## Title

Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells

**Permalink**
https://escholarship.org/uc/item/0x8772xv

**Journal**
Nature, 526(7571)

**ISSN**
0028-0836

**Authors**
Lawson, Devon A
Bhakta, Nirav R
Kessenbrock, Kai
et al.

**Publication Date**
2015-10-01

**DOI**
10.1038/nature15260

**Supplemental Material**
https://escholarship.org/uc/item/0x8772xv#supplemental

Peer reviewed

*Paper draft*

# Single-cell analysis reveals a distinct stem cell program in early human metastatic breast cancer cells

Devon A. Lawson[1], Nirav Bhakta[2], Kai Kessenbrock[1,3], Karin Prummel[1,†], Ying Yu[1], Ken Takai[1,‡], Alicia Zhou[3], Henok Eyob[3], Paul Yaswen[4], Alana Welm[5], Andrei Goga[2,3]*, Zena Werb[1*]

[1]Department of Anatomy, [2]Medicine, [3]Cell and Tissue Biology, University of California, San Francisco,
[4]Department of Cancer and DNA Damage Responses, Lawrence Berkeley National Laboratory,
[5]Department of Oncological Sciences, University of Utah,

Present addresses:
[†]Cancer Genetics and Developmental Biology, Utrecht University, Netherlands;
[‡]Laboratory of Virus Control, Institute for Virus Research, Kyoto University, Japan;

*Address correspondence to:

Zena Werb, Ph.D.
Department of Anatomy
HSW1323
University of California
513 Parnassus Ave.
San Francisco, CA 94143-0452
Tel. 415-476-4622
Fax  415-476-4565
email: zena.werb@ucsf.edu

Andrei Goga, MD, Ph.D.
Department of Cell and Tissue Biology
Department of Medicine
HSW601
University of California
513 Parnassus Ave.
San Francisco, CA 94143-0452
Tel. 415-476-4622
Fax  415-476-4565
email: Andrei.Goga@ucsf.edu

Running title: Early DTCs as stem cells

**Despite major advances in understanding the molecular and genetic basis of cancer, disease progression to metastasis remains the cause of >90% of cancer-related mortality[1]. Understanding the mechanisms underlying metastasis initiation is critical for the development of new therapeutic strategies to specifically treat and prevent progression to metastatic disease. Dogma suggests that metastatic lesions are seeded by rare tumor cells with unique biological properties[2]. This is supported by studies in human colon and pancreatic cancer showing metastases are initiated by cells with phenotypic and functional properties of cancer stem cells (CSCs).[3-5] However, the identity of metastasis-initiating cells in human breast cancer remains elusive[2]. Here we show at the single-cell level that early stage human disseminated tumor cells (DTCs) possess a distinct stem cell-like gene expression signature. By developing a highly sensitive FACS-based assay to isolate DTCs from patient-derived xenograft (PDX) models of human breast cancer, we were able to compare gene expression patterns in DTCs from early vs. later stage mice. While early DTCs express basal/stem cell, EMT, pro-survival, and dormancy-associated genes, later DTCs express higher levels of markers associated with luminal cell differentiation, exit from dormancy and entry into cell cycle, and more closely resemble primary tumor cells. This led us to test whether treatment with cyclin dependent kinase (CDK) inhibitors to block cell cycle progression could attenuate metastatic progression. CDK inhibition resulted in a significant decrease in the number of DTCs detected in PDX mice, and also blocked seeding and colonization of MDA-MB-231 breast cancer cells in drug treated mice. These findings support a hierarchical model for metastatic cell initiation and progression, and present the cell cycle program as an attractive new target for the management of metastatic disease.**

To investigate the differentiation state of early stage metastatic cells, we utilized a microfluidics-based platform (Fluidigm) for multiplex gene expression analysis in individual cells. This facilitated a systems-level approach, where we were able to study the simultaneous expression of groups of genes, and resolve cellular diversity only achievable at the single-cell level. We designed dynamic arrays to examine genes involved in stem cell function, differentiation, cell cycle and dormancy, epithelial-to-mesenchymal transition (EMT), and signaling pathways commonly dysregulated in breast cancer.

To generate a reference for analyzing metastatic cells, we developed a single-cell gene expression signature from normal breast epithelium from reduction mammoplasty patients. The breast contains two epithelial lineages: the basal/myoepithelial lineage that contains stem cells, and a luminal lineage that contains progenitor and mature cell populations. We sorted single basal/stem, luminal, and luminal progenitor cells from three individuals and processed them according to established protocols (Fig. 1a).[6-9] As expected, Principal component analysis (PCA) and unsupervised hierarchical clustering showed that basal and luminal cells represent distinct populations in each individual (Fig. 1b,d). 49 of the 123 genes tested showed differential expression between the populations, and were used to generate a 49-gene differentiation signature to distinguish them. This signature included established lineage-specific genes such as KRT5, TP63, MUC1, CD24, and GATA3 (Fig. 1c, Extended Data Fig. 1 and Extended Data Table 1), validating our multiplex PCR approach. Of note, we observed significant heterogeneity within the luminal lineage, where we identified at least two distinct populations: i) stereotypic 'mature' luminal cells high for luminal genes and low for basal/stem genes, and ii) 'intermediate' cell populations showing varying levels of basal/stem and luminal genes. Further interrogation of diversity in this lineage may reveal cells with distinct roles in mammary gland differentiation and homeostasis.

We analyzed mice from ten breast cancer PDX models. In this study, we focused on the triple negative subtype since it is the most aggressive, metastasis is frequent, and there are no targeted therapeutics to treat it.[10] We report data from three triple-negative (ER-PR-HER2-), basal-like models here (HCI-001, HCI-002, and HCI-010) (Fig. 2a). These PDX models maintain the essential properties of the original patient tumors, including histopathology, clinical markers, global gene expression patterns, hormone responsiveness, and metastatic tropism, making them authentic experimental systems for studying human cancer metastasis.[11]

To isolate DTCs from PDX mice, we first developed a highly sensitive, species-specific FACS-based assay. We annotated published microarray data to identify cell surface genes highly expressed in PDX breast cancer cells.[11] This revealed as a top candidate CD298 (ATP1B3), which is a beta subunit of Na+/K+ ATPases that are essential for basic cellular function (Extended Data Fig. 2a)[12]. Using a human species-specific antibody, we found that CD298 is expressed by >99% of cells in three different human mammary cell lines, with no background in mouse lines or control mouse peripheral tissues (Fig. 2d and Extended Data Fig. 2b). In dissociated PDX primary tumors, CD298 was expressed by the majority (60-95%) of cells (Fig. 2b) and was superior to other common markers, such as EpCAM, CD24, and MHCI (Extended Data Fig. 2c). We therefore expected this assay would capture the majority of DTCs in PDX mice with negligible false-positive rates.

We detected DTCs in peripheral tissues of 44/72 (61%) PDX mice using this assay, including the lung, lymph node, bone marrow, liver, brain, and peripheral blood (Fig. 2a,d). We analyzed DTCs from 21 mice, and report comprehensive analysis of nine here and five different distant sites (Extended Data Table 2). Since our goal was to study metastasis-initiating cells (MICs), we hypothesized that early stage mice with very low numbers of cells in distant tissues would be enriched for putative MICs. Conversely, tissues with large numbers of DTCs were more likely to contain later stage, growing metastatic lesions. We designated tissues containing (i) <250 DTCs as 'early', (ii) tissues with an intermediate number (300-800) as 'mid', and (iii) tissues with $>10^3$ as 'late' (Fig. 2d). A similar range of metastatic lesions were found by histology in the lungs (Fig. 2c). DTCs in the peripheral blood and bone marrow were not designated since they may be transient and rarely develop into overt metastases. Gene expression levels in DTCs were compared to primary tumor cells from the same animal, to focus on differences between DTCs and the primary tumor cells they originate from.

Remarkably, PCA plots for individual animals showed that 'early' DTCs were more distinct from primary tumor cells than 'mid' and 'late' DTCs. (Fig. 3a). This was also observed by unsupervised hierarchical clustering of pooled data from all nine animals, which showed that 'early' DTCs form a unique cluster, while 'mid' and 'late' DTCs cluster more closely with primary tumor cells (Extended Data Fig. 3). Most striking was the finding that this was due to a conserved basal/stem cell signature in 'early' DTCs across all animals and models. Analysis of genes comprising the 49-gene differentiation signature showed that 'early' DTCs expressed higher levels of 22 basal/stem cell genes, including LGR5, BMI1, BCL2, NOTCH4, and JAG1. They also expressed lower levels of seven luminal genes, including MUC1, EMP1, and CD24. 'Late' DTCs, by contrast, more closely resembled primary tumor cells. 'Early' DTCs also expressed very high levels of the pluripotency genes OCT4, and SOX2, suggesting they may exploit embryonic programs for self-renewal and maintenance (Fig. 3b). 'Early' DTCs may also utilize an EMT program, which has been shown to promote stemness in the mammary gland[13,14]. 'Early' DTCs expressed higher levels of SNAI2, SKP2, and TWIST1/2, and lower levels of CDH1, which is consistent with an EMT phenotype and was also observed in normal basal/stem cells, with the exception of TWIST1/2 (Fig.

1c and Extended Data Table 3). However, ZEB1 and ZEB2 were not differentially expressed in ‘early’ DTCs, and VIM was substantially lower (Extended Data Table 3). “Late” DTCs did not possess this distinct EMT profile and more closely resembled primary tumor cells (Fig. 3c).

Focusing on clustering of the DTCs, we further found that they organize into a hierarchy reminiscent of the normal mammary gland. ‘Early’ DTCs formed one distinct cluster and possessed a basal/stem-like expression pattern. ‘Late’ DTCs formed another cluster distant from ‘early’ DTCs and expressed a more luminal-like pattern (Fig. 3c). A third cluster localized between them, which was comprised of many ‘mid’ stage DTCs that possessed an intermediate phenotype between basal/stem- and luminal-like. These findings suggest that early, putative metastasis-initiating cells possess a basal/stem-like differentiation state and progressively differentiate to a more luminal-like state as they proliferate and colonize. Extended Data Fig. 4 and Extended Data Table 3 show all 55 genes differentially expressed in ‘early’ DTCs relative to primary tumor cells.

We also discovered several specific differences between DTCs from different tissues. Brain DTCs were the most distinct, both from primary tumor cells and other DTCs. They were the most homogeneous and formed their own cluster (Fig. 3c, green), and expressed the highest levels of several stem cell, quiescence, and anti-apoptosis genes including LGR5, BMI1, TGFB2, CDKN1B, and BCL2L1 (Extended Data Fig. 5a). They also expressed the lowest levels of differentiation genes (CD24, MUC1, KRT19) and pro-proliferation genes like cMYC (Fig. 3c). Although only rare CTCs could be recovered, they most closely resembled lung DTCs ($r^2$=0.217), were least similar to brain DTCs ($r^2$=0.027), and expressed high levels of VEGFA, JAG1, and SNAI2 (Extended Data Fig. 5b). Bone marrow DTCs were distinguished by their heightened expression of TGFBR2 and ITPKB1 (Extended Data Fig. 5a). Of note, most CTCs and bone marrow DTCs localized to the intermediate cluster with ‘mid’ stage DTCs (Fig. 3c), raising questions regarding their relationship to early stage DTCs and role in metastasis initiation.

Interestingly, we also observed a shift towards a proliferative signature associated with metastatic progression. ‘Early’ DTCs expressed higher levels of quiescence and dormancy-associated genes, including CDKN1B, TGFBR2, TGFBR3, TGFB2, and MTOR (Fig. 3c, Fig 4b, and Extended Data Fig. 4 and Table 4).[15,16] ‘Late’ DTCs appeared to enter cell cycle, expressing lower levels of quiescence and dormancy-associated genes and higher levels of cell cycle-promoting genes such as cMYC and CDK2, as well as MMP1 and CD24, which have been associated with reactivation following dormancy. Moreover, unsupervised hierarchical clustering showed ‘early’ and ‘late’ DTCs in distinct clusters based on differential expression of these genes (Fig. 4c).

These findings prompted us to test whether blocking this switch from dormancy into cycle could inhibit metastatic progression in our models. Management of residual disease by maintaining dormancy, or inducing apoptosis of reactivated DTCs have been proposed as viable treatment strategies in patients, but have not yet been thoroughly tested in this setting[17,18]. Since we observed high levels of both cMYC and CDK2 in later stage metastatic cells, we chose to test Dinaciclib, a CDK inhibitor that selectively targets CDK1, CDK2, CDK5, and CDK9. We have also shown that Dinaciclib can target high Myc-expressing cancer cells through synthetic lethality, and several reports suggest additional efficacy through bromodomain inhibition of Myc regulatory factors.[19-21] We tested Dinaciclib on a total of 49 mice from two PDX models, HCI-001 and HCI-002, which were from drug-naïve patients. Drug was administered thrice weekly (30 mg/kg) for four weeks starting when tumors first became palpable (Fig. 4d). We analyzed the mice for DTCs at the conclusion of the four week course, or earlier if tumors reached endpoint (20 mm). Remarkably, we found that only 1 of 24 drug treated animals developed DTCs, in comparison to 44% (11/25) of vehicle treated mice (Fig. 4d). Metastatic burden was also lower in treated vs. control mice (Fig. 4d). To test whether Dinaciclib could specifically block early phases of metastasis, we utilized an i.v. injection model for experimental metastasis of MDA-MB-231 triple-negative breast cancer cells. Luciferase and GFP-labeled cells were injected i.v. on day 0, followed by three daily i.p. injections of drug. When animals were imaged on day 17, we found significantly less luciferase signal in drug treated mice (Fig. 4e). Furthermore, we found only very rare (<3) GFP+ foci in the lungs of drug treated animals (Extended Data Fig. 5c), in contrast to control mice which developed widespread metastases in the lungs, liver, and peritoneum. This suggests that Dinaciclib induces death in metastasizing tumor cells, rather than simply by delaying growth. Since Dinaciclib is already being tested in clinical trials for multiple cancer types, including a Phase I trial for triple-negative breast cancer[22], further experiments to define its mechanism of action against metastasis may be of immediate clinical benefit.

By investigating gene expression in animals at varying stages of metastasis, we have found clear evidence that earlier stage DTCs possess basal/stem cell characteristics, and later stage cells express more luminal-like ones. We also found that the majority of cells from the primary tumor and circulation are luminal-like, with rare cells that appear more stem-like (2/>350 primary tumor cells, Extended Data Fig. 3; 1/17 CTCs, Fig. 3c). This is consistent with a model where rare stem-like cells escape the primary tumor, survive circulation, and seed in a distant tissue where they eventually become competent for proliferation and colonization to produce lesions that recapitulate the cellular hierarchy in the primary tumor (Extended Data Fig. 6). However, further experiments will be necessary to thoroughly define the ontogeny of ‘early’ stem-like DTCs.

## Methods

### Animal studies

The University of California, San Francisco Institutional Animal Care and Use Committee reviewed and approved all animal experiments. Following established protocols by Welm and colleagues[11], ~8mm$^3$ tumor fragments from frozen PDX tumors were prepared and individually transplanted into the cleared inguinal fat pads of pre-pubescent NOD/SCID mice. Tumor fragments were stored by freezing in 90% FBS and 10% DMSO in liquid nitrogen. Clinical details of patients the PDX models are derived from are included in DeRose et al., (2011)[11]. All animals for Fluidigm experiments were euthanized when tumors reached 2-2.5cm.

### Dinaciclib treatment experiments and IVIS imaging

Dinaciclib was prepared and administered according to previously established protocols in mice[19,23]. Dinaciclib was reconstituted in 20% HPBCD (hydroxypropyl β cyclodextrin). For PDX animal studies, drug treatment course was initiated when tumors became palpable. A total of 49 animals were treated by i.p. injection three times per week at 30 mg/kg of drug, or vehicle (HPBCD). Mice were euthanized at the conclusion of a 4-week treatment course, or earlier if their tumors reached 20 mm in diameter. Animals that developed adverse effects (e.g., >20% weight loss) were excluded from the study. Statistical significance between drug and vehicle treated groups was examined by two-tailed, un-paired t-test. For experimental metastasis experiments, two cohorts of 10 animals (5 per treatment group) were injected i.v. with $10^5$ MDA-MB-231 cells on day 0 and given three consecutive daily treatments (day 0, 1, and 2) of

30mg/kg Dinaciclib or vehicle by i.p. injection. Animals were imaged by i.p. injection of firefly D-luciferase (Gold Biotechnology) using an IVIS Spectrum. Emitted light signal was quantified by drawing Regions of Interest (ROIs) around the lung region of each animal in Living Image software (Caliper LifeScience). Statistical significance between treatment groups was assessed by a two-tailed, un-paired t-test.

**Microarray analysis**

Published microarray data from Welm and colleagues[11] was analyzed for plasma membrane genes highly expressed across all 15 PDX tumor samples and 12 original patient tumor samples included in their study. Genes were rank ordered for expression across all the samples using x software. CD298/ATP1B3 ranked number 35 out of over 6,000 genes annotated.

**Tissue dissociation**

All solid tissues, including primary tumor, lungs, lymph nodes (axillary, brachial, cervical, sciatic, and lumbar) and brain were dissociated for FACS using the same protocol[24]. Briefly, tissues were mechanically chopped with a scalpel, placed in culture medium (DMEM/F12 with 5 ng/ml insulin (UCSF Cell Culture Facility), 50 ng/ml gentamycin (UCSF Cell Culture Facility) containing 2 mg/ml collagenase-1 (Sigma). They were then digested for 45 min at 37°C. The resulting suspension was resuspended in 2 U/µl DNAse for 3 min at RT, washed and dissociated with 2 ml 0.05% trypsin/EDTA (UCSF Cell Culture Facility) for 10 min at 37°C and filtered through a 70 µm filter. Erythrocytes in lung and tumor samples were lysed with Red Blood Cell Lysis Buffer for 5 min at room temperature.

**Flow cytometry**

Antibodies for the human antigens CD45 (Alexa-450, eBioscience), Ter119 (Alexa-450, eBioscience), CD31 (Alexa-450, eBioscience), CD298 (PE, Biolegend), EpCAM (PE, eBioscience), CD49f (APC, eBioscience), and CD117/cKit (FITC, eBioscience) were purchased commercially. For mouse antigens, CD45 (FITC, eBioscience), Ter119 (FITC, eBioscience), and CD31 (FITC, eBioscience) were used. Antibody staining was performed in DMEM/5%FBS supplemented with penicillin and streptomycin. After 15 min on ice, stained cells were washed of excess unbound antibodies and resuspended in media. Flow sorting was done using a BD FACSAriaII cell-sorter (Becton Dickinson), and analysis was done on an LSRII (Becton Dickinson). Forward-scatter height versus forward-scatter width (FSC-H versus FSC-W) and side-scatter area versus side-scatter width (SSC-A versus SSC-W) were used to eliminate cell aggregates and ensure single cell sorting. Dead cells were eliminated by excluding Sytox positive (SYTOX Blue dead cell stain, Molecular Probes) cells, which increased the efficiency of sorting robust, live cells for single-cell experiments. Contaminating human or mouse hematopoietic and endothelial cells were excluded by gating out Lin+ (CD45, Ter119, CD31) cells. In Fluidigm experiments where the number of DTCs identified was listed (Extended Data Table 2), the entire tissue sample was run. A consistent number of live cells was found in tissues from each animal. In any case where live cell yields were below average, mice were excluded from the study. In Fig 2a, animals or tissues were designated as positive for DTCs if >10 hCD298+mLin- cells were identified in the entire sample.

**Fluidigm dynamic array experiments**

Single-cell gene-expression experiments were performed using Fluidigm's 96.96 qPCR DynamicArray microfluidic chips. Single cells were sorted by FACS into individual wells of 96-well PCR plates, using the FACSAriaII single-cell sorting protocol with specific adjustments (device: 96-well PCR plate; precision: single-cell; nozzle: 100 µm). Experiments were performed according to Fluidigm's Advanced

Development Protocol 41. Each well of 96-well PCR plates were preloaded with 9 µl volume of RT-STA solution: 5 µl of CellsDirect PCR mix (Invitrogen), 0.2 µl of SuperScript-III RT/Platinum Taq mix (Invitrogen), 1.0 µl of a mixture of all pooled primer assays (500nM), and 2.8 µl of DNA suspension buffer (TEKnova). All primer assays used in array experiments are listed in Supplemental Information (Table 1). After sorting, PCR plates were frozen (-20°C) or placed into a thermocycler for combined reverse transcription (50°C for 15 min, 95°C for 2 min) and target-specific amplification (each cycle: 95°C for 15 s, 58°C for 4 min). 3.6 µl of Exonuclease reaction solution (2.52 µl $H_2$0, 0.36 Exo reaction buffer, and 0.72 µl ExoI, New England BioLabs) was then added to remove unincorporated primers (37°C for 30 min, 80°C for 15 min). Subsequently, each well was diluted 1:3 with TE buffer (TEKnova). In a separate plate, a 2.7 µl aliquot from each sample well was then mixed with 2.5 µl of SsoFast EvaGreen Supermix with Low Rox (Bio-Rad) and 0.25 µl of Fluidigm's DNA Binding Dye Sample Loading Reagent and plates were centrifuged to mix solutions. In another separate plate, each primer assay mix was generated by loading 2.5 µl of Assay Loading Reagent (Fluidigm), 2.25 uL DNA Suspension Buffer, and 0.25 µl of each 100 µM primer pair mix. Before loading primer assays and sample mixes into each chip, chips were primed by injecting control line fluid (Fluidigm) and running the 'Prime' program in the IFX Controller HX. After priming, 5 µl of each sample and primer mix were loaded into each well of the chips. Samples and assays were then mixed in the chip by running the 'Load Mix' program in the IFC Controller HX. Chips were transferred into the BioMark real-time PCR reader (Fluidigm) and run according to the manufacturer's instructions. A list of the 123 primer assays used in this study is provided in Supplemental Information. All primer sequences were acquired through the Harvard Primer bank, and synthesized by Integrated DNA Technologies. Thorough technical evaluations of the microfluidics chip technology, limits of detection, and efficiency of multiplex PCR in this platform have been reported by Fluidigm and several independent reports[25-27].

## Computational analysis, display, and statistical assessment of single-cell PCR datasets

All single-cell PCR data was analyzed using Fluidigm's Real-time PCR analysis software, using the Linear (Derivative) and User (Detectors) settings to generate Ct values for each gene. Ct values were further processed in the R statistical language, using algorithms we generated. All code is provided in the Supplemental Information, and published in Github (https://github.com/) for upload into R. 268 normal mammary cells from reduction mammoplasty samples, and 399 DTCs and 383 primary tumor cells from PDX mice were analyzed. In normal mammary cell experiments, Ct values were normalized by subtracting the average value of the basal/stem cell population on a per-gene, per-array basis to correct for batch-to-batch differences in reverse transcription, pre-amplification, and real-time PCR. In PDX experiments, Ct values were normalized by subtracting the average primary tumor expression from the same individual animal on a per-gene basis, in order to identify conserved differences in gene expression in DTCs relative to the primary tumor cells they derive from, in addition to correction of batch-to-batch differences. Normalization using housekeeping genes was not performed, as is it not recommended for single-cell qPCR. Normalized Ct values were converted to relative log2 expression values simply through multiplication by -1. Low-quality samples were identified and removed from further analysis if less than 80% of the assayed genes amplified. Gene expression data was displayed by PCA, unsupervised hierarchical clustering, and box plots. Clustering was performed on both cells and genes, based on Pearson's correlation distance metric and average linkage. For all clustering heatmaps, the blue/red color scale is set from -1.5 to 1.5 corresponding to log2 gene expression. For PCA, in which missing data are not easily accommodated, failed reactions were set to a value 1 lower than the lowest observed value across all samples for each gene separately.

To identify gene expression differences between pre-defined populations, several statistical tests were performed. For normal mammary cell experiments, we first performed three-group comparisons between basal/stem, luminal, and luminal progenitor cells (both parametric: ANOVA; and non-parametric: Kruskal-Wallis). This yielded a list of 49 differentially expressed genes (Fig. 1c and Extended Data Table 1). To determine which genes were characteristic of each population, we subsequently performed pair-wise tests (parametric: moderated t-test (ref); and non-parametric: Mann-Whitney U test). In DTC vs. T experiments, only pair-wise and no three-group comparisons were performed. Our algorithm selected the most appropriate test from which to report a p-value based on the type of data observed for that gene (non-parametric if >50% of samples failed for either group, parametric otherwise). This criterion was chosen in an attempt to prevent a high proportion of failed values masking group differences. All p-values were also adjusted for the fact that many genes were being simultaneously analyzed by controlling the false discovery rate (FDR) with the Benjamini-Hochberg method. To identify basal/stem cell characteristic genes, we compared basal/stem (B) to both luminal (L) and luminal progenitor cells (LP) (i.e., B vs. (L+LP)). Luminal genes were identified by performing L vs. B, and luminal progenitor genes by performing LP vs. L (since they are a subset of the L lineage). As we are only analyzing assays for which at least 1 cell yielded amplification, undetectable amplification represents non-expression rather than technical error in the PCR reaction. In order to capture non-expression in the statistical tests, failed reactions were set to a value 0.01 lower than the lowest observed value across all samples for each gene separately. For the non-parametric tests above, the specific value chosen is not important, while for the parametric tests, this method is comparable to using a lower limit of detection.

### Histological analysis

For histological analyses, tissues were fixed overnight in 4% paraformaldehyde and processed for paraffin embedding. Sections were stained with hematoxylin and eosin (H&E) using standard methods.

## Supplemental Information

### Table 1: Primer assays used in Fluidigm dynamic array experiments

### Document 1: Code (R statistical language) for analysis of single-cell real-time qPCR experiments using Fluidigm dynamic arrays

## Figure legends

**Figure 1: Single-cell analysis of normal human mammary epithelial cells. a)** FACS plots show basal/stem (Lin$^{neg}$CD49f$^{hi}$EpCAM$^{lo}$cKit$^{neg}$, blue), luminal (Lin$^{neg}$CD49f$^{lo}$EpCAM$^{hi}$cKit$^{neg}$, yellow), and luminal progenitor (Lin$^{neg}$CD49f$^{med}$EpCAM$^{med}$cKit$^{pos}$, red) cells from a reduction mammoplasty patient. (Lin=human CD45/Ter119/CD31) **b)** PCA plots show distinct cell populations identified in three patients. **c)** Bar graph shows 49 genes differentially expressed between the populations The p-value and fold-change for each gene are listed in Extended Data Table 1. B=basal/stem, LP=Luminal progenitor, and L=Luminal. **d)** Heatmap and dendogram show unsupervised hierarchical clustering of individual cells and genes from the 49-gene signature that were run on all arrays.

**Figure 2: Identifiation of human DTCs in PDX mice. a)** Dissemination frequencies and tropism by FACS in each model. **b)** FACS plots show hCD298+mLin- human cancer cells in tumors from each model (n>8), and control mammary gland (n=3). **c)** H&E stains show lesions of varying size and cell number in lung tissues of PDX mice. (top bar = 100 µm; bottom bar = 200 µm) **d)** FACS plots show dissemination patterns of hCD298+mLin- DTCs in representative mice, which exemplify (i) 'early' (<250), (ii) 'mid' (300-800), and (iii) 'late' (>10$^4$) stage designations. Top panels show no background was observed in control mice (n>3).

**Figure 3: 'Early' DTCs possess a distinct basal/stem cell program. a)** PCA plots show 'early', 'mid', and 'late' DTC and primary tumor cell populations in five representative mice. **b)** Bar graph shows genes from the 49-gene differentiation signature that were differentially expressed in 'early' DTCs, relative to primary tumor (T=0). Pluripotency genes SOX2 and OCT4 are also shown. (*$p<0.05$; p-values and fold change are listed in Extended Data Table 3) **c)** Heatmap and dendogram show unsupervised hierarchical clustering of DTCs and each gene from the 49-gene signature that were run on all arrays. LU=Lung, LN=Lymph node, BM=Bone marrow, PB=Peripheral blood, BR=brain

**Figure 4: Metastatic progression is blocked by cell cycle inhibition. a)** Schematic of cell cycle and dormancy/quiescence. **b)** Bar graph shows expression for dormancy and cell cycle regulatory genes in 'late' relative to 'early' DTCs. (*$p<0.05$) **c)** Unsupervised hierarchical clustering of DTCs and regulatory genes. **d)** Dinaciclib treatment course in PDX mice (top). Bar graphs show percent of mice with DTCs, and metastatic burden per animal (*$p<0.05$). **e)** Dinaciclib treatment course in mice i.v. injected with MDA-MB-231 cells (top). IVIS images (bottom left) show luciferase signal in vehicle vs. drug treated mice. Bar graph (bottom right) shows quantification of luciferase signal (*$p<0.05$).

**Extended Data Figure 1: Gene expression in individual basal/stem, luminal, and luminal progenitor cells.** Box plots show expression levels for each gene from the 49-gene differentiation signature in individual cells. P-values and fold change for each gene are shown in Extended Data Table 1. Black dots=single cells; red dots=single cells with no expression.

**Extended Data Figure 2: Identification and validation of CD298 for detection of human cells. a)** Analysis of published microarray data identifies CD298 as highly expressed on many PDX breast cancer models and corresponding original patient tumors. The heatmap shows genes rank ordered from highest to lowest for raw expression values across all samples. The inset highlights expression for CD298. **b)** FACS for CD298 on human and mouse mammary cell lines to establish species specificity. **c)** FACS on primary PDX tumors comparing CD298 expression with other markers used in related applications (EpCAM, CD24, MHCI; percentages indicate dual positive cells). EpCAM is used to identify CTCs in the clinic; CD24 is a pan-epithelial marker; and MHCI is used as a ubiquitous marker on all nucleated cells. These markers were not used in this study because they were not robustly expressed on all PDX models.

**Extended Data Figure 3: 'Early' DTCs are distinct from 'late' DTCs and primary tumor cells.** Unsupervised hierarchical clustering of DTCs and primary tumor cells and genes from the 49-gene signature that were run on all arrays. Dendogram shows a distinct cluster (left) containing the majority of 'early' DTCs. Most 'mid' and 'late' DTCs cluster with primary tumor cells (right).

**Extended Data Figure 4: Gene expression in individual 'early,' 'mid,' 'late,' and primary tumor cells.** Box plots show expression levels for each differentially expressed gene ($p<0.05$) in individual cells. Statistics for each gene are shown in Extended Data Table 3. Black dots=single cells; red dots=single cells with no expression.

**Extended Data Figure 5: a)** Box plots show genes differentially expressed genes in brain, peripheral blood and bone marrow DTCs. Each colored dot represents expression in a single cell, and grey dots represent non-expressing cells. Y-axis shows log2 fold change relative to primary tumor. **b)** Pearson correlations indicate similarity of CTCs to other DTC types across all genes analyzed. Each dot represents an individual gene. LU=Lung, LN=Lymph node, BM=Bone marrow, PB=Peripheral blood (CTC), BR=brain **c)** Fluorescene microscopy images (left) of lungs from mice injected i.v. with MDA-MB-231 cells, and treated with either vehicle or Dinaciclib. Bar graph (right) enumerates GFP+ metastatic lesions in the lungs of vehicle and drug treated mice.

**Extended Data Figure 6: Model for tumor cell dissemination and progression.** Based on gene expression signatures we have generated from metastatic cells of varying stages, we propose a model where 'early' DTCs (blue) are enriched for genes involved in stemness, protection from apoptosis, dormancy/quiescence, and EMT (top panel). By contrast, later stage cells (yellow/red) are more heterogeneous, expressing higher levels of genes charactersitic of mammary differentiation, dormancy exit and proliferation, and lower levels of EMT genes; which more closely recapitulates the heterogeneity observed in primary tumor cells. This strongly suggests that 'early' stage DTCs – and putative MICs – differentiate to a more luminal-like state as they proliferate and colonize. The origin of 'early' DTCs, however, remains unclear. Prevailing theories suggest that MICs are present as a rare subpopulation in primary tumors and are selected during the metastatic process due to their unique biology[2,28,29]. We found evidence of rare 'early' DTC-like cells amongst the primary tumor cells analyzed (2/>350 primary tumor cells cluster with 'early' DTCs; Extended Data Fig. 2). Furthermore, 1 of 17 CTCs analyzed clustered with 'early' DTCs (Fig. 3c). The identification of these cells at low frequency in the primary tumor and circulation is most consistent with a model where rare, stem-like primary tumor cells escape the primary tumor, enter circulation, and seed in distant tissue sites (lower panels). This is also supported by previous work showing that invasive 'leader' cells on the peripheral edges of primary tumors uniquely express basal/stem cell markers[30]. However, alternative models where differentiated primary tumor cells invade and de-differentiate in metastatic sites (e.g. through interaction with the new microenvironment) cannot be excluded. New technologies will be needed in order to screen large numbers of primary tumor cells to distinguish between these alternative models for the origin of ‘early’ stem-like DTCs.

**Extended Data Table 1: Genes comprising the 49-gene differentiation signature that are differentially expressed between normal human mammary epithelial populations**

**Extended Data Table 2: Designation of each tissue as ‘early’, ‘mid’, or ‘late’ based on the number of DTCs detected**

**Extended Data Table 3: All genes differentially expressed in ‘early’ DTCs relative to primary tumor cells**

## References

1 Weigelt, B., Peterse, J. L. & van 't Veer, L. J. Breast cancer metastasis: markers and models. *Nature reviews. Cancer* **5**, 591-602, doi:10.1038/nrc1670 (2005).

2 Oskarsson, T., Batlle, E. & Massague, J. Metastatic stem cells: sources, niches, and vital pathways. *Cell stem cell* **14**, 306-321, doi:10.1016/j.stem.2014.02.002 (2014).

3 Hermann, P. C. *et al.* Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. *Cell stem cell* **1**, 313-323, doi:10.1016/j.stem.2007.06.002 (2007).

4 Pang, R. *et al.* A subpopulation of CD26+ cancer stem cells with metastatic capacity in human colorectal cancer. *Cell stem cell* **6**, 603-615, doi:10.1016/j.stem.2010.04.001 (2010).

5 Dieter, S. M. *et al.* Distinct types of tumor-initiating cells form human colon cancer tumors and metastases. *Cell stem cell* **9**, 357-365, doi:10.1016/j.stem.2011.08.010 (2011).

6 Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast cancer research : BCR* **14**, R134, doi:10.1186/bcr3334 (2012).

7 Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84-88, doi:10.1038/nature04372 (2006).

8 Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993-997, doi:10.1038/nature04496 (2006).

9 Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature medicine* **15**, 907-913, doi:10.1038/nm.2000 (2009).

10 Dent, R. *et al.* Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 4429-4434, doi:10.1158/1078-0432.CCR-06-3045 (2007).

11 DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nature medicine* **17**, 1514-1520, doi:10.1038/nm.2454 (2011).

12 Malik, N., Canfield, V. A., Beckers, M. C., Gros, P. & Levenson, R. Identification of the mammalian Na,K-ATPase 3 subunit. *The Journal of biological chemistry* **271**, 22754-22758 (1996).

13 Mani, S. A. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704-715, doi:10.1016/j.cell.2008.03.027 (2008).

14 Guo, W. *et al.* Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* **148**, 1015-1028, doi:10.1016/j.cell.2012.02.008 (2012).

15 Bragado, P. *et al.* TGF-beta2 dictates disseminated tumour cell fate in target organs through TGF-beta-RIII and p38alpha/beta signalling. *Nature cell biology* **15**, 1351-1361, doi:10.1038/ncb2861 (2013).

16 Kim, R. S. *et al.* Dormancy signatures and metastasis in estrogen receptor positive and negative breast cancer. *PloS one* **7**, e35569, doi:10.1371/journal.pone.0035569 (2012).

17 Aguirre-Ghiso, J. A., Bragado, P. & Sosa, M. S. Metastasis awakening: targeting dormant cancer. *Nature medicine* **19**, 276-277, doi:10.1038/nm.3120 (2013).

18 Giancotti, F. G. Mechanisms governing metastatic dormancy and reactivation. *Cell* **155**, 750-764, doi:10.1016/j.cell.2013.10.029 (2013).

19 Horiuchi, D. *et al.* MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. *The Journal of experimental medicine* **209**, 679-696, doi:10.1084/jem.20111512 (2012).

20 Martin, M. P., Olesen, S. H., Georg, G. I. & Schonbrunn, E. Cyclin-dependent kinase inhibitor dinaciclib interacts with the acetyl-lysine recognition site of bromodomains. *ACS chemical biology* **8**, 2360-2365, doi:10.1021/cb4003283 (2013).

21 Delmore, J. E. *et al.* BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* **146**, 904-917, doi:10.1016/j.cell.2011.08.017 (2011).

22 Jo Chien, U. o. C., San Francisco; Merck Sharp & Dohme Corp. *ClinicalTrials.gov*, <http://clinicaltrials.gov/show/NCT01676753> (

23 Parry, D. *et al.* Dinaciclib (SCH 727965), a novel and potent cyclin-dependent kinase inhibitor. *Molecular cancer therapeutics* **9**, 2344-2353, doi:10.1158/1535-7163.MCT-10-0324 (2010).

24 Welm, B. E., Dijkgraaf, G. J., Bledau, A. S., Welm, A. L. & Werb, Z. Lentiviral transduction of mammary stem cells for analysis of gene function during development and cancer. *Cell stem cell* **2**, 90-102, doi:10.1016/j.stem.2007.10.002 (2008).

25 Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature biotechnology* **29**, 1120-1127, doi:10.1038/nbt.2038 (2011).

26 Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell* **18**, 675-685, doi:10.1016/j.devcel.2010.02.012 (2010).

27 Devonshire, A. S., Elaswarapu, R. & Foy, C. A. Applicability of RNA standards for evaluating RT-qPCR assays and platforms. *BMC genomics* **12**, 118, doi:10.1186/1471-2164-12-118 (2011).

28 Sampieri, K. & Fodde, R. Cancer stem cells and metastasis. *Seminars in cancer biology* **22**, 187-193 (2012).

29 Shiozawa, Y., Nie, B., Pienta, K. J., Morgan, T. M. & Taichman, R. S. Cancer stem cells and their role in metastasis. *Pharmacology & therapeutics* **138**, 285-293, doi:10.1016/j.pharmthera.2013.01.014 (2013).

30 Cheung, K. J., Gabrielson, E., Werb, Z. & Ewald, A. J. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **155**, 1639-1651, doi:10.1016/j.cell.2013.11.029 (2013).

## Acknowledgements

Annie Le, Richard Daneman, Chih-Yang Wang, John Stingl, Mikhail Binneweis, Owen Witte. This study was supported by funds from the National Cancer Institute (CA180039), Stand Up To Cancer/ AACR (DT0409) and the US Department of Defense Congressionally Directed Medical Research Program (CDMRP) postdoctoral fellowship (11-1-0742).