermo Fisher Scientific) containing 5 µl lysis buffer containing 1 µg/µl BSA supplied in 2.5% glycerol (Thermo Scientific, Thermo Fisher Scientific) and 0.2% Triton X-100 (Sigma-Aldrich). Three wells were kept with only lysis buffer as cell-free controls. After sorting, plates were immediately frozen on dry ice and stored in −80°C.

RNA Sequencing

The Smart-seq2 protocol was used to generate a sequencing library (Picelli et al., 2014). Reverse transcription was performed directly on the 96-well plates with collected cells. First, 1 µM of an adapter sequence-containing oligo-dT30VN (5′-AAGCAGTGGTATCAACGCAGAGTACT30VN-3′), 1 mM dNTP (both Sigma-Aldrich), and 0.04 µl of a 1:100,000 dilution of ERCC RNA Spike-In Mix 1 (Ambion, Thermo Fisher Scientific) were added to the sample followed by incubation at 72°C for 3 min and cooling to 4°C. Next, 1 × first-strand buffer (50 mM Tris–HCl pH 8.3, 75 mM KCl, and 3 mM $MgCl_2$), 5 mM dithiothreitol (both Invitrogen, Thermo Fisher Scientific), 10 mM $MgCl_2$ (Ambion, Thermo Fisher Scientific), 1 M betaine (Sigma-Aldrich), 0.6 µM adapter sequence-containing template switching oligonucleotides (5′-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′ with rG = riboguanosine and +G = locked nucleic acid modified guanosine, Eurogentec), 15 U RNaseOUT and 150 U SuperScript II (both Invitrogen, Thermo Fisher Scientific) were added to a final volume of 15 µl. Reverse transcription was performed in a PTC-200 instrument (MJ Research) at 42°C for 90 min and 70°C for 15 min. cDNA was stored at −20°C. Concentrations indicated refer to the final reverse transcription reaction.

Preamplification was performed in a 50 µL reaction containing 1 × KAPA Hifi HotStart ReadyMix (KAPA Biosystems), 0.1 µM primer (5′-AAGCAGTGGTATCAACGC AGAGT-3′; Sigma-Aldrich) and 7.5 µl cDNA. Preamplification was performed in a PTC-200 instrument at 98°C for 3 min followed by 23 cycles of amplification at 98°C for 20 s, 67°C for 15 s, and 72°C for 6 min and a final additional incubation at 72°C for 5 min. Samples were transferred from 72°C directly to dry ice and stored at −20°C.

Concentration and quality of preamplified cDNA were assessed with Agilent High Sensitivity DNA Kit on a 2100 Bioanalyzer Instrument (Agilent Technologies). To determine the concentration range of the samples, 33 randomly selected samples were analyzed. Based on these concentrations, 2 µl preamplified cDNA was diluted 1:100 in RNase/DNase-free water (Invitrogen, Thermo Fisher Scientific) and 2 µl diluted sample was used for tagmentation and indexing, in order to not exceed 100 pg cDNA. In total, 52 each of high and low intensity of PKH26 and 31 G1 cells, were further processed. Tagmentation and indexing were performed using Nextera XT DNA Library Preparation Kit and Nextera XT Index Kit v2 (Illumina). To each sample, 10 µl Tagment DNA Buffer and 5 µl Amplicon Tagment Mix was added in a total volume of 20 µl and tagmentation was performed in a PTC-200 instrument at 55°C for 5 min followed by cooling to 10°C. 5 µl Neutralize Tagment Buffer was added followed by centrifugation for 1 min at 1100 rpm and 5 min

incubation at room temperature. A mixture of 15 µl Nextera PCR Master Mix and 5 µl of each index 1 (i7) and index 2 (i5) adapters was added to a total volume of 50 µl. Library amplification was performed in a PTC-200 instrument at 72°C for 3 min, 95°C for 30 s followed by 16 cycles of amplification at 95°C for 10 s, 55°C for 30 s, and 72°C for 30 s and a final additional incubation at 72°C for 5 min followed by cooling to 10°C.

Amplified samples were purified using Agencourt AMPure XP beads (BD Biosciences). All sample volume was added to 30 µl beads generating a beads-to-sample ratio of 0.6 and suspension was mixed by pipetting. DNA binding to beads was performed for 5 min at room temperature followed by 5 min incubation on a magnet (DynaMag, Thermo Fisher Scientific). Supernatant was discarded and beads washed twice with 200 µl 80% ethanol. Beads were left to dry and sample was eluted with 17.5 µl RNase/DNase-free water for 2 min followed by 2 min incubation on magnet before eluted sample was retrieved.

The mass concentration of each sample was analyzed using Qubit dsDNA High Sensitivity Assay Kit (Invitrogen, Thermo Fisher Scientific). To assess the quality and molarity, 35 selected samples were analyzed with Agilent High Sensitivity DNA Kit. The average mean size was used to calculate the molar concentration of each sample. Samples with lower concentration than 5 nM were excluded and remaining samples were diluted to 5 nM and pooled. The library pool was purified once more using Agencourt AMPure XP beads with a beads-to-sample ratio of 0.6 and diluted to 3 nM.

Sequencing of pooled single-cell libraries were performed by Genomics Core facility at the University of Gothenburg on a NextSeq 500 instrument (Illumina) using a Nextseq500 Kit High Output V2 with paired-end sequencing and a read length of 2 × 150 bp.

Single-Cell Transcriptome Data Analysis

Reads were aligned to the human genome (hg19), with ERCC spike-in sequences included, using STAR (Dobin et al., 2012) with splice junctions supplied from the GENCODE (Harrow et al., 2012) V17 annotation. Gene expression levels were assessed by binning reads to genes using HTseq (Anders et al., 2015) with the options "-s no" and "-m intersection-strict." The quality of each sample was assessed using FastQC (Andrews, 2010). In total, four samples failed either in the sequencing process or during quality assessment.

Genes were filtered based on variability in comparison to a noise level estimated from the ERCC spike-ins as described (Brennecke et al., 2013), utilizing the dependence of technical noise on the average read count and fitting a model to the ERCC spike-ins. All genes with squared coefficient of variation above the noise level were selected for further analyses. An additional filtering step was included only selecting genes expressed in at least 5% of all cells.

Cells were clustered using a modified version of the RaceID algorithm (Grun et al., 2015). Shortly, read counts were normalized to the median transcript number across cells. Cells were clustered using k-means on the modified Pearson correlation matrix with Euclidian metric and data was visualized using t-distributed stochastic neighbor embedding. Cluster

number $k = 3$ was chosen and the stability of the clusters were assessed with Jaccard's similarity. Only genes selected using the filtering method described above were included in the clustering, without any additional filtering. The RaceID algorithm also includes an outlier detection method which was omitted.

Pseudotemporal ordering of cells was performed using the TSCAN algorithm (Ji and Ji, 2016), which first groups cells into clusters and then orders cells along a pseudotemporal path using a minimum spanning tree approach. For this method, all genes were used and filtered according to the default filtering in TSCAN. Normalized values received from the RaceID algorithm were used.

Differentially expressed genes between clusters were identified using the SCDE algorithm (Kharchenko et al., 2014). The raw data matrix, including all genes, was used for this algorithm. Adjusted *p*-values were calculated from the cZ-values received as output from the algorithm. Genes with an adjusted *p*-value less than 0.05 were considered significant.

Gene set analysis was performed using the molecular signatures database (MSigDB) (Subramanian et al., 2005). The gene lists obtained from differential expression analysis was compared with the Hallmark, Reactome and KEGG gene sets as well as transcription factor targets. Top 100 significantly enriched gene sets (cutoff of FDR $q = 0.05$) were identified.

Survival data was assessed using the Kaplan-Meier plotter tool for breast cancer[1] which is based on available microarray data (Gyorffy et al., 2010). Relapse-free survival was assessed using JetSet to select the optimal probe (Li et al., 2011) and auto-selection of best cutoff for dividing the patients into low and high expression of each gene. The analysis was performed separately on patients belonging to different intrinsic subgroups. Furthermore, ERα positive and negative breast cancer patients were analyzed as well as ERα negative patients with or without systemic treatment. Apart from this, no other selections of patients were done. To distinguish significantly altered survival, *p*-values were adjusted for multiple testing using false discovery rate (Benjamini-Hochberg procedure). The *p*-value adjustment was performed separately for each subgroup of patients including *p*-values for all analyzed genes.

RESULTS

Enrichment of Breast Cancer Stem Cells and Identification of Biologically Variable Genes

It has been shown that low proliferation (Dembinski and Krauss, 2009) and the ability to form spheres under non-adherent conditions (Dontu et al., 2003) are traits linked to the CSC populations. To functionally enrich for breast cancer cells with these characteristics, we applied the mammosphere assay in combination with the label-retention PKH26 assay on breast cancer cell line MDA-MB-231, as shown in **Figure 1A**. Cells were first labeled with the membrane dye PKH26 before being cultured as mammospheres for 5 days. Low (PKH26 high cells)

---

[1]http://kmplot.com

and high (PKH26 low cells) proliferating single cells dissociated from mammospheres were collected (**Figure 1B**). As control, normal adherent monolayer cells in the G1 cell cycle phase were collected using the Vybrant DyeCycle Violet DNA-binding dye (**Figures 1A,C**). We performed whole transcriptomic analysis at single-cell level applying the Smart-Seq2 protocol (Picelli et al., 2014). After sequencing and sample quality control we used 46 low proliferating mammosphere cells (PKH26 high cells), 45 high proliferating mammosphere cells (PKH26 low cells) and 30 cells in G1 cell cycle phase cultured in adherent monolayer conditions (G1 cells) for further analyses.

Next, we identified relevant genes with biological variation above the technical noise (Brennecke et al., 2013). Briefly, the technical noise level was assessed from the variation and expression of ERCC spike-in controls which were added to each sample before library preparation (**Figure 1D**). In total, 576 genes were identified with a variability above this technical noise level and expressed in at least 5% of all cells.

Identification of a Cancer Stem Cell Subpopulation

To identify subpopulations of cells we applied parts of the RaceID algorithm (Grun et al., 2015), which groups cells using *k*-means clustering and visualizes the obtained subpopulations with *t*-distributed stochastic neighbor embedding (*t*-SNE). This approach identified three cell clusters (**Figure 2A**).
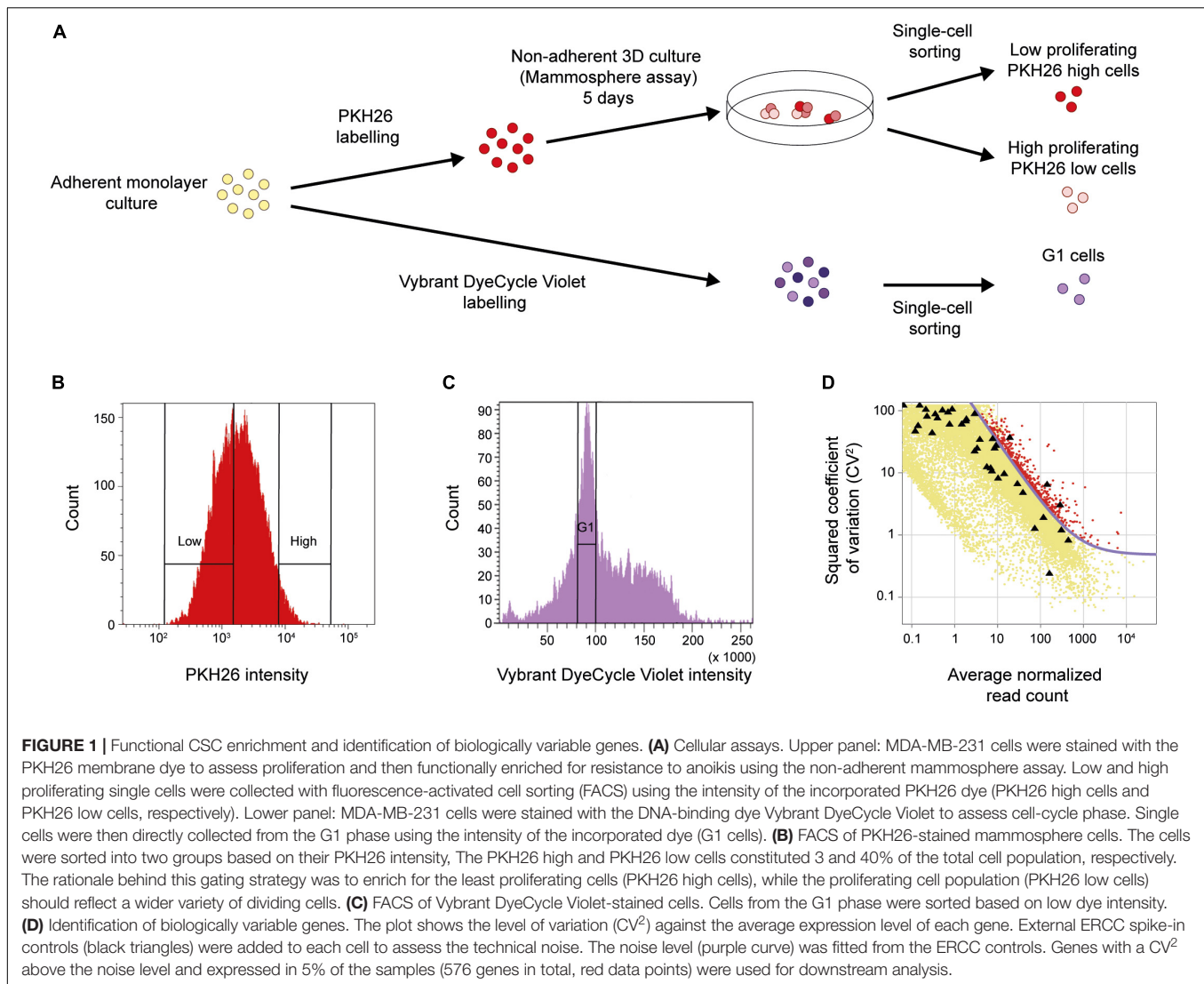
To compare the clustering based on gene expression to the functional properties assessed based on the culture and staining procedures, we overlaid the cellular phenotype in the *t*-SNE plot (**Figure 2B**). Cluster A mainly consisted of PKH26 high cells (76%), 69% of the cluster B cells were PKH26 low cells, while 73% of the cluster C cells were G1 cells. These results show that the enriched cellular phenotypes were also reflected at gene expression level. To further confirm this relationship, we overlaid the mean expression of cell-cycle related genes in the *t*-SNE plot (**Figure 2C**). High proliferating (PKH26 low cells) and G1 cells were the cells that expressed cell-cycle related genes to highest extent.

To define the gradual transition of the cells between different subpopulations we constructed a pseudotemporal ordering of all cells using the TSCAN algorithm (Ji and Ji, 2016). This algorithm orders cells depending on their gradual changes in the transcriptome profile. The pseudotemporal cell order was visualized in the *t*-SNE plot with the previously identified clusters (**Figure 2D**). Early ordered cells were more prone to belong to cluster A whereas cells toward the end mainly belonged to cluster C. Altogether, these data show that cluster A is the most CSC-like subpopulation.

Subpopulation-Specific Gene Expression Profiles

To characterize the three identified subpopulations we defined differentially expressed genes using the SCDE algorithm (Kharchenko et al., 2014). Gene lists for all pair-wise comparisons are shown in **Supplementary Table S1**. The regulated genes were compared with the Hallmark, Reactome and KEGG gene

FIGURE 1 | Functional CSC enrichment and identification of biologically variable genes. **(A)** Cellular assays. Upper panel: MDA-MB-231 cells were stained with the PKH26 membrane dye to assess proliferation and then functionally enriched for resistance to anoikis using the non-adherent mammosphere assay. Low and high proliferating single cells were collected with fluorescence-activated cell sorting (FACS) using the intensity of the incorporated PKH26 dye (PKH26 high cells and PKH26 low cells, respectively). Lower panel: MDA-MB-231 cells were stained with the DNA-binding dye Vybrant DyeCycle Violet to assess cell-cycle phase. Single cells were then directly collected from the G1 phase using the intensity of the incorporated dye (G1 cells). **(B)** FACS of PKH26-stained mammosphere cells. The cells were sorted into two groups based on their PKH26 intensity, The PKH26 high and PKH26 low cells constituted 3 and 40% of the total cell population, respectively. The rationale behind this gating strategy was to enrich for the least proliferating cells (PKH26 high cells), while the proliferating cell population (PKH26 low cells) should reflect a wider variety of dividing cells. **(C)** FACS of Vybrant DyeCycle Violet-stained cells. Cells from the G1 phase were sorted based on low dye intensity. **(D)** Identification of biologically variable genes. The plot shows the level of variation (CV^2) against the average expression level of each gene. External ERCC spike-in controls (black triangles) were added to each cell to assess the technical noise. The noise level (purple curve) was fitted from the ERCC controls. Genes with a CV^2 above the noise level and expressed in 5% of the samples (576 genes in total, red data points) were used for downstream analysis.
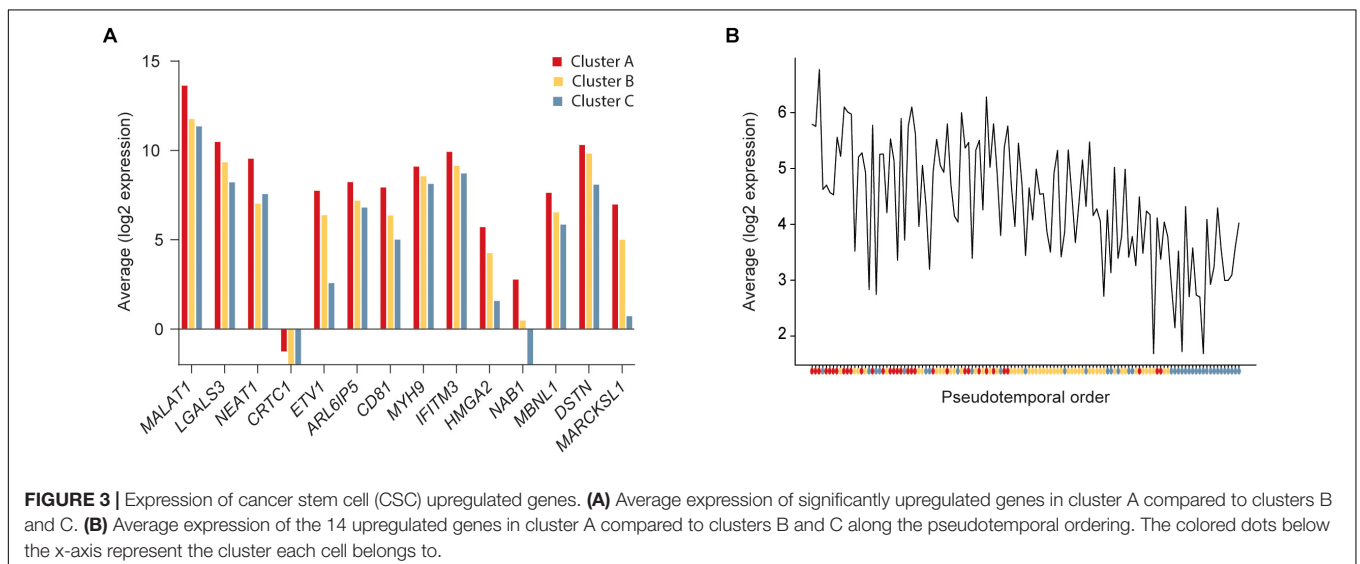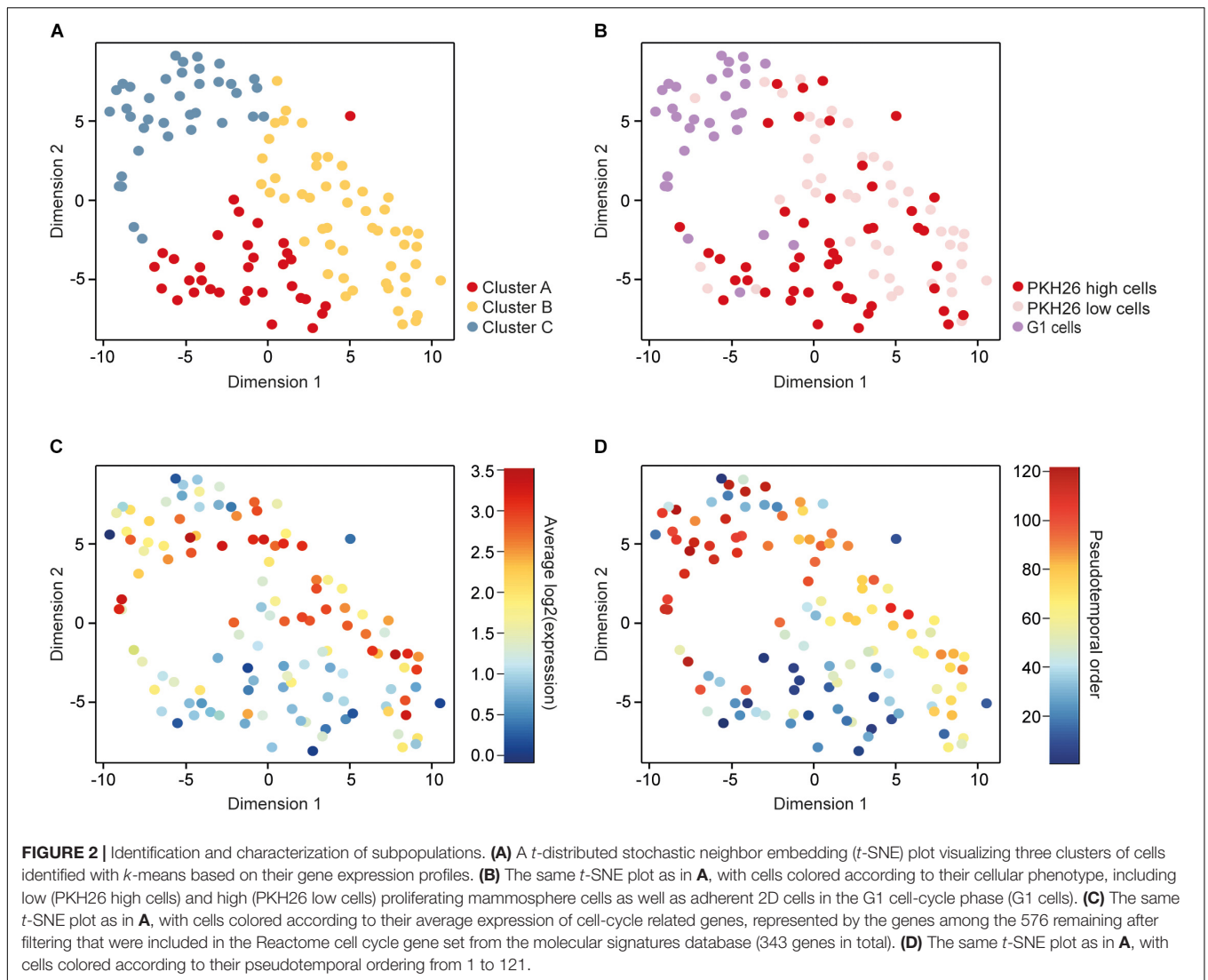
sets in the Molecular Signatures Database (**Supplementary Table S2**; Subramanian et al., 2005). Furthermore, over-representation of targets for transcription factors was identified using the same database (**Supplementary Table S2**). Another difference between the subpopulations was that the relative amount of total transcripts was higher in cluster C compared to clusters A and B (**Supplementary Figure S1A**). The relative amount of total transcripts also increased toward the end of the pseudotemporal ordering (**Supplementary Figure S1B**).

To further define the CSC-like cells in cluster A we identified 14 genes that were significantly upregulated in this cluster compared to the other two clusters (**Table 1**). The average expression of each gene in the three clusters is illustrated in **Figure 3A** and the average expression of all 14 genes along the pseudotemporal ordering is shown in **Figure 3B**. To determine the association between the 14 genes and clinical survival data for ERα positive and negative tumors we used a Kaplan-Meier analysis tool based on a collection of publicly available data (Gyorffy et al., 2010; **Table 1** and **Supplementary Figure S2**). Furthermore, we performed the same survival analyses in patients with tumors belonging to different intrinsic subgroups (basal-like, luminal A, luminal B and HER2-enriched) as well as in patients with ERα negative tumors with or without systemic treatment (**Supplementary Table S3**). The data showed variable results among different genes and patient groups but three genes, *LGALS3*, *MYH9*, and *DSTN*, were significantly related to worse survival in ERα negative tumors.

DISCUSSION

Breast cancer tumors are heterogeneous and harbor CSCs. Several pathways related to self-renewal are associated to breast CSCs, including the Wnt, Notch, and Hedgehog pathways (Liu et al., 2006; Harrison et al., 2010; Khramtsov et al., 2010). These signaling pathways are also connected to each other, indicating a complex system for CSC maintenance (Takebe et al., 2011). There have been attempts of targeting the CSC population through

FIGURE 2 | Identification and characterization of subpopulations. **(A)** A *t*-distributed stochastic neighbor embedding (*t*-SNE) plot visualizing three clusters of cells identified with *k*-means based on their gene expression profiles. **(B)** The same *t*-SNE plot as in **A**, with cells colored according to their cellular phenotype, including low (PKH26 high cells) and high (PKH26 low cells) proliferating mammosphere cells as well as adherent 2D cells in the G1 cell-cycle phase (G1 cells). **(C)** The same *t*-SNE plot as in **A**, with cells colored according to their average expression of cell-cycle related genes, represented by the genes among the 576 remaining after filtering that were included in the Reactome cell cycle gene set from the molecular signatures database (343 genes in total). **(D)** The same *t*-SNE plot as in **A**, with cells colored according to their pseudotemporal ordering from 1 to 121.

FIGURE 3 | Expression of cancer stem cell (CSC) upregulated genes. **(A)** Average expression of significantly upregulated genes in cluster A compared to clusters B and C. **(B)** Average expression of the 14 upregulated genes in cluster A compared to clusters B and C along the pseudotemporal ordering. The colored dots below the x-axis represent the cluster each cell belongs to.

TABLE 1 | Genes upregulated in cluster A.

Gene name	Log2 fold change[a]	Adj. *p*-value[b]	Survival data[c]	Function[d]
MALAT1 (non-coding)	0.95	4.8×10^{-9}	ERα neg. – ERα pos. –	mRNA regulation of genes connected to cell cycle regulation and cell migration
LGALS3	1.4	3.8×10^{-6}	ERα neg. ↓ ERα pos. ↑	Involved in apoptosis and cell adhesion
NEAT1 (non-coding)	1.9	5.3×10^{-6}	ERα neg. – ERα pos. –	Paraspeckle formation, mRNA regulation
CRTC1	4.1	3.0×10^{-4}	ERα neg. ↑ ERα pos. –	Transcriptional coactivator of CREB1, involved in several pathways
ETV1	1.5	2.1×10^{-3}	ERα neg. ↑ ERα pos. –	Transcription factor of the ETS family, involved in many biological processes
ARL6IP5	1.3	2.7×10^{-3}	ERα neg. – ERα pos. ↑	Expression affected by vitamin A. May be associated with the cytoskeleton and glutamate transport
CD81	1.2	9.0×10^{-3}	ERα neg. – ERα pos. –	Possibly involved in muscle cell fusion and signal transduction, might be connected to cell growth
MYH9	1.1	9.5×10^{-3}	ERα neg. ↓ ERα pos. –	Cytokinesis, cell motility and maintenance of cell shape
IFITM3	0.91	1.3×10^{-2}	ERα neg. – ERα pos. ↑	Interferon-induced membrane protein. Inhibits entry of viruses
HMGA2	1.2	1.5×10^{-2}	ERα neg. – ERα pos. –	Transcriptional regulation, involved in cell cycle regulation
NAB1	2.2	2.2×10^{-2}	ERα neg. – ERα pos. ↑	Transcriptional repressor for zinc finger transcription factors EGR1 and EGR2
MBNL1	1.1	2.2×10^{-2}	ERα neg. – ERα pos. ↑	Alternative splicing
DSTN	1.1	2.5×10^{-2}	ERα neg. ↓ ERα pos. –	Actin depolymerisation
MARCKSL1	1.5	2.7×10^{-2}	ERα neg. ↑ ERα pos. –	Cell movement and migration

[a]*Upregulation, comparing cluster A to both other clusters.* [b]*p-value adjusted for multiple testing by false discovery rate.* [c]*Correlation of gene expression to relapse-free survival assessed by Kaplan-Meier analysis using http://kmplot.com. Significantly (FDR < 0.05) better (↑) or worse (↓) survival indicated for ERα negative and positive breast cancer.* [d]*Functions of gene and/or encoded protein. Information from NCBI (NCBI Resource Coordinators, 2016) and/or The UniProt Consortium (2017).*

these pathways but this is associated with several challenges (Takebe et al., 2015) and no such therapy is in clinical use.

To further characterize breast CSCs we performed single-cell RNA sequencing on cellular populations functionally enriched for different CSC properties. The mammosphere assay enriches for cells that are able to self-renew and differentiate and the intensity of the PKH26 dye allows cells to be selected based on their proliferation rate (Pece et al., 2010). However, the mammosphere assay has been shown to select for a larger cell population than the CSCs only (Pastrana et al., 2011) and PKH26 is a membrane dye whose intensity readout could be affected by other cellular properties than cell division, such as cell size (Dolatabadi et al., 2017). For comparison, we included cells cultured under adherent monolayer conditions, selected from the G1 cell cycle phase only to minimize the effect of cell cycle specific gene expression (Whitfield et al., 2002). Our three single-cell defined subpopulations overlapped to a large extent with our three predefined groups, i.e., PKH26 high, PKH26 low, and G1 cells. Gene expression analysis showed proliferation to be a strong dividing factor between clusters and the cell group showing a decrease in expression of cell-cycle related genes also showed an over-representation of cells with high PKH26 intensity. These data support the accuracy of PKH26 in separating cells based on proliferation rate. In order to study the differentiation process, we ordered the cells along a pseudotemporal path.

The pseudotemporal ordering fitted well with the clustering, indicating a gradual transition of cells between different states along the different subpopulations, proposedly from a more CSC-like state to a more differentiated state.

To determine the cellular functions of identified cell clusters we defined enriched gene sets (**Supplementary Table S2**). As expected, when comparing cluster A with cluster B this analysis generated overrepresentation of gene sets related to cell cycle, confirming the use of label-retention PKH26 dye assay. Also, enriched transcription factor sequence motifs included several targets of E2Fs, known cell-cycle regulators (Morgan, 2007) as well as NFY, shown to interact with E2Fs but also involved in stem cell self-renewal (Zhu et al., 2004, 2005; Blum et al., 2009). When comparing cluster B to cluster C, overrepresented gene sets included several processes related to CSCs, such as epithelial to mesenchymal transition (Mani et al., 2008; Morel et al., 2008), hypoxia (Schwab et al., 2012; Harrison et al., 2013) and apoptosis (Madjd et al., 2009; Kruyt and Schuringa, 2010). Also, gene sets related to protein and RNA metabolism were upregulated. Among the transcription factor targets, CREB and ATF were highly represented, which are involved in diverse physiological processes, including proliferation, survival (Mayr and Montminy, 2001) and differentiation (Masson et al., 1993; Kingsley-Kallesen et al., 1999). Finally, the cellular functions describing the differences between clusters A and C reflected a

combination of functions describing differences of the other two cluster comparisons, including cell-cycle regulation, stem-cell properties and differentiation. Additionally, among transcription factor targets, AP1 was enriched, a transcription factor complex consisting of members of the JUN, FOS and ATF families that has been found to be connected to proliferation, invasion and apoptosis (Chen et al., 1996; Eferl and Wagner, 2003; Milde-Langosch et al., 2004). Global changes in RNA and protein have been seen when comparing stem cells to more differentiated cell types. Our results showed an increase in total amount of transcripts in cluster C, in line with earlier studies reporting decreased amount of total RNA in stem cells (Sampath et al., 2008; Sanchez et al., 2016).

We propose that cluster A contains the most CSC-like cells and that, for this cluster, upregulated genes are potential CSC biomarkers (**Table 1**). Among these genes, two were non-coding, *MALAT1* and *NEAT1*, of which both are connected to CSC properties in several cancer types, including breast cancer (Arun et al., 2016; Jen et al., 2017; Li et al., 2017). The coding genes are also related to CSC properties but also other cellular functions. For example, non-muscle myosin IIA, encoded by *MYH9*, is involved in cell motility and has been shown to be involved in migration of MDA-MB-231 cells (Betapudi et al., 2006; Dulyaninova et al., 2007). MYH9 also interacts with Galectin-3, encoded by *LGALS3*, in bone metastasis (Nakajima et al., 2016). Galectin-3 is connected to several tumor properties (Newlaczyl and Yu, 2011), including chemoresistance and stem cell properties in breast cancer (Newlaczyl and Yu, 2011; Guha et al., 2014). Furthermore, *HMGA2* is involved in metastasis through epithelial to mesenchymal transition (Thuault et al., 2006; Morishita et al., 2013). We also determined the connection between the expressions of these genes to clinical outcome using publicly available data. The varying relationships between gene and patient selection was not surprising, since it is well known that breast cancer subgroups, like ERα positive and negative breast cancer, often display divergent expression pattern (Reis-Filho and Pusztai, 2011). The three genes; *LGALS3*, *MYH9*, and *DSTN*, that were significantly associated with worse relapse-free survival in ERα negative breast cancers are of extra interest for future studies. Two of those genes, *LGALS3* and *DSTN*, were significantly related to worse survival specifically in patients that had not received systemic treatment, whereas the opposite were true for *MYH9*. Additionally, *NEAT1* was connected to worse survival in systemically treated ERα negative patients, although the analysis was performed in a smaller patient cohort (**Supplementary Table S3**). Looking into the intrinsic subgroups, two of the above-mentioned genes, *LGALS3* and *DSTN*, were significantly associated to worse survival in several subgroups including the basal group, in which also *NAB1* was connected to worse survival.

CONCLUSION

In conclusion, we have identified potential breast cancer biomarkers related to CSC properties, especially associated with ERα negative breast cancer, using functional cellular assays combined with single-cell gene expression profiling. Further experimental validation, using more cell lines as well as other model systems, is needed to confirm their cellular functions and potential clinical use.

DATA AVAILABILITY STATEMENT

RNA sequencing data have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) database with accession number GSE124989.

AUTHOR CONTRIBUTIONS

EJ and AS conceived the study. EJ, EP, and TK performed the experiments. EJ, SG, JK, EL, GL, and AS analyzed the data. EJ and AS drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00500/full#supplementary-material

REFERENCES

Akrap, N., Andersson, D., Bom, E., Gregersson, P., Ståhlberg, A., and Landberg, G. (2016). Identification of distinct breast cancer stem cell populations based on single-cell analyses of functionally enriched stem and progenitor pools. *Stem Cell Rep.* 6, 121–136. doi: 10.1016/j.stemcr.2015.12.006

Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J., and Clarke, M. F. (2003). Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3983–3988.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Andrews, S. (2010). *FastQC: A Quality Control Application for FastQ data*. Avilable at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed December 29, 2016).

Arun, G., Diermeier, S., Akerman, M., Chang, K. C., Wilkinson, J. E., and Hearn, S. (2016). Differentiation of mammary tumors and reduction in metastasis upon Malat1 lncRNA loss. *Genes Dev.* 30, 34–51. doi: 10.1101/gad.270959.115

Beck, B., and Blanpain, C. (2013). Unravelling cancer stem cell potential. *Nat. Rev. Cancer* 13, 727–738. doi: 10.1038/nrc3597

Betapudi, V., Licate, L. S., and Egelhoff, T. T. (2006). Distinct roles of nonmuscle myosin II isoforms in the regulation of MDA-MB-231 breast cancer cell spreading and migration. *Cancer Res.* 66, 4725–4733. doi: 10.1158/0008-5472.can-05-4236

Bhola, N. E., Balko, J. M., Dugger, T. C., Kuba, M. G., Sánchez, V., Sanders, M., et al. (2013). TGF-β inhibition enhances chemotherapy action against triple-negative breast cancer. *J. Clin. Investig.* 123, 1348–1358. doi: 10.1172/JCI65416

Blum, R., Gupta, R., Burger, P. E., Ontiveros, C. S., Salm, S. N., Xiong, X., et al. (2009). Molecular signatures of prostate stem cells reveal novel signaling pathways and provide insights into prostate cancer. *PLoS One* 4:e5722. doi: 10.1371/journal.pone.0005722

Brennecke, P., Anders, S., Kim, J. K., Kolodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Meth.* 10, 1093–1095. doi: 10.1038/nmeth.2645

Chen, T. K., Smith, L. M., Gebhardt, D. K., Birrer, M. J., and Brown, P. H. (1996). Activation and inhibition of the AP-1 complex in human breast cancer cells. *Mol. Carcinog.* 15, 215–226. doi: 10.1002/(sici)1098-2744(199603)15:3<215::aid-mc7>3.0.co;2-g

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8:15081. doi: 10.1038/ncomms1508

NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 44, D7–D19. doi: 10.1093/nar/gkv1290

Dembinski, J. L., and Krauss, S. (2009). Characterization and functional analysis of a slow cycling stem cell-like subpopulation in pancreas adenocarcinoma. *Clin. Exp. Metastasis* 26, 611–623. doi: 10.1007/s10585-009-9260-0

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dolatabadi, S., Candia, J., Akrap, N., Vannas, C., Tesan Tomic, T., Losert, W., et al. (2017). Cell cycle and cell size dependent gene expression reveals distinct subpopulations at single-cell level. *Front. Genet.* 8:1. doi: 10.3389/fgene.2017.00001

Dontu, G., Abdallah, W. M., Foley, J. M., Jackson, K. W., Clarke, M. F., Kawamura, M. J., et al. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev.* 17, 1253–1270. doi: 10.1101/gad.1061803

Dulyaninova, N. G., House, R. P., Betapudi, V., and Bresnick, A. R. (2007). Myosin-IIA heavy-chain phosphorylation regulates the motility of MDA-MB-231 carcinoma cells. *Mol. Biol. Cell* 18, 3144–3155. doi: 10.1091/mbc.e06-11-1056

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Eferl, R., and Wagner, E. F. (2003). AP-1: a double-edged sword in tumorigenesis. *Nat. Rev. Cancer* 3, 859–868. doi: 10.1038/nrc1209

Fillmore, C. M., and 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. doi: 10.1186/bcr1982

Ginestier, C., Hur, M. H., 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. doi: 10.1016/j.stem.2007.08.014

Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966

Guha, P., Bandyopadhyaya, G., Polumuri, S. K., Chumsri, S., Gade, P., Kalvakolanu, D. V., et al. (2014). Nicotine promotes apoptosis resistance of breast cancer cells and enrichment of side population cells with cancer stem cell-like properties via a signaling cascade involving galectin-3, α9 nicotinic acetylcholine receptor and STAT3. *Breast Cancer Res. Treat.* 145, 5–22. doi: 10.1007/s10549-014-2912-z

Gyorffy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., et al. (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* 123, 725–731. doi: 10.1007/s10549-009-0674-9

Harrison, H., Farnie, G., Howell, S. J., Rock, R. E., Stylianou, S., Brennan, K. R., et al. (2010). Regulation of breast cancer stem cell activity by signaling through the Notch4 receptor. *Cancer Res.* 70, 709–718. doi: 10.1158/0008-5472.CAN-09-1681

Harrison, H., Rogerson, L., Gregson, H. J., Brennan, K. R., Clarke, R. B., and Landberg, G. (2013). Contrasting hypoxic effects on breast cancer stem cell hierarchy is dependent on er-a status. *Cancer Res.* 73, 1420–1433. doi: 10.1158/0008-5472.CAN-12-2505

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., and Kokocinski, F. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Jen, J., Tang, Y. A., Lu, Y. H., Lin, C. C., Lai, W. W., and Wang, Y. C. (2017). Oct4 transcriptionally regulates the expression of long non-coding RNAs NEAT1 and MALAT1 to promote lung cancer progression. *Mol. Cancer* 16:104. doi: 10.1186/s12943-017-0674-z

Ji, Z., and Ji, H. (2016). TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44:e117. doi: 10.1093/nar/gkw430

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967

Khramtsov, A. I., Khramtsova, G. F., Tretiakova, M., Huo, D., Olopade, O. I., and Goss, K. H. (2010). Wnt/β-catenin pathway activation is enriched in basal-like breast cancers and predicts poor outcome. *Am. J. Pathol.* 176, 2911–2920. doi: 10.2353/ajpath.2010.091125

Kingsley-Kallesen, M. L., Kelly, D., and Rizzino, A. (1999). Transcriptional regulation of the transforming growth factor-β2 promoter by cAMP-responsive element-binding protein (CREB) and activating transcription factor-1 (ATF-1) is modulated by protein kinases and the coactivators p300 and CREB-binding protein. *J. Biol. Chem.* 274, 34020–34028. doi: 10.1074/jbc.274.48.34020

Kreso, A., and Dick, J. E. (2014). Evolution of the cancer stem cell model. *Cell Stem Cell* 14, 275–291. doi: 10.1016/j.stem.2014.02.006

Kruyt, F. A. E., and Schuringa, J. J. (2010). Apoptosis and cancer stem cells: implications for apoptosis targeted therapy. *Biochem. Pharmacol.* 80, 423–430. doi: 10.1016/j.bcp.2010.04.010

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* 526, 131–135. doi: 10.1038/nature15260

Lee, M. C. W., Lopez-Diaz, F. J., Khan, S. Y., Tariq, M. A., Dayn, Y., Vaske, C., et al. (2014). Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4726–E4735. doi: 10.1073/pnas.1404656111

Li, Q., Birkbak, N. J., Gyorffy, B., Szallasi, Z., and Eklund, A. C. (2011). Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics* 12:474. doi: 10.1186/1471-2105-12-474

Li, X., Wang, S., Li, Z., Long, X., Guo, Z., Zhang, G., et al. (2017). The lncRNA NEAT1 facilitates cell growth and invasion via the miR-211/HMGA2 axis in breast cancer. *Int. J. Biol. Macromol.* 105, 346–353. doi: 10.1016/j.ijbiomac.2017.07.053

Liu, S., Dontu, G., Mantle, I. D., Patel, S., Ahn, N. S., Jackson, K. W., et al. (2006). Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells. *Cancer Res.* 66, 6063–6071. doi: 10.1158/0008-5472.can-06-0054

Lyle, S., and Moore, N. (2011). Quiescent, slow-cycling stem cell populations in cancer: a review of the evidence and discussion of significance. *J. Oncol.* 2011:396076. doi: 10.1155/2011/396076

Madjd, Z., Mehrjerdi, A. Z., Sharifi, A. M., Molanaei, S., Shahzadi, S. Z., and Asadi-Lari, M. (2009). CD44+ cancer cells express higher levels of the anti-apoptotic protein Bcl-2 in breast tumours. *Cancer Immun.* 9:4.

Mani, S. A., Guo, W., Liao, M. J., Eaton, E. N., Ayyanan, A., and Zhou, A. Y. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* 133, 704–715. doi: 10.1016/j.cell.2008.03.027

Masson, N., Hurst, H. C., and Lee, K. A. W. (1993). Identification of proteins that interact with CREB during differentiation of f9 embryonal carcinoma cells. *Nucleic Acids Res.* 21, 1163–1169. doi: 10.1093/nar/21.5.1163

Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor creb. *Nat. Rev. Mol. Cell Biol.* 2, 599–609. doi: 10.1038/35085068

Milde-Langosch, K., Röder, H., Andritzky, B., Aslan, B., Hemminger, G., Brinkmann, A., et al. (2004). The role of the AP-1 transcription factors c-Fos, FosB, Fra-1 and Fra-2 in the invasion process of mammary carcinomas. *Breast Cancer Res. Treat.* 86, 139–152. doi: 10.1023/b:brea.0000032982.49024.71

Morel, A. P., Lièvre, M., Thomas, C., Hinkal, G., Ansieau, S., and Puisieux, A. (2008). Generation of breast cancer stem cells through epithelial-mesenchymal transition. *PLoS One* 3:e2888. doi: 10.1371/journal.pone.0002888

Morgan, D. O. (2007). *The Cell Cycle: Principles of Control.* London: New Science Press.

Morishita, A., Zaidi, M. R., Mitoro, A., Sankarasharma, D., Szabolcs, M., Okada, Y., et al. (2013). HMGA2 is a driver of tumor metastasis. *Cancer Res.* 73, 4289–4299. doi: 10.1158/0008-5472.CAN-12-3848

Nakajima, K., Kho, D. H., Yanagawa, T., Harazono, Y., Hogan, V., Chen, W., et al. (2016). Galectin-3 cleavage alters bone remodeling: different outcomes in breast and prostate cancer skeletal metastasis. *Cancer Res.* 76, 1391–1402. doi: 10.1158/0008-5472.CAN-15-1793

Newlaczyl, A. U., and Yu, L. G. (2011). Galectin-3–a jack-of-all-trades in cancer. *Cancer Lett.* 313, 123–128. doi: 10.1016/j.canlet.2011.09.003

Nguyen, A., Yoshida, M., Goodarzi, H., and Tavazoie, S. F. (2016). Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. *Nat. Commun.* 7:11246. doi: 10.1038/ncomms11246

Pastrana, E., Silva-Vargas, V., and Doetsch, F. (2011). Eyes wide open: a critical review of sphere-formation as an assay for stem cells. *Cell Stem Cell* 8, 486–498. doi: 10.1016/j.stem.2011.04.007

Pece, S., Tosoni, D., Confalonieri, S., Mazzarol, G., Vecchi, M., Ronzoni, S., et al. (2010). Biological and molecular heterogeneity of breast cancers correlates with th eir cancer stem cell content. *Cell* 140, 62–73. doi: 10.1016/j.cell.2009.12.007

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* 9, 171–181. doi: 10.1038/nprot.2014.006

Reis-Filho, J. S., and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* 378, 1812–1823. doi: 10.1016/s0140-6736(11)61539-0

Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature* 414, 105–111.

Sampath, P., Pritchard, D. K., Pabon, L., Reinecke, H., Schwartz, S. M., Morris, D. R., et al. (2008). A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* 2, 448–460. doi: 10.1016/j.stem.2008.03.013

Sanchez, C. G., Teixeira, F. K., Czech, B., Preall, J. B., Zamparini, A. L., Seifert, J. R. K., et al. (2016). Regulation of ribosome biogenesis and protein synthesis controls germline stem cell differentiation. *Cell Stem Cell* 18, 276–290. doi: 10.1016/j.stem.2015.11.004

Savage, P., Blanchet-Cohen, A., Revil, T., Badescu, D., Saleh, S. M. I., and Wang, Y. C. (2017). A Targetable EGFR-dependent tumor-initiating

program in breast cancer. *Cell Rep.* 21, 1140–1149. doi: 10.1016/j.celrep.2017.10.015

Schwab, L. P., Peacock, D. L., Majumdar, D., Ingels, J. F., Jensen, L. C., Smith, K. D., et al. (2012). Hypoxia-inducible factor 1α promotes primary tumor growth and tumor-initiating cell activity in breast cancer. *Breast Cancer Res.* 14:R6.

Senkus, E., Kyriakides, S., Penault-Llorca, F., Poortmans, P., Thompson, A., Zackrisson, S., et al. (2013). Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 24(Suppl.6), vi12–vi24. doi: 10.1093/annonc/mdr371

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Takebe, N., Miele, L., Harris, P. J., Jeong, W., Bando, H., Kahn, M., et al. (2015). Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. *Nat. Rev. Clin. Oncol.* 12, 445–464. doi: 10.1038/nrclinonc.2015.61

Takebe, N., Warren, R. Q., and Ivy, S. P. (2011). Breast cancer growth and metastasis: interplay between cancer stem cells, embryonic signaling pathways and epithelial-to-mesenchymal transition. *Breast Cancer Res.* 13:211. doi: 10.1186/bcr2876

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi: 10.1093/nar/gkw1099

Thuault, S., Valcourt, U., Petersen, M., Manfioletti, G., Heldin, C. H., and Moustakas, A. (2006). Transforming growth factor-β employs HMGA2 to elicit epithelial-mesenchymal transition. *J. Cell Biol.* 174, 175–183. doi: 10.1083/jcb.200512110

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *Cancer J. Clin.* 65, 87–108. doi: 10.3322/caac.21262

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* 13, 1977–2000. doi: 10.1091/mbc.02-02-0030

Zhu, J., Zhang, Y., Joe, G. J., Pompetti, R., and Emerson, S. G. (2005). NF-Ya activates multiple hematopoietic stem cell (HSC) regulatory genes and promotes HSC self-renewal. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11728–11733. doi: 10.1073/pnas.0503405102

Zhu, W., Giangrande, P. H., and Nevins, J. R. (2004). E2Fs link the control of G1/S and G2/M transcription. *EMBO J.* 23, 4615–4626. doi: 10.1038/sj.emboj.7600459