bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Dissecting the contributions of tumor heterogeneity on metastasis at single-cell resolution

Juliane Winkler[1,2,✉], **Weilun Tan**[3], **Catherine M. M. Diadhiou**[1], **Christopher S. McGinnis**[4], **Aamna Abbasi**[1], **Saad Hasnain**[1], **Sophia Durney**[2], **Elena Atamaniuc**[2], **Daphne Superville**[2], **Leena Awni**[2], **Joyce V. Lee**[2], **Johanna H. Hinrichs**[1,5], **Marco Y. Hein**[3], **Michael Borja**[3], **Angela Detweiler**[3], **Su-Yang Liu**[1], **Ankitha Nanjaraj**[1], **Vaishnavi Sitarama**[1], **Hope S. Rugo**[6], **Norma Neff**[3], **Zev J. Gartner**[4,7], **Angela Oliveira Pisco**[3,✉], **Andrei Goga**[2,✉], **Spyros Darmanis**[3,8,✉], and **Zena Werb**[1,†]

[1]Department of Anatomy, University of California, San Francisco, San Francisco CA 94143, USA
[2]Department of Cell and Tissue Biology, University of California, San Francisco, San Francisco CA 94143, USA
[3]Chan Zuckerberg Biohub, San Francisco CA 94143, USA
[4]Department of Pharmaceutical Chemistry, University of California, San Francisco San Francisco CA 94143, USA
[5]Institute of Internal Medicine D, Medical Cell Biology, University Hospital Münster, Münster, Germany
[6]Department of Medicine, University of California, San Francisco, San Francisco CA 94143, USA
[7]Chan Zuckerberg Biohub Investigator, San Francisco CA 94143, USA
[8]Genentech Inc., South San Francisco, CA 94080, USA
[†]deceased 6/2020

Metastasis is the leading cause of cancer-related deaths, but metastasis research is challenged by limited access to patient material and a lack of experimental models that appropriately recapitulate tumor heterogeneity. Here, we analyzed single-cell transcriptomes of matched primary tumor and metastasis from patient-derived xenograft models of breast cancer, demonstrating that primary tumor and metastatic cells show profound transcriptional differences across heterogeneous tumors. While primary tumor cells upregulated several metabolic genes, metastatic cells displayed a motility phenotype in micrometastatic lesions and increased stress response signaling during metastatic progression. Additionally, we identified gene signatures that are associated with the metastatic potential and correlated with patient outcomes. Poorly metastatic primary tumors showed increased immune-regulatory control that may prevent metastasis, whereas highly metastatic primary tumors upregulated markers of epithelial-mesenchymal transition (EMT). We found that intra-tumor heterogeneity is dominated by epithelial-mesenchymal plasticity (EMP) which presented as a dynamic continuum with intermediate cell states that were characterized by novel, specific markers. These intermediate EMP markers correlated with worse patient outcomes and could serve as potential new therapeutic targets to block metastatic development.

Correspondence: *juliane.winkler@ucsf.edu, angela.pisco@czbiohub.org, andrei.goga@ucsf.edu, darmanis.spyridon@gene.com*

Current cancer treatment is most effective in attacking the primary tumor but has little effect on metastatic cells. This is a substantial problem because metastases account for the vast majority of cancer-related deaths (1). During the multistep process of metastasis, tumor cells adapt to various microenvironments that are distinct from their site of origin, but our understanding of the processes that lead to these adaptations is limited. Moreover, phenotypic alterations of metastatic cells may also cause resistance to therapeutics that cannot be accounted for by just genotypic changes (2).

The reason why some cancers metastasize while others do not is poorly understood. For example, specific genetic alterations are not necessarily required for metastatic progression (3), highlighting the importance of phenotypic adaptations of individual tumor cells to microenvironmental influences. In order to metastasize, tumor cells have to acquire complex traits; some of these include the ability to invade, intravasate and survive in circulation until they reach the metastatic site, where tumor cells extravasate into a new tissue and give rise to a secondary tumor. One concept aiming to explain these complex phenotypic changes is that tumor cells undergo epithelial-to-mesenchymal transition (EMT) and gain mesenchymal features. Thus, EMT has been suggested to play a fundamental role for tumor cells to disseminate to distant organs (4). However, to form overt metastasis, these disseminated tumor cells need to revert the EMT process, undergo mesenchymal-to-epithelial transition (MET), and gain epithelial features again. Epithelial-mesenchymal plasticity (EMP) therefore describes the ability of tumor cells to dynamically switch between epithelial and mesenchymal cell states. EMT is often described by the loss or gain of a few canonical markers involved in cell adhesion and motility (e.g. VIM, EPCAM, CDH1, CDH2), the expression of which are regulated by a set of core transcription factors (e.g. SNAI1, SNAI2, TWIST1, ZEB1). However, these commonly used markers are context- and tissue-dependent and change dynamically during the EMT process (5–7) leading to controversies in the field that rely on these few markers (8–14). Moreover, tumor tissues are heterogeneous, displaying various phenotypes and cell states within one tumor and thus require the analysis of individual cells within one tumor. To better understand the contributions of EMP to the metastatic process we need to comprehensively analyze individual heterogeneous tumor cells both at the primary tumor and metastatic site.

Advances in single-cell transcriptomics have enabled investigation into intra-tumor heterogeneity in breast cancer (BC) (15–18) and many other cancers (19–21). For

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

instance, an integration of these studies across multiple different tumor entities has highlighted both the importance and the context-dependency of the EMT process in tumor biology (22). However, with a few notable exceptions (23, 24), these studies focus on characterizing primary tumors. Moreover, they lack information about patient outcomes and metastatic phenotypes due to the necessary long-term follow-up. Comparing tumor heterogeneity at single-cell resolution between matched primary and metastatic tumors is logistically difficult; patient metastatic tumor samples are often collected years after the primary tumor was resected. Moreover, analyzing metastatic lesions is also technically difficult because they may consist of individual or small numbers of metastatic cells within complex tissues, which are hard to locate and isolate from patients. It is particularly challenging to investigate EMP *in vivo*, in both primary tumors and metastatic lesions, using an unperturbed system that resamples heterogeneous human tumor tissue. Finally, while xenograft models of metastasis can alleviate many of these limitations, such models that rely on transplanted cell lines do not faithfully reproduce the heterogeneity present in primary tumors. Thus, a detailed understanding of the involvement of the dynamic and context-dependent EMP process in metastasis is lacking (14).

Here, we characterized the metastatic potential of a large panel of patient-derived xenograft (PDX) models of human BC that spontaneously metastasize and preserve the heterogeneity of the primary human tumor. We analyzed the transcriptional profiles of individual primary tumor and matched metastatic cells using different single-cell RNA sequencing (scRNA-seq) approaches. Our study provides a rich dataset that allows us to investigate the impact of tumor heterogeneity on metastatic phenotypes both at the primary tumor and metastatic site at single-cell resolution. We identified gene signatures that are associated with metastatic potential. Specifically, we found that highly metastatic tumors express elevated EMT markers and demonstrate that EMP is a key factor of intra-tumor heterogeneity both at the primary tumor and the metastatic site. Within the continuum of EMP, we identified intermediate EMP cell states that are characterized by specific marker genes. High expression of those EMP marker genes was correlated with worse outcomes in a subset of BC patients.

BC PDX models with varying metastatic potential show transcriptional heterogeneity

To investigate intrinsic factors that impact a tumor's ability to metastasize, we analyzed the transcriptional heterogeneity of primary tumors as well as matched metastases at single-cell resolution using PDX BC models (Figure 1A). Human breast tumors were orthotopically transplanted into the cleared mouse mammary fat pad and spontaneously metastasized to the lung and other organs. They thereby preserve the heterogeneity of the primary human tumor, fully recapitulate the metastatic cascade, and mimic the metastatic pattern of the patient (25–27). We characterized PDX models derived from 13 BC patients, belonging to different BC subtypes (three luminal B, ten basal) with varying metastatic potential. Our PDXs included two estrogen receptor (ER) and progesterone receptor (PR) positive, one triple-positive (ER, PR, HER2), and ten triple-negative BC (TNBC) models (25, 28); three of the basal TNBC PDX models were newly established in this study (Supplementary Table 1).

First, we characterized the metastatic phenotype of the different tumor models once primary tumors reached a size of 2.5 cm in diameter. Based on the number and size of metastatic foci in the lungs of recipient mice the tumor models were grouped into those with low (n=6 models), moderate (n=3), and high metastatic potential (n=4). PDX models with low metastatic potential form no or very few micrometastases (< 10 cells), moderate models show more micro- and intermediate-sized (10 - 100 cells) metastases, and highly metastatic models develop either a high number of micrometastases and/or many macrometastases (> 100 cells) resulting in a substantial metastatic burden (proportion of metastatic cells in the lung) (Figure 1B, C, Supplementary Figure S1A, B). The metastatic potential based on this classification was independent of the primary tumor growth rate (Supplementary Figure S1C). For example, the fast-growing but poorly metastatic HCI002 model developed very few but larger metastatic foci even after primary tumor resection with subsequent tumor recurrence (Supplementary Figure S1E, F). Tumor resection allowed HCI002 to grow for a similar period as the slower growing but highly metastatic HCI010 model (Supplementary Figure S1D), indicating that HCI002's low metastatic potential is independent of primary tumor growth rate.

To investigate the transcriptional landscape of primary tumor and metastatic cells, individual tumor cells were isolated from primary tumors and matched metastatic lungs from 12 PDX models for scRNA-seq. Tumor cells stained with a human-specific antibody directed against a ubiquitous cell surface marker (CD298) (29) were isolated by fluorescent activated cell sorting (FACS) and subjected to scRNA-seq (Smart-Seq2). High-quality single-cell transcriptome data were collected for 2,090 cells (1,395 primary tumor and 695 metastatic cells). Of note, we were not able to isolate a sufficient number of metastatic cells from the poorly metastatic HCI002 model. The PAM50 BC subtype (Supplementary Table 1) and receptor status was confirmed for most samples (Figure 1D) according to ESR1 (ER), PGR (PR), and ERBB2 (HER2) transcript detection. Interestingly, ERBB2 was detected in all tumors including those not clinically classified as HER2-positive. This was potentially due to the required threshold for the clinical classification of the original tumor by histochemistry and/or single region sampling of the heterogeneous original tumor. In addition, receptor expression was maintained in metastatic cells in our data (Figure 1D and Supplementary Figure S1G). These results are in contrast to studies that reported a change of receptor status during tumor progression and recurrence occurring in up to 40 % of patients depending on the specific receptor, which had implications for treatment options and poor patient outcomes (30–32). However, our data indicate

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

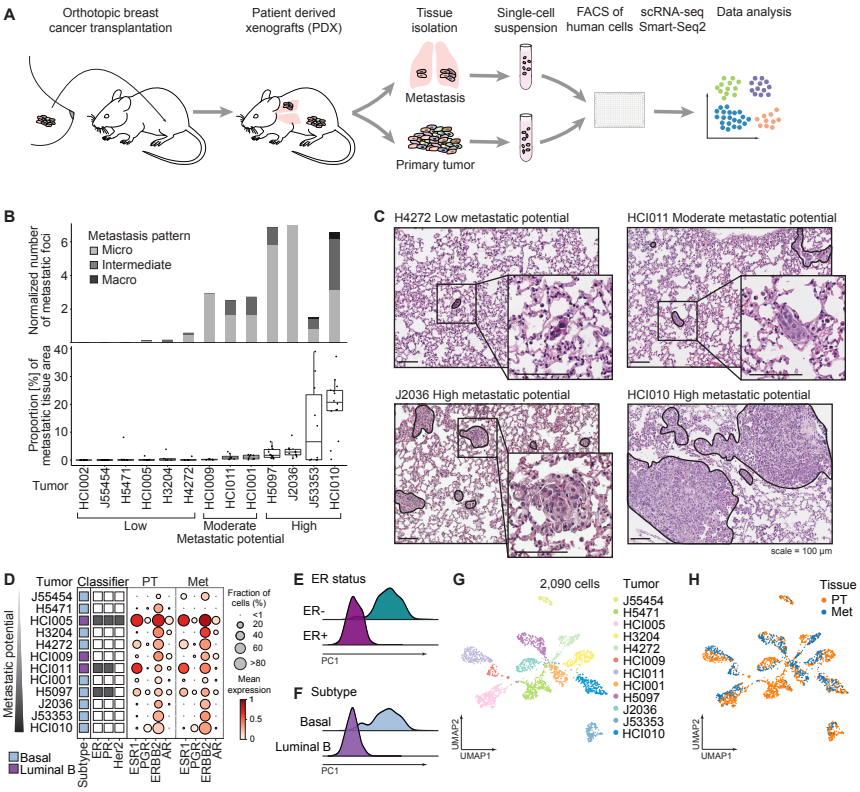


Figure 1. BC PDX models with varying metastatic potential show transcriptional heterogeneity. (**A**) Schematic overview of the experimental setup. Metastatic lung and primary tumor tissue were isolated from BC PDX models and dissociated. The resulting single-cell suspensions were enriched for human cells, sorted into 384 well plates and scRNA-Seq was performed using Smart-Seq2. (**B**) Bar chart shows the median number of metastatic foci per mm^2 lung tissue area per tumor model (upper panel) determined by histology. The size of metastatic foci is colored in shades of gray (micrometastasis: < 10 cells, intermediate: 10–100 cells and macrometastasis: > 100 cells). Boxplot shows the fraction of metastatic tissue per total lung tissue area determined by histology. Annotations indicate the metastatic potential of the tumor models. (**C**) Representative H&E images of metastatic lung tissues of tumor models for low, moderate and high metastatic potential. Scale = 100 µm. (**D**) Bubble plot shows the expression of receptors in primary tumor (PT) and metastatic cells (Met) per tumor model. The size of dots indicates the fraction of expressing cells and the red color indicates the magnitude of gene expression. Box annotations show BC subtype and receptor classification. Tumor models are ordered by increasing metastatic potential as determined in (B). (**E**) Ridgeplot shows the normalized number of cells along Principal Component 1 (PC) coordinates color-coded by ER status. (**F**) Ridgeplot shows the normalized number of cells along PC1 coordinates color-coded by BC subtype. (**G**) UMAP projection of single-cell transcriptomes color-coded by individual tumor models. (**H**) UMAP projection of single-cell transcriptomes color-coded by primary tumor (PT, orange) and metastatic cells (Met, blue).

that changes in receptor status are not caused by the metastatic process but are likely a consequence of selection during receptor-targeted therapy.

The ER status and BC subtype were the major sources of variation in our dataset. This is illustrated by principal component (PC) analysis, which showed a clear separation of ER status and BC subtypes along PC1 (Figure 1E, F, Supplementary Figure S1H). Moreover, individual tumors clustered separately from other tumors, reflecting the effect of inter-patient heterogeneity on gene expression (Figure 1G). Notably, variability between technical batches (individual plates) or biological replicates (same tumor implanted into different animals) was not observed (Supplementary Figure S1I, J). Finally, within individual tumor models, primary and metastatic cells clustered separately in all cases, highlighting the transcriptional differences between primary and metastatic cells from a particular tumor model (Figure 1H).

Taken together, we established and characterized PDX models of different BC subtypes with varying metastatic potentials that were independent of their primary tumor growth rate. Receptor status was maintained between primary tumor and metastatic cells. In addition to inter-patient heterogeneity, primary tumor and metastatic cells showed strong transcriptional differences within individual tumors.

Differential gene expression analysis reveals metastasis-associated gene signatures and heterogeneity between cells

To characterize general transcriptional programs unique to metastatic cells, cells were grouped across all samples by tissue source (primary tumor or metastatic lung) and differential gene expression was determined using MAST (33) with the tumor model as a covariate. We found 132 differentially expressed genes (DEGs), 79 of which were upregulated in metastatic cells conserved across all 12 tumor models ($\log_2$ fold change > 0.5; Figure 2A, Supplementary Figure S2A, Supplementary Table 2). Among the top metastasis-associated genes were several cytokeratins (KRT5, KRT6B, KRT14, KRT17, KRT81), calcium-binding S100 proteins (S100A16, S100A14), heat shock protein HSP1, cell-surface proteins such as TSPAN1, serine proteases (KLK6, KLK7), and the glycoproteins CEACAM6 and PSCA. Pathway-level analysis revealed that metastatic cells were enriched in MYC, E2F, PI3K/AKT/MTOR signaling and oxidative phosphorylation (Figure 2B). This observation is consistent with studies showing enrichment of MYC signaling and oxidative phosphorylation pathways in metastatic BC cells found in the lung (29, 34). Interestingly, metastatic cells additionally upregulated genes involved in immune response pathways (IL6/JAK/STAT3), presumably as an adaptation to the metastatic microenvironment (Figure 2B, Supplementary Figure S2B). In contrast, hypoxia, EMT, angiogenesis, and glycolysis were pathways enriched in primary tumor cells (Figure 2B). To determine whether the identified pathways were upregulated in all individual primary tumor

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

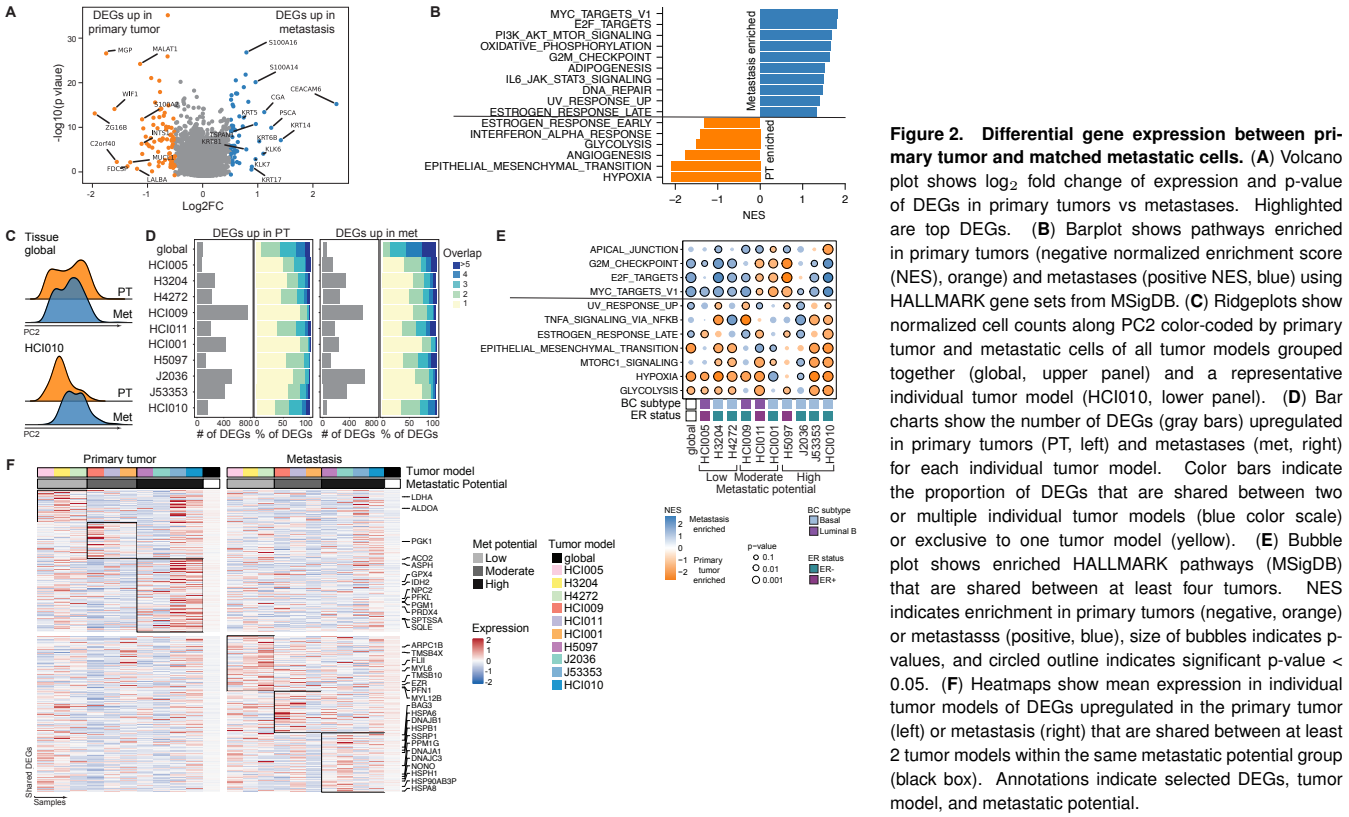


Figure 2. Differential gene expression between primary tumor and matched metastatic cells. (**A**) Volcano plot shows $\log_2$ fold change of expression and p-value of DEGs in primary tumors vs metastases. Highlighted are top DEGs. (**B**) Barplot shows pathways enriched in primary tumors (negative normalized enrichment score (NES), orange) and metastases (positive NES, blue) using HALLMARK gene sets from MSigDB. (**C**) Ridgeplots show normalized cell counts along PC2 color-coded by primary tumor and metastatic cells of all tumor models grouped together (global, upper panel) and a representative individual tumor model (HCI010, lower panel). (**D**) Bar charts show the number of DEGs (gray bars) upregulated in primary tumors (PT, left) and metastases (met, right) for each individual tumor model. Color bars indicate the proportion of DEGs that are shared between two or multiple individual tumor models (blue color scale) or exclusive to one tumor model (yellow). (**E**) Bubble plot shows enriched HALLMARK pathways (MSigDB) that are shared between at least four tumors. NES indicates enrichment in primary tumors (negative, orange) or metastasss (positive, blue), size of bubbles indicates p-values, and circled outline indicates significant p-value < 0.05. (**F**) Heatmaps show mean expression in individual tumor models of DEGs upregulated in the primary tumor (left) or metastasis (right) that are shared between at least 2 tumor models within the same metastatic potential group (black box). Annotations indicate selected DEGs, tumor model, and metastatic potential.

or metastatic cells or only in a subset of cells we examined the expression of DEGs associated with the top enriched pathways in either primary tumor (hypoxia) or metastatic cells (MYC). The analysis revealed a profound heterogeneity both between and within tumor models (Supplementary Figure S2C). However, when analyzed individually, primary tumor and metastatic cells displayed strong transcriptional differences illustrated by separation along PC2 (Figure 2C; Supplementary Figure S2D–G) mirroring our previous observations (Figure 1H).

To control for this pronounced variability amongst our tumor models, we next analyzed DEGs between primary tumor and metastatic cells for each model separately (Supplementary Table 2) and compared these across tumor models. Due to insufficient metastatic cell numbers, two tumor models with low metastatic potential (J55454, H5471) were excluded from this analysis (Supplementary Figure S2E, F). The different tumor models showed a wide range of numbers of DEGs (Figure 2D). Notably, more than 50 % of DEGs were tumor model-specific and only a few (< 5 %) were shared between more than 5 tumor models, highlighting again the magnitude of inter-patient heterogeneity (Figure 2D). We focused on enriched pathways that were shared between tumor models (Figure 2E). Although most shared pathways were also identified in the previous analysis across tumor models (Figure 2B, Supplementary Figure S2B), some pathways showed intriguing enrichment differences between tumor models. For example, whereas the combined analysis revealed an overall suppression of the estrogen-response pathway in the primary tumor, the individual analyses showed that this pathway was specifically upregulated in ER+ primary tumors (HCI005, HCI011, H5097) compared to matched metastatic samples. This suggests that estrogen signaling is impaired in the metastatic cells despite maintained ESR1 expression (Figure 2E, Figure 1D, Supplementary Figure S1G). Additionally, while this analysis showed that metastatic cells of some tumor models were enriched in the G2M checkpoint pathway, we could not confirm an overall more active proliferation or substantial cell cycle shifts of metastatic cells in our data (Supplementary Figure S2H, I). Owing to their larger size, primary tumors have limited access to nutrients; thus, it is not surprising that enrichment of glycolysis and hypoxia seemed to be a general feature in primary tumors. Moreover, EMT was enriched either in primary tumors or metastasis in the majority of the analyzed tumor models indicating a dynamic activity of this pathway in both compartments.

Since individual DEGs were shared only between a few tumor models, we focused on DEGs that were common between tumor models with a similar metastatic phenotype (Figure 2F, Supplementary Table 3). We found 74 upregulated genes in metastatic cells that were shared between at least two tumors of low metastatic potential. Among these were many genes involved in cytoskeleton assembly and cell motility (e.g. MYL12B, MYL6, PFN1, TMSB4X, TMSB10, ARPC1B, EZR, FLII). In contrast, among the 91 genes upregulated in metastatic cells from high metastatic tumor models were many genes indicative of high stress-response signaling, including several heat shock proteins (HSPB1, HSPA8, HSPA6, HSPH1, HSP90AB3P, DnaJs A1,

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B1, C3, and BAG3), PPM1G and genes involved in DNA damage repair (SSRP1, NONO). Several genes involved in glycolysis (ALDOA, LDHA, PGK1, PFKL, PGM1) and other metabolic processes (GPX4, PRDX4, ACO2, ASPH, IDH2, SQLE, NPC2, SPTSSA) were upregulated in primary tumor cells suggesting differential metabolism in primary tumors as compared to the metastasis.

In summary, we observed strong transcriptional differences between primary tumor and metastatic cells on an individual tumor level with the majority of DEGs being specific to each tumor. Shared features across models are upregulation of hypoxia, glycolysis and other metabolic-related genes in primary tumor cells. Shared upregulated genes among metastatic cells are involved in cytoskeleton assembly and motility and stress response signaling.

Metastatic signatures are correlated with patient outcomes

The tumor models used in this study exhibit consistent metastatic behaviors and were classified into tumors with low, moderate, and high metastatic potential (Figure 1B, C, Supplementary Figure S1A). One fundamental question is whether intrinsic features of the primary tumor are predictive of the observed different metastatic potential of those tumor models. To address this question we generated an additional, larger scRNA-Seq dataset, which better reflected the intra-tumor heterogeneity of the primary tumors. To this end, we performed high-throughput, droplet-based scRNA-Seq with MULTI-Seq (36) sample multiplexing on 10 different primary tumors with varying metastatic potential (Figure 3A), resulting in 16,861 tumor cells (Supplementary Figure S3A–C).

To identify signatures that were associated with the metastatic potential of the primary tumor, we looked for DEGs between different metastatic potential groups (Figure 3B, Supplementary Table 4, Supplementary Table 5). For each metastatic potential group, we selected genes that were shared between both scRNA-Seq methods that were used in this study (Figure 3C, Supplementary Table 6). Among the shared genes upregulated in primary tumors with a low metastatic potential were genes related to immune regulation processes such as antigen processing and cross-presentation (e.g. HLA-A, HLA-B, HLA-C, HLA-E, B2M, TAP1), and innate immunity (e.g. NFKBIA, PSMB3, SQSTM1, LAMP2, IFI6, IFI35) (Figure 3D, Supplementary Figure S3D). As our model used immunocompromised mice that lack B, T and NK cells, these findings potentially reflect a tumor-intrinsic, anti-metastatic feature independent of the canonical function of these genes in immune regulation. Genes upregulated in highly metastatic primary tumors included known metastasis-related genes such as S100A4 (37–39), MUC1 (40) and genes associated with EMT (VIM, PLOD1, BGN), including the common EMT marker vimentin (VIM) (Figure 3D). MYC signaling was among the top 5 enriched pathways in highly metastatic primary tumors (Supplementary Figure S3E). MYC signaling can lead to evasion from immune surveillance by the suppression of interferon signaling and antigen-presentation pathways including the down-regulation of B2M and MHC-I (41, 42). This anti-correlation could explain the observed upregulation of immune regulatory pathways in poorly metastatic compared to highly metastatic primary tumors that showed elevated MYC signaling (Supplementary Figure S3F). Supporting our experimental data (Supplementary Figure S1C) a highly metastatic phenotype is not the result of more proliferation since proliferation rate or cell cycle phase distributions were not significantly changed between primary tumors of different metastatic potentials (Supplementary Figure S3G, H).

Next, we tested whether the observed metastasis-associated signatures were correlated with patient-related outcomes using publicly available bulk gene expression data of BC patients across different subtypes (35) (Figure 3E). Indeed, patients with a high expression of the poorly metastatic signature exhibit improved distant metastasis-free survival (DMFS). A high expression of moderate metastatic genes was associated with worse recurrence-free survival (RFS) and a high expression of the highly metastatic signature showed the worst outcome for patients.

In summary, we identified intrinsic metastasis-associated gene signatures in primary tumors that were correlated with patient-related outcomes of an external dataset. While genes upregulated in poorly metastatic primary tumors are involved in immune regulation presenting potential non-canonical anti-metastatic functions, genes present in the highly metastatic signature were associated with EMT.

Epithelial-mesenchymal plasticity is a key feature of tumor heterogeneity and is associated with metastatic potential

Markers of EMT were upregulated in primary tumors of highly metastatic tumor models as compared to models with low metastatic potential. However, we also found EMT to be enriched in either primary tumor or metastatic cells in different tumor models indicating a dynamic process during metastatic progression. Tumor cells must switch phenotypes multiple times during the metastatic process to adapt to different environments. While numerous studies have established a role for the epithelial-mesenchymal transition (EMT) in cancer progression, fewer have examined the role of epithelial-mesenchymal plasticity (EMP) in this process (14). Studying the latter is challenging, as most studies focus on a limited set of end-point markers to distinguish epithelial and mesenchymal cell states and/or to perturb these markers to test the role of EMT either *in vitro* or in mouse models *in vivo*. Here, using single-cell transcriptomics, we seek to identify those cell states across the spectrum of EMT and in multiple heterogeneous human tumor populations that correlate with their metastatic potential *in vivo*. To this end, we used a pan-cancer gene signature of 303 mesenchymal and epithelial markers to characterize the EMP state of individual cells (43). Individual canonical epithelial markers (EPCAM and CDH1) were highly expressed in cells with a high epithelial signature and also mesenchymal

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

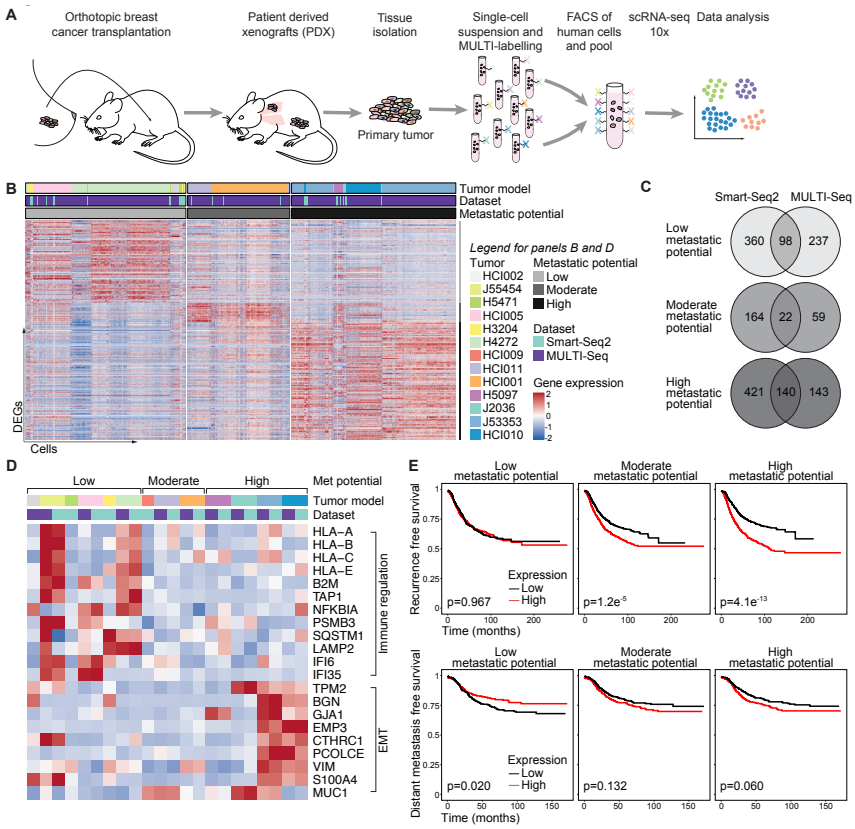


Figure 3. Metastatic signatures are correlated with patient outcomes. (**A**) Schematic workflow about the experimental setup using MULTI-Seq. (**B**) Heatmaps show DEGs between individual tumors and tumors of the other metastatic potential groups that are shared between at least 2 tumors. Annotations show the tumor model, dataset and metastatic potential group. (**C**) Venn diagram shows the number of DEGs shared between the Smart-Seq2 and MULTI-Seq datasets for the different metastatic potential groups. (**D**) Heatmaps show the mean expression of selected metastasis-associated genes. Legends for annotations are the same as in Figure 4B. (**E**) Recurrence-free survival (RFS, top, n = 2,032 patients) and distant metastasis-free survival (DMFS, bottom, n = 958 patients) of BC patients using the mean expression of the metastasis-associated gene signatures (generated with KM-plotter (35)).

markers (VIM, FN1, CDH2) showed the expected expression patterns (Supplementary Figure S4A). However, some of these commonly used markers, such as FN1 and CDH2, were minimally detected on an individual cell level, indicating the importance of using multi-gene signatures to define cell states. To illustrate that cells can express epithelial and mesenchymal markers dynamically, we combined epithelial and mesenchymal signatures to define the overall EMP cell state; e.g. an EMP signature >0 reflects cells with a higher mesenchymal signature than epithelial signature. These EMP signatures of individual tumor models were strongly correlated (R^2 = 0.78) between our two datasets (Smart-Seq2 and MULTI-Seq), demonstrating reproducibility of results across different sequencing methods and experiments (Supplementary Figure S4B). Tumors that consistently metastasized expressed a significantly higher ($p<0.001$) EMP signature compared to those that poorly metastasized (Figure 4A, Supplementary Figure S4C). For individual tumor models, a high EMP state was associated with metastatic potential (Figure 4B, Smart-Seq2 R = 0.336, Supplementary Figure S4D, MULTI-Seq R = 0.606). Surprisingly, the overall EMP state of each tumor model was similar for both primary tumor and metastatic cells (Figure 4C) suggesting an intrinsic determinant of EMP that is potentially independent of environmental influences which were likely very different between the tissues. However, across individual cells, the EMP state was highly variable within one tumor model. Indeed, EMP signatures of individual tumor models were strongly correlated with PC1 coordinates, indicating that the EMP cell state is a major source of variation between cells within one tumor model and significantly contributes to intra-tumor heterogeneity (Supplementary Figure S4E, F). Finally, we observed that EMP state was gradually changing in transcriptional space, further illustrating that EMP is a continuum of cell states (Figure 4 D).

Next, we asked whether the metastatic potential is associated with the EMP state of individual cells or their proportion within the tumor. Cells were classified by the magnitude of their EMP signature expression into three different cell states: epithelial-like (epithelial > mesenchymal signature), intermediate EMP (epithelial = mesenchymal signature) and mesenchymal-like cells (epithelial < mesenchymal signature) (Figure 4E, Supplementary Figure S4G). The proportion of mesenchymal-like cells largely aligned with the metastatic potential, with no classified mesenchymal-like cells present in poorly metastatic tumors and almost no epithelial-like cells present in tumors with high metastatic potential (Figure 4F, Supplementary Figure S4H). ER+ and luminal B tumors showed the highest proportion of epithelial-like cells. However, even within this group, the proportion of mesenchymal-like cells was associated with increased metastatic potential. Similar associations were observed for the group of TNBC basal tumors which showed an overall higher fraction of mesenchymal-like cells.

EMP is a continuum of cell states with intermediate EMP cells expressing distinct marker genes

Studies suggest that both mesenchymal and epithelial functions are necessary for the metastatic cascade (44). Therefore, the intermediate EMP cells were of special

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

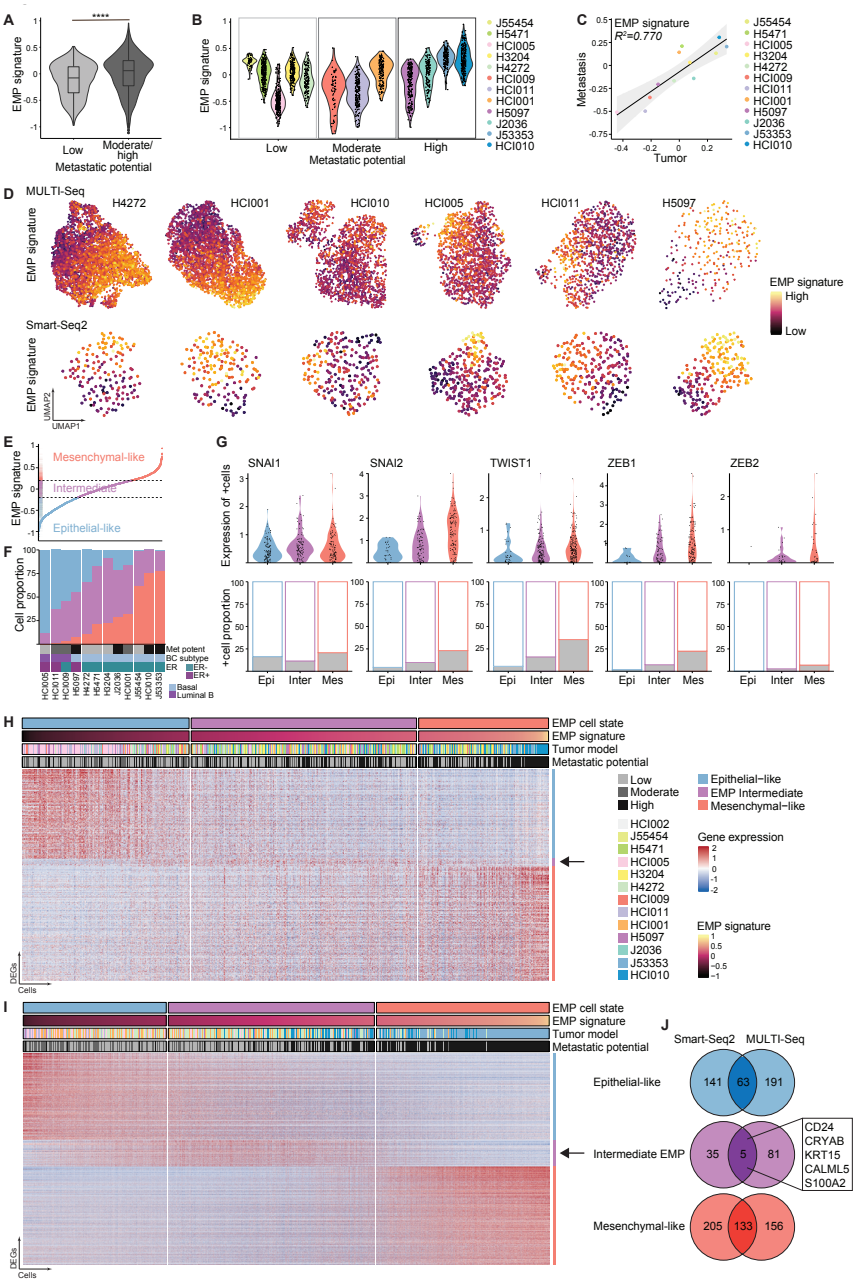


Figure 4. EMP is a key feature of tumor heterogeneity. (**A**) Violin plot shows EMP signature expression of tumor models with low and intermediate/high metastatic potential using the Smart-Seq2 dataset. Boxplot showing median, significance p<0.001 by Wilcox test. (**B**) Violin plot shows EMP signature per tumor model ordered by metastatic potential using the Smart-Seq2 dataset. (**C**) Scatter plot shows the correlation of the mean EMP signature of the primary tumor and metastatic cells colored by the tumor model. Linear regression with 95 % confidence intervals and Pearson correlation coefficient are shown. (**D**) UMAP projections of single-cell transcriptomes for individual tumor models. The color scale indicates the magnitude of EMP signature expression. (**E**) Cells ranked by EMP signature define three cell states: epithelial-like (blue), intermediate EMP (purple) and mesenchymal-like cells (red) using the Smart-Seq2 dataset. (**F**) Bar chart shows the proportion of the three different EMP cell states in each tumor model ranked by the increasing proportion of mesenchymal-like cells. Gray-scale boxes indicate the metastatic potential. Other annotations indicate ER status and BC subtype. Showing the Smart-Seq2 dataset. (**G**) Violin plots (top) show expression of EMT-associated TFs in expressing cells grouped by EMP cell states (Epi = epithelial-like, Inter = Intermediate EMP, Mes = mesenchymal-like cells). Bar charts (bottom) show the fraction of expressing cells colored in gray. Showing the Smart-Seq2 dataset. (**H**) Heatmap shows DEGs for epithelial-like, mesenchymal-like, and intermediate EMP cells for the Smart-Seq2 data. Cells are ordered by increasing EMP signature. Annotations indicate EMP cell state, EMP signature expression, tumor model and metastatic potential. The arrow highlights intermediate EMP cell marker genes. (**I**) Same as in (H) using the MULTI-Seq data. (**J**) Venn diagrams show overlap DEGs of epithelial-like, mesenchymal-like, and intermediate EMP cells between Smart-Seq2 and MULTI-Seq data. Highlighted are overlapped markers for intermediate EMP cells.

interest as they may represent cells with both epithelial and mesenchymal capabilities and a high degree of plasticity and therefore might contribute to the pool of cells that are more likely to metastasize (45, 46). However, the identified intermediate EMP cells, which expressed both epithelial and mesenchymal signatures at similar levels (Figure 4E), were present in every tumor although their abundance did not correlate with metastatic potential (Figure 4F). Intermediate EMP cells expressed core transcription factors (TFs) promoting EMT such as SNAI2, TWIST1, ZEB1 and ZEB2 (13) (upper panels of Figure 4G and Supplementary Figure S4I) at higher levels than epithelial-like cells but lower than mesenchymal-like cells highlighting their intermediate character. Moreover, the fraction of cells expressing these TFs also increased from epithelial-like to intermediate EMP to mesenchymal-like cells (lower panels of Figure 4G and Supplementary Figure S4I). To further characterize this intermediate EMP cell state we performed a more comprehensive differential gene expression analysis between the three EMP cell states in both datasets (Smart-Seq2 and MULTI-Seq) and identified genes upregulated in epithelial-like, intermediate EMP and mesenchymal-like cells (Figure 4H, I, Supplementary Table 7, Supplementary Table 8). For each EMP cell state, we focused on marker genes that were shared between the two datasets (Figure 4J). Surprisingly, only 13 % (40/303, MULTI-Seq) – 18 % (56/303, Smart-Seq2) of DEGs were shared with the published markers (43) that were used to classify the three EMP states. Most identified DEGs were exclusive to one or both of our datasets and were not found in the set

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of published markers (Supplementary Figure S4J). Genes shared across all three sets included common mesenchymal (e.g. VIM, BGN, SNAI2, LOX) and epithelial (e.g. KRT18, KRT8) markers; whereas other broadly used markers such as EPCAM, CDH1, CDH2 and FN1 were not included. Only 5 intermediate EMP cell marker genes were shared between our datasets (Figure 4J). The expression of all 5 intermediate EMP markers (CRYAB, KRT15, S100A2, CD24, CALML5) peaked in intermediate EMP cells and decreased in epithelial-like and mesenchymal-like cells (Figure 5A, Supplementary Figure S5A).

The five markers of the intermediate EMP cell state have been previously implicated in EMT, cancer stemness, and metastasis pathways. For example, we identified the cell surface protein CD24, for which there are conflicting results as to its role in tumor progression. For example, whereas $CD24^{-/low}/CD44^{+}$ have been shown to initiate breast tumors in NOD/scid mice (47), other studies reported that high CD24 expression increased metastasis (48) and that $CD24^{+}/CD90^{+}$ cells initiate metastases that display a mesenchymal phenotype (49, 50). $CD24^{+}/CD44^{+}$ cells were shown to be plastic and express epithelial and mesenchymal markers forming mammospheres more efficiently than epithelial-like $CD24^{+}/CD44^{-}$ or mesenchymal-like $CD24^{-}/CD44^{+}$ cells (51–54). Interestingly, a mixture of $CD24^{+}/CD44^{-}$ and $CD24^{-}/CD44^{+}$ was more efficient in mammosphere formation than either population alone (52). Collectively, these data suggest that CD24 could indeed mark a plastic intermediate EMP cell state with potential stem-like properties and that cooperativity may exist between cell populations with different EMP characteristics.

Another identified intermediate EMP marker, CRYAB, encodes the small heat shock protein α-basic-crystallin (αB-crystallin), which protects cells from apoptosis by inhibiting caspase-3 activation under various stress conditions such as oxidative stress (55, 56). Importantly, αB-crystallin confers anoikis resistance and thereby enables metastatic dissemination (57). CRYAB is overexpressed in various tumors (58, 59) including the basal BC subtype (60) and has been associated with poor patient outcomes and metastasis (61–64). CRYAB expression levels in hepatocellular carcinoma cell lines were accompanied by EMT marker expression indicating that CRYAB could promote a mesenchymal phenotype (59). Importantly, a study investigating brain metastasis found that CRYAB is expressed in a small non-proliferative metastatic cell population and might be required for the survival of single metastatic cells and micrometastasis (65). Additionally, CRYAB is highly expressed in dormant micrometastasis in the lung compared to proliferative macrometastasis (65, 66). These combined features could also indicate an intermediate EMP cell state, although not explored in these studies.

Additional intermediate EMP markers were KRT15, CALML5 and S100A2. KRT15, an intermediate filament protein belonging to the epithelial Keratin Type I family, has been suggested to be an epidermal stem cell marker (67). One study reported that high KRT15 expression correlated with better outcomes for BC patients (68); whereas others reported that KRT15 was upregulated in advanced stage BC and BC with high-relapse risk (69, 70), as well as being associated with poor prognosis in other types of cancers (71, 72). CALML5 (Calmodulin-Like Protein 5) is a calcium-binding protein that is predominantly expressed in keratinocytes. It has recently been shown that calmodulin may mediate the induction of a partial EMP cell state (measured by loss of E-cadherin surface expression and a partial upregulation of mesenchymal markers) through calcium signaling (73). Also involved in calcium signaling is S100A2, which belongs to the S100 protein family that can both bind calcium and function extracellularly. S100A2 is deregulated in cancers suggesting both tumor-promoting and suppressing roles (74–76) and may also have dual roles with regard to EMT. S100A2 was shown to be regulated by TGF-β1 (77) and partially mediate TGF-β1-induced EMT (78, 79) but seems to repress EMT in other contexts (80).

Recent studies describing the existence of intermediate EMP cells have associated these with an increased ability to form metastases after tail vein injection using genetically engineered mouse models of skin squamous cell carcinoma (45). Here, we identified 5 novel markers, CD24, CRYAB, KR1T15, S100A2, and CALML5, that are expressed by human intermediate EMP cells in BC *in vivo* (Figure 5A, Supplementary Figure S5A). These genes could serve as biomarkers to identify BC patients with an increased proportion of potentially more aggressive tumor cells. To test the clinical significance of our findings we analyzed two BC gene expression datasets. In the first dataset (35), patients across different BC subtypes that showed a high expression of the epithelial-like gene signature had a better RFS, whereas patients with a high expression of the intermediate EMP or mesenchymal-like gene signature showed worse RFS (Supplementary Figure S5B). In an independent dataset (METABRIC (81)), intermediate EMP cell gene signature showed a BC subtype-dependent correlation with patient-related outcome (Figure 5B). Whereas luminal tumors did not show a correlation, high expression of the intermediate EMP cell signature in patients with basal and Her2-like classified BC showed worse RFS. These subtypes also showed the worst outcomes and resistance to therapy compared to the other subtypes (82).

Taken together, we identified cells co-expressing epithelial and mesenchymal markers that belong to an intermediate EMP cell state. These intermediate EMP cells were present in primary tumors and metastases of all tumor models studied and were characterized by low expression of EMT-associated TFs. Specific marker genes could identify intermediate EMP cells and a high expression of these markers was associated with worse patient-related outcomes. These novel intermediate EMP cell marker genes could serve as targets to block the dynamic process of EMP by directly targeting the potentially most plastic cells and thereby interfering with the metastatic cascade.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

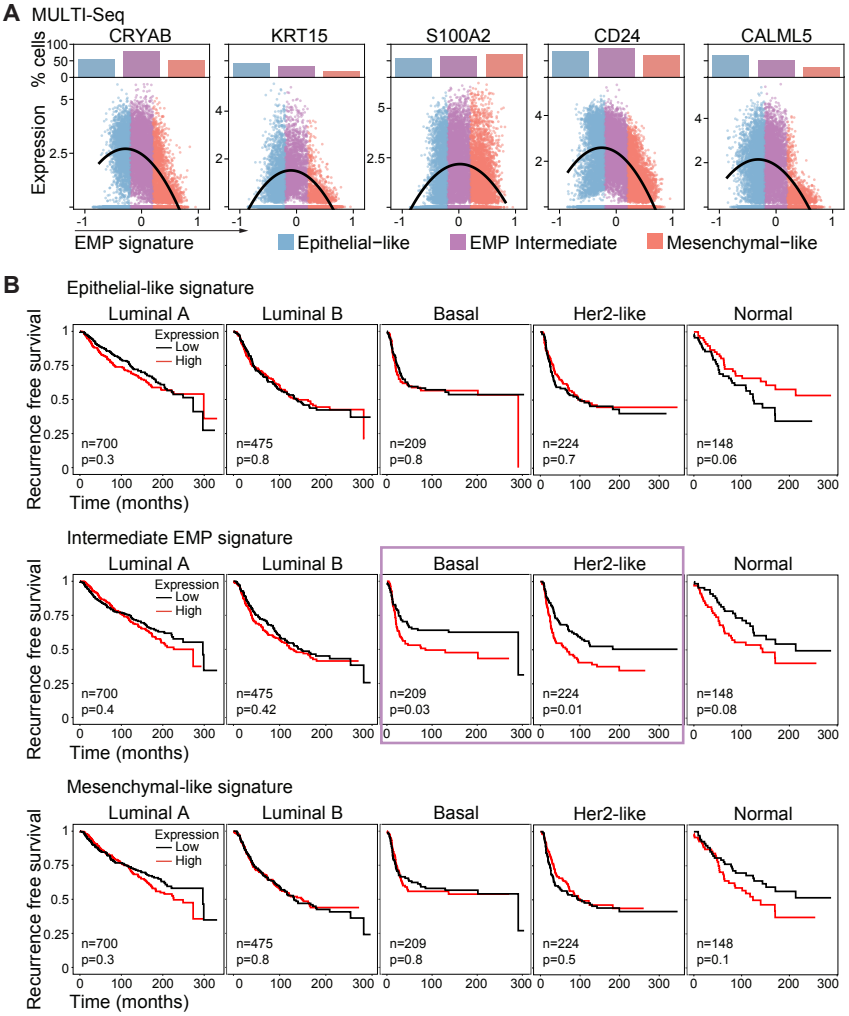


Figure 5. Intermediate EMP cell markers were correlated with patient outcome. (**A**) Scatter plots show the expression of indicated genes ordered by increasing EMP signature expression. Dots show the expression for individual cells and lines show smoothed expression of expressing cells. Bar charts on top show the proportion of positive expressing cells for the three EMP cell states (blue=epithelial-like, purple=intermediate EMP, red=mesenchymal-like cells). Showing the MULTI-Seq dataset. (**B**) Recurrence-free survival of BC patients (METABRIC) separated by PAM50 BC subtype using the mean expression of the epithelial-like (top panel), intermediate EMP (middle panel), and mesenchymal-like signatures (lower panel). The number of patients (n) and p-value (p) are shown. The purple box indicates a significant p-value.

Discussion

Metastasis is responsible for the majority of cancer-related deaths but the underlying processes that drive metastasis are not fully elucidated. Recent advances in single-cell biology shed light on the profound heterogeneity of tumors between and within patients that likely contributes to the complexity of the metastatic phenotype. By analyzing matched primary tumor and metastatic cells we found significant differences in the transcriptional profiles of metastatic cells compared to their primary tumor of origin in distinct BC subtypes that showed strong patient-to-patient variability. Specifically, we found that primary tumor cells consistently upregulated genes involved in hypoxia, glycolysis and other metabolic pathways across all tumor models. Moreover, metastatic cells frequently upregulated genes involved in cytoskeleton assembly, cell motility, cellular stress and immune response signaling. These transcriptional differences presumably are necessary to acquire traits for dissemination and are a result of the adaptations to different environments. A better understanding of this observed heterogeneity and the transcriptional differences between primary tumor and metastatic cells could have implications for therapy response.

One of these cellular traits is EMT, which has long been suggested as being an important driver of metastasis. Our current work highlights the complexity of the process (i.e. EMP) and its associated cell states. Recent research suggests that epithelial and mesenchymal cell states are the edges of a wider dynamic continuum of EMP including intermediate cell states (13, 83). However, these intermediate EMP cells remain poorly characterized. Here, we report that EMP is a dominant feature of tumor heterogeneity observed in different human BC tumors *in vivo*. We identified epithelial- and mesenchymal-like cells as well as intermediate EMP cells that surprisingly co-exist in every tumor. Intermediate EMP cells (described previously also as partial-EMT, hybrid-EMT, or EMT-transition cells) have recently been reported to exhibit the greatest metastatic potential when compared to mesenchymal or epithelial cells using tail vein injection of skin squamous carcinoma cells or orthotopic injection of highly metastatic pancreatic ductal adenocarcinoma cells, both derived from genetic mouse models (45, 46). On the contrary, we did not observe a correlation between the abundance of intermediate EMP cells in tumors and their metastatic potential. Indeed, we found that intermediate EMP cells were also present in very poorly metastatic tumor models. Instead, we found that a stronger mesenchymal

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenotype (high EMP signature) and a higher proportion of mesenchymal-like cells correlated with the metastatic potential and this correlation could be further influenced by the BC subtype. Overall, our data suggest that the propensity to metastasize is a function of the entire tumor cell population, as opposed to the presence or absence of a 'rogue' and a potentially small subset of cells. It will be important for future studies to investigate the interactions of a heterogeneous tumor cell population with non-malignant cells of the environment and their involvement in the metastatic process.

Notably, our findings build upon, but do not necessarily contradict, the recently described importance of intermediate EMP cells for metastasis. Potentially, a higher proportion of epithelial-like cells may prevent intermediate EMP cells from metastasizing. Conversely, a higher proportion of mesenchymal cells may support the metastatic capabilities of intermediate EMP cells. One example of this proposed cooperativity between different EMP cells states is the observation that an admix culture of (epithelial-like) $CD24^{+}/CD44^{-}$ and (mesenchymal-like) $CD24^{-}/CD44^{+}$ immortalized normal human mammary epithelial (HMLER) cells was more efficient in mammosphere formation than either population alone (52). Thus, one hypothesis is that the critical factor determining the metastatic potential of a tumor is a combination of its composition of cells with varying EMP states and the level of cooperativity between them. The observation that metastasis can have polyclonal origins (84) and that circulating tumor cell (CTC) clusters are more effective in metastasis formation than individual CTCs (85) supports our idea that cooperativity between different EMP cell states may result in more effective metastasis formation.

EMP is likely a highly context and tumor type-specific process involving different signaling (6). Although the presence or proportion of intermediate EMP cells was not correlated with more metastasis in our models, we did find that high expression of intermediate EMP marker genes was associated with poorer outcomes in a subset of BC patients whereas a mesenchymal or epithelial gene expression did not show a subtype-specific correlation. This observation not only highlights the potential importance of the intermediate EMP cell state for patient outcomes but also indicates that the EMP process and its involvement in metastatic disease might be subtype-specific. Other markers of intermediate EMP cell states have been identified and linked to tumorigenesis, metastasis and stemness (such as CD104, $EPCAM^{-}/CD106^{+}$, ALDH1) (45, 54, 86, 87). These and our study highlight the need for a deeper understanding of the involvement of the intermediate EMP cell state in metastasis and its potential spatio-temporal context specify (23).

Surprisingly, there seems to be a predetermined, intrinsic equilibrium of EMP cell states within the tumor that appears to be independent of extrinsic signals and microenvironmental adaptations. Thus, primary tumor and metastatic cells exhibit very similar levels of the EMP signature expression despite showing remarkable differences in their overall transcriptomes. Sustaining plasticity and reversing the mesenchymal into a more epithelial-like cell state (MET) is proposed to occur during the formation of overt metastasis (88, 89). In this context, we would expect to detect more mesenchymal-like cells isolated from micrometastases and more epithelial-like phenotypes in macrometastases. Based on our histological characterization, poorly metastatic models show primarily micrometastases but also a few intermediate sized foci and potentially even rare macrometastases. Since metastatic cells were isolated from whole lung tissue, we were unable to distinguish whether cells obtained from poorly metastatic models were associated with intermediate-sized foci or (very rare) macrometastases or compare the transcriptome of micro- and macrometastases. Nonetheless, our single-cell analysis of primary tumor and metastasis revealed that EMP represents a continuum during spontaneous metastasis of a large panel of patient-derived breast tumors. Recent technology developments in spatial transcriptomics and multiplexed antibody-based imaging will be perfectly suited for future studies to investigate the dynamics of the various EMP cell states as metastatic tumors form.

ACKNOWLEDGMENTS

We thank S. Schmid, H. Goodarzi, L. Murrow, Z. Gardell and H. Kortbai and all members of the Werb, Goga and Gartner labs for their valuable discussions and input. We thank M. Owyong, A. Abisoye-Ogunniyan, N. Ataii, and K. Salari for their technical assistance. We thank M.T. Lewis, L.E. Dobrolecki, and A. Welm for sharing their PDX models. We thank the UCSF flow core for their assistance, in particular A. Carlos, S. Kraus, and V. Nguyen. UCSF flow core is supported by RRID:SCR_018206 and DRC Center Grant NIH P30 DK063720. We thank E. Chow for sequencing assistance. We thank the UCSF Cancer Center Tissue Core. This study was supported by funds from EMBO long-term post-doctoral fellowship (EMBO ALTF 159-2017 to JW), Program for Breakthrough Biomedical Research Award (to JW), ImmunoX Bakar Trainee Momentum Award (to JW), Mark foundation (Endeavor grant to AG), Gazarian Foundation (to AG), Breast Cancer Research Foundation (to HR, AG), National Institutes of Health (U01 CA199315 to ZW, JW, AG), US National Institutes of Health 1R01CA223817 (AG), and the Chan Zuckerberg Biohub.

AUTHOR CONTRIBUTIONS

JW conceptualized the study. WT, JW, CSM, MYH, DS, AA analyzed data. JW, CMD, SH, AN, VS, EA performed animal studies. JW, AA, JHH, WT, CMD, JVL performed tissue processing. JW, CSM prepared MULTI-Seq libraries. WT, MB, JW prepared Smart-Seq2 libraries. AD, NN performed sequencing. JW, CMD, AA, LA, SH performed tissue stainings. JW, SDu and SYL performed tissue imaging and analysis. JW, WT wrote the manuscript. SDa, MYH, AOP, AG, CSM edited the manuscript. SDa, AOP, AG, ZG, ZW provided guidance and funding.

References

1. H. Dillekås, M. S. Rogers, and O. Straume. Are 90% of deaths from cancer caused by metastases? *Cancer Medicine*, 8(12):5574–5576, 2019.
2. C. Kim, R. Gao, E. Sei, R. Brandt, J. Hartman, T. Hatschek, N. Crosetto, T. Foukakis, and N. E. Navin. Chemoresistance evolution in Triple-Negative breast cancer delineated by Single-Cell sequencing. *Cell*, 173(4):879–893.e13, 2018.
3. B. Nguyen, C. Fong, A. Luthra, S. A. Smith, R. G. DiNatale, S. Nandakumar, H. Walch, W. K. Chatila, R. Madupuri, R. Kundra, C. M. Bielski, B. Mastrogiacomo, M. T. A. Donoghue, A. Boire, S. Chandarlapaty, K. Ganesh, J. J. Harding, C. A. Iacobuzio-Donahue, P. Razavi, E. Reznik, C. M. Rudin, D. Zamarin, W. Abida, G. K. Abou-Alfa, C. Aghajanian, A. Cercek, P. Chi, D. Feldman, A. L. Ho, G. Iyer, Y. Y. Janjigian, M. Morris, R. J. Motzer, E. M. O'Reilly, M. A. Postow, N. P. Raj, G. J. Riely, M. E. Robson, J. E. Rosenberg, A. Safonov, A. N. Shoushtari, W. Tap, M. Y. Teo, A. M. Varghese, M. Voss, R. Yaeger, M. G. Zauderer, N. Abu-Rustum, J. Garcia-Aguilar, B. Bochner, A. Hakimi, W. R. Jarnagin, D. R. Jones, D. Molena, L. Morris, E. Rios-Doria, P. Russo, S. Singer, V. E. Strong, D. Chakravarty, L. H. Ellenson, A. Gopalan, J. S. Reis-Filho, B. Weigelt, M. Ladanyi, M. Gonen, S. P. Shah, J. Massague, J. Gao, A. Zehir, M. F. Berger, D. B. Solit, S. F. Bakhoum, F. Sanchez-Vega, and N. Schultz. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell*, 185 (3):563–575.e11, 2022.
4. C. L. Chaffer, B. P. San Juan, E. Lim, and R. A. Weinberg. EMT, cell plasticity and metastasis. *Cancer and Metastasis Reviews*, 35(4):645–654, 2016.
5. D. P. Cook and B. C. Vanderhyden. Context specificity of the EMT transcriptional response. *Nat. Commun.*, 11(1):2142, 2020.
6. D. P. Cook and B. C. Vanderhyden. Transcriptional census of epithelial-mesenchymal plasticity in cancer. *Sci Adv*, 8(1):eabi7640, 2022.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7. G. S. Kinker, A. C. Greenwald, R. Tal, Z. Orlova, M. S. Cuoco, J. M. McFarland, A. Warren, C. Rodman, J. A. Roth, S. A. Bender, B. Kumar, J. W. Rocco, P. A. C. M. Fernandes, C. C. Mader, H. Keren-Shaul, A. Plotnikov, H. Barr, A. Tsherniak, O. Rozenblatt-Rosen, V. Krizhanovsky, S. V. Puram, A. Regev, and I. Tirosh. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.*, 52(11):1208–1218, 2020.
8. N. M. Aiello, T. Brabletz, Y. Kang, M. A. Nieto, R. A. Weinberg, and B. Z. Stanger. Upholding a role for EMT in pancreatic cancer metastasis. *Nature*, 547(7661):E7–E8, 2017.
9. X. Ye, T. Brabletz, Y. Kang, G. D. Longmore, M. A. Nieto, B. Z. Stanger, J. Yang, and R. A. Weinberg. Upholding a role for EMT in breast cancer metastasis. *Nature*, 547(7661): E1–E3, 2017.
10. X. Zheng, J. L. Carstens, J. Kim, M. Scheible, J. Kaye, H. Sugimoto, C.-C. Wu, V. S. LeBleu, and R. Kalluri. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature*, 527(7579):525–530, 2015.
11. K. R. Fischer, A. Durrans, S. Lee, J. Sheng, F. Li, S. T. C. Wong, H. Choi, T. El Rayes, S. Ryu, J. Troeger, R. F. Schwabe, L. T. Vahdat, N. K. Altorki, V. Mittal, and D. Gao. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature*, 527(7579):472–476, 2015.
12. J. Yang, P. Antin, G. Berx, C. Blanpain, T. Brabletz, M. Bronner, K. Campbell, A. Cano, J. Casanova, G. Christofori, S. Dedhar, R. Derynck, H. L. Ford, J. Fuxe, A. García de Herreros, G. J. Goodall, A.-K. Hadjantonakis, R. Y. J. Huang, C. Kalcheim, R. Kalluri, Y. Kang, Y. Khew-Goodall, H. Levine, J. Liu, G. D. Longmore, S. A. Mani, J. Massagué, R. Mayor, D. McClay, K. E. Mostov, D. F. Newgreen, M. A. Nieto, A. Puisieux, R. Runyan, P. Savagner, B. Stanger, M. P. Stemmler, Y. Takahashi, M. Takeichi, E. Theveneau, J. P. Thiery, E. W. Thompson, R. A. Weinberg, E. D. Williams, J. Xing, B. P. Zhou, G. Sheng, and EMT International Association (TEMTIA). Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.*, 21(6):341–352, 2020.
13. P. B. Gupta, I. Pastushenko, A. Skibinski, C. Blanpain, and C. Kuperwasser. Phenotypic plasticity: Driver of cancer initiation, progression, and therapy resistance. *Cell Stem Cell*, 24(1):65–78, 2019.
14. E. D. Williams, D. Gao, A. Redfern, and E. W. Thompson. Controversies around epithelial-mesenchymal plasticity in cancer metastasis. *Nat. Rev. Cancer*, 19(12):716–732, 2019.
15. W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.*, 8:15081, 2017.
16. S. Z. Wu, G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy, N. Bartonicek, T. Wang, L. Larsson, D. Kaczorowski, N. I. Weisenfeld, C. R. Uytingco, J. G. Chew, Z. W. Bent, C.-L. Chan, V. Gnanasambandapillai, C.-A. Dutertre, L. Gluch, M. N. Hui, J. Beith, A. Parker, E. Robbins, D. Segara, C. Cooper, C. Mak, B. Chan, S. Warrier, F. Ginhoux, E. Millar, J. E. Powell, S. R. Williams, X. S. Liu, S. O'Toole, E. Lim, J. Lundeberg, C. M. Perou, and A. Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.*, 53(9):1334–1347, 2021.
17. B. Pal, Y. Chen, F. Vaillant, B. D. Capaldo, R. Joyce, X. Song, V. L. Bryant, J. S. Penington, L. Di Stefano, N. T. Ribera, S. Wilcox, G. B. Mann, A. T. Papenfuss, G. J. Lindeman, G. K. Smyth, J. E. Visvader, and kConFab. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *The EMBO Journal*, 40(11), 2021.
18. R. Gao, S. Bai, Y. C. Henderson, Y. Lin, A. Schalck, Y. Yan, T. Kumar, M. Hu, E. Sei, A. Davis, F. Wang, S. F. Shaitelman, J. R. Wang, K. Chen, S. Moulder, S. Y. Lai, and N. E. Navin. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.*, 39(5):599–608, 2021.
19. A. Maynard, C. E. McCoach, J. K. Rotow, L. Harris, F. Haderk, D. L. Kerr, E. A. Yu, E. L. Schenk, W. Tan, A. Zee, M. Tan, P. Gui, T. Lea, W. Wu, A. Urisman, K. Jones, R. Sit, P. K. Kolli, E. Seeley, Y. Gesthalter, D. D. Le, K. A. Yamauchi, D. M. Naeger, S. Bandyopadhyay, K. Shah, L. Cech, N. J. Thomas, A. Gupta, M. Gonzalez, H. Do, L. Tan, B. Bacaltos, R. Gomez-Sjoberg, M. Gubens, T. Jahan, J. R. Kratz, D. Jablons, N. Neff, R. C. Doebele, J. Weissman, C. M. Blakely, S. Darmanis, and T. G. Bivona. Therapy-Induced evolution of human lung cancer revealed by Single-Cell RNA sequencing. *Cell*, 182(5):1232–1251.e22, 2020.
20. L. N. G. Castro, L. Nicolas Gonzalez Castro, I. Tirosh, and M. L. Suvà. Decoding cancer biology one cell at a time. *Cancer Discovery*, 11(4):960–970, 2021.
21. A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.
22. A. Gavish, M. Tyler, D. Simkin, D. Kovarsky, L. N. Gonzalez Castro, D. Halder, R. Chanoch-Myers, J. Laffy, M. Mints, A. R. Greenwald, A. Wider, R. Tal, A. Spitzer, T. Hara, A. Tirosh, S. V. Puram, M. L. Suva, and I. Tirosh. The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors. 2021.
23. S. V. Puram, I. Tirosh, A. S. Parikh, A. P. Patel, K. Yizhak, S. Gillespie, C. Rodman, C. L. Luo, E. A. Mroz, K. S. Emerick, D. G. Deschler, M. A. Varvares, R. Mylvaganam, O. Rozenblatt-Rosen, J. W. Rocco, W. C. Faquin, D. T. Lin, A. Regev, and B. E. Bernstein. Single-Cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624.e24, 2017.
24. A. M. Laughney, J. Hu, N. R. Campbell, S. F. Bakhoum, M. Setty, V.-P. Lavallée, Y. Xie, I. Masilionis, A. J. Carr, S. Kottapalli, V. Allaj, M. Mattar, N. Rekhtman, J. B. Xavier, L. Mazutis, J. T. Poirier, C. M. Rudin, D. Pe'er, and J. Massagué. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat. Med.*, 26(2):259–269, 2020.
25. Y. S. DeRose, G. Wang, Y.-C. Lin, P. S. Bernard, S. S. Buys, M. T. W. Ebbert, R. Factor, C. Matsen, B. A. Milash, E. Nelson, L. Neumayer, R. L. Randall, I. J. Stijleman, B. E. Welm, and A. L. Welm. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.*, 17(11): 1514–1520, 2011.
26. J. W. Cassidy, A. S. Batra, W. Greenwood, and A. Bruna. Patient-derived tumour xenografts for breast cancer drug discovery. *Endocr. Relat. Cancer*, 23(12):T259–T270, 2016.
27. L. E. Dobrolecki, S. D. Airhart, D. G. Alferez, S. Aparicio, F. Behbod, M. Bentires-Alj, C. Brisken, C. J. Bult, S. Cai, R. B. Clarke, H. Dowst, M. J. Ellis, E. Gonzalez-Suarez, R. D. Iggo, P. Kabos, S. Li, G. J. Lindeman, E. Marangoni, A. McCoy, F. Meric-Bernstam, H. Piwnica-Worms, M.-F. Poupon, J. Reis-Filho, C. A. Sartorius, V. Scabia, G. Sflomos, Y. Tu, F. Vaillant, J. E. Visvader, A. Welm, M. S. Wicha, and M. T. Lewis. Patient-derived xenograft (PDX) models in basic and translational breast cancer research. *Cancer Metastasis Rev.*, 35(4):547–573, 2016.
28. X. Zhang, S. Claerhout, A. Prat, L. E. Dobrolecki, I. Petrovic, Q. Lai, M. D. Landis, L. Wiechmann, R. Schiff, M. Giuliano, H. Wong, S. W. Fuqua, A. Contreras, C. Gutierrez, J. Huang, S. Mao, A. C. Pavlick, A. M. Froehlich, M.-F. Wu, A. Tsimelzon, S. G. Hilsenbeck, E. S. Chen, P. Zuloaga, C. A. Shaw, M. F. Rimawi, C. M. Perou, G. B. Mills, J. C. Chang, and M. T. Lewis. A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.*, 73(15):4885–4897, 2013.
29. D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C.-Y. Wang, P. Yaswen, A. Goga, and Z. Werb. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 526 (7571):131–135, 2015.
30. P. F. McAnena, A. McGuire, A. Ramli, C. Curran, C. Malone, R. McLaughlin, K. Barry, J. A. L. Brown, and M. J. Kerin. Breast cancer subtype discordance: impact on post-recurrence survival and potential treatment options. *BMC Cancer*, 18(1):203, 2018.
31. L. S. Lindström, E. Karlsson, U. M. Wilking, U. Johansson, J. Hartman, E. K. Lidbrink, T. Hatschek, L. Skoog, and J. Bergh. Clinically used breast cancer markers such as estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 are unstable throughout tumor progression. *J. Clin. Oncol.*, 30(21):2601–2608, 2012.
32. G. Aurilio, D. Disalvatore, G. Pruneri, V. Bagnardi, G. Viale, G. Curigliano, L. Adamoli, E. Munzone, A. Sciandivasci, F. De Vita, A. Goldhirsch, and F. Nolè. A meta-analysis of oestrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 discordance between primary breast cancer and metastases. *Eur. J. Cancer*, 50(2): 277–289, 2014.
33. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, 16:278, 2015.
34. R. T. Davis, K. Blake, D. Ma, M. B. I. Gabra, G. A. Hernandez, A. T. Phung, Y. Yang, D. Maurer, A. E. Y. T. Lefebvre, H. Alshetaiwi, Z. Xiao, J. Liu, J. W. Locasale, M. A. Digman, E. Mjolsness, M. Kong, Z. Werb, and D. A. Lawson. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nat. Cell Biol.*, 22(3):310–320, 2020.
35. B. Győrffy. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput. Struct. Biotechnol. J.*, 19:4101–4109, 2021.
36. C. S. McGinnis, D. M. Patterson, J. Winkler, D. N. Conrad, M. Y. Hein, V. Srivastava, J. L. Hu, L. M. Murrow, J. S. Weissman, Z. Werb, E. D. Chow, and Z. J. Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, 2019.
37. W.-Y. Lee, W.-C. Su, P.-W. Lin, H.-R. Guo, T.-W. Chang, and H. H. W. Chen. Expression of S100A4 and met: potential predictors for metastasis and survival in early-stage breast cancer. *Oncology*, 66(6):429–438, 2004.
38. M. P. Davies, P. S. Rudland, L. Robertson, E. W. Parry, P. Jolicoeur, and R. Barraclough. Expression of the calcium-binding protein S100A4 (p9ka) in MMTV-neu transgenic mice induces metastasis of mammary tumours. *Oncogene*, 13(8):1631–1637, 1996.
39. H. Xu, M. Li, Y. Zhou, F. Wang, X. Li, L. Wang, and Q. Fan. S100A4 participates in epithelial-mesenchymal transition in breast cancer via targeting MMP2. *Tumour Biol.*, 37 (3):2925–2932, 2016.
40. T. M. Horm and J. A. Schroeder. MUC1 and metastatic cancer. *Cell Adhesion & Migration*, 7(2):187–198, 2013.
41. N. Muthalagu, T. Monteverde, X. Raffo-Iraolagoitia, R. Wiesheu, D. Whyte, A. Hedley, S. Laing, B. Kruspig, R. Upstill-Goddard, R. Shaw, S. Neidler, C. Rink, S. A. Karim, K. Gyuraszova, C. Nixon, W. Clark, A. V. Biankin, L. M. Carlin, S. B. Coffelt, O. J. Sansom, J. P. Morton, and D. J. Murphy. Repression of the type I interferon pathway underlies MYC- and KRAS-Dependent evasion of NK and B cells in pancreatic ductal adenocarcinoma. *Cancer Discovery*, 10(6):872–887, 2020.
42. J. V. Lee, F. Housley, C. Yau, R. Nakagawa, J. Winkler, J. M. Anttila, P. M. Munne, M. Savelius, K. E. Houlahan, D. Van de Mark, G. Hemmati, G. A. Hernandez, Y. Zhang, S. Samson, C. Baas, L. J. Esserman, L. J. van 't Veer, H. S. Rugo, C. Curtis, J. Klefström, M. Matloubian, and A. Goga. Combinatorial immunotherapies overcome MYC-driven immune evasion in triple negative breast cancer. *Nat Commun*, 13(1):3671, 2022.
43. T. Z. Tan, Q. H. Miow, Y. Miki, T. Noda, S. Mori, R. Y.-J. Huang, and J. P. Thiery. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.*, 6(10):1279–1293, 2014.
44. V. Padmanaban, I. Krol, Y. Suhail, B. M. Szczerba, N. Aceto, J. S. Bader, and A. J. Ewald. E-cadherin is required for metastasis in multiple models of breast cancer. *Nature*, 573 (7774):439–444, 2019.
45. I. Pastushenko, A. Brisebarre, A. Sifrim, M. Fioramonti, T. Revenco, S. Boumahdi, A. Van Keymeulen, D. Brown, V. Moers, S. Lemaire, S. De Clercq, E. Minguijón, C. Balsat, Y. Sokolow, C. Dubois, F. De Cock, S. Scozzaro, F. Sopena, A. Lanas, N. D'Haene, I. Salmon, J.-C. Marine, T. Voet, P. A. Sotiropoulou, and C. Blanpain. Identification of the tumour transition states occurring during EMT. *Nature*, 556(7702):463–468, 2018.
46. K. P. Simeonov, C. N. Byrns, M. L. Clark, R. J. Norgard, B. Martin, B. Z. Stanger, J. Shendure, A. McKenna, and C. J. Lengner. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell*, 39(8):1150–1162.e9, 2021.
47. M. Al-Hajj, M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, and M. F. Clarke. Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, 100(7):3983–3988, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

48. P. Baumann, N. Cremers, F. Kroese, G. Orend, R. Chiquet-Ehrismann, T. Uede, H. Yagita, and J. P. Sleeman. CD24 expression causes the acquisition of multiple cellular properties associated with tumor growth and metastasis. *Cancer Res.*, 65(23):10783–10793, 2005.
49. I. Malanchi, A. Santamaria-Martínez, E. Susanto, H. Peng, H.-A. Lehr, J.-F. Delaloye, and J. Huelsken. Interactions between cancer stem cells and their niche govern metastatic colonization. *Nature*, 481(7379):85–89, 2011.
50. Y. Del Pozo Martin, D. Park, A. Ramachandran, L. Ombrato, F. Calvo, P. Chakravarty, B. Spencer-Dene, S. Derzsi, C. S. Hill, E. Sahai, and I. Malanchi. Mesenchymal cancer Cell-Stroma crosstalk promotes niche activation, epithelial reversion, and metastatic colonization. *Cell Rep.*, 13(11):2456–2469, 2015.
51. M. J. Meyer, J. M. Fleming, M. A. Ali, M. W. Pesesky, E. Ginsburg, and B. K. Vonderhaar. Dynamic regulation of CD24 and the invasive, CD44posCD24neg phenotype in breast cancer cell lines. *Breast Cancer Res.*, 11(6):R82, 2009.
52. A. Grosse-Wilde, A. Fouquier d'Hérouël, E. McIntosh, G. Ertaylan, A. Skupin, R. E. Kuestner, A. del Sol, K.-A. Walters, and S. Huang. Stemness of the hybrid Epithelial/Mesenchymal state in breast cancer and its association with poor survival. *PLoS One*, 10(5):e0126522, 2015.
53. S. A. Mani, W. Guo, M.-J. Liao, E. N. Eaton, A. Ayyanan, A. Y. Zhou, M. Brooks, F. Reinhard, C. C. Zhang, M. Shipitsin, L. L. Campbell, K. Polyak, C. Brisken, J. Yang, and R. A. Weinberg. The Epithelial-Mesenchymal transition generates cells with properties of stem cells. *Cell*, 133(4):704–715, 2008.
54. C. Kröger, A. Afeyan, J. Mraz, E. N. Eaton, F. Reinhardt, Y. L. Khodor, P. Thiru, B. Bierie, X. Ye, C. B. Burge, and R. A. Weinberg. Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.*, 116(15): 7353–7362, 2019.
55. J.-H. Shin, S.-W. Kim, C.-M. Lim, J.-Y. Jeong, C.-S. Piao, and J.-K. Lee. alphab-crystallin suppresses oxidative stress-induced astrocyte apoptosis by inhibiting caspase-3 activation. *Neurosci. Res.*, 64(4):355–361, 2009.
56. M. C. Kamradt, M. Lu, M. E. Werner, T. Kwan, F. Chen, A. Strohecker, S. Oshita, J. C. Wilkinson, C. Yu, P. G. Oliver, C. S. Duckett, D. J. Buchsbaum, A. F. LoBuglio, V. Craig Jordan, and V. L. Cryns. The small heat shock protein αB-crystallin is a novel inhibitor of TRAIL-induced apoptosis that suppresses the activation of caspase-3. *Journal of Biological Chemistry*, 280(12):11059–11066, 2005.
57. D. Malin, E. Strekalova, V. Petrovic, H. Rajanala, B. Sharma, A. Ugolkov, W. J. Gradishar, and V. L. Cryns. ERK-regulated αB-crystallin induction by matrix detachment inhibits anoikis and promotes lung metastasis in vivo. *Oncogene*, 34(45):5626–5634, 2015.
58. Y. Mao, D.-W. Zhang, H. Lin, L. Xiong, Y. Liu, Q.-D. Li, J. Ma, Q. Cao, R.-J. Chen, J. Zhu, and Z.-Q. Feng. Alpha b-crystallin is a new prognostic marker for laryngeal squamous cell carcinoma. *J. Exp. Clin. Cancer Res.*, 31:101, 2012.
59. X.-Y. Huang, A.-W. Ke, G.-M. Shi, X. Zhang, C. Zhang, Y.-H. Shi, X.-Y. Wang, Z.-B. Ding, Y.-S. Xiao, J. Yan, S.-J. Qiu, J. Fan, and J. Zhou. αB-crystallin complexes with 14-3-3ζ to induce epithelial-mesenchymal transition and resistance to sorafenib in hepatocellular carcinoma. *Hepatology*, 57(6):2235–2247, 2013.
60. J. V. Moyano, J. R. Evans, F. Chen, M. Lu, M. E. Werner, F. Yehiely, L. K. Diaz, D. Turbin, G. Karaca, E. Wiley, T. O. Nielsen, C. M. Perou, and V. L. Cryns. AlphaB-crystallin is a novel oncoprotein that predicts poor clinical outcome in breast cancer. *J. Clin. Invest.*, 116 (1):261–270, 2006.
61. C. van de Schootbrugge, J. Bussink, P. N. Span, F. C. G. J. Sweep, R. Grénman, H. Stegeman, G. J. M. Pruijn, J. H. A. M. Kaanders, and W. C. Boelens. αB-crystallin stimulates VEGF secretion and tumor cell migration and correlates with enhanced distant metastasis in head and neck squamous cell carcinoma. *BMC Cancer*, 13:128, 2013.
62. K. D. Voduc, T. O. Nielsen, C. M. Perou, J. C. Harrell, C. Fan, H. Kennecke, A. J. Minn, V. L. Cryns, and M. C. U. Cheang. αB-crystallin expression in breast cancer is associated with brain metastasis. *NPJ Breast Cancer*, 1, 2015.
63. D. Chelouche-Lev, H. M. Kluger, A. J. Berger, D. L. Rimm, and J. E. Price. ?b-crystallin as a marker of lymph node involvement in breast carcinoma. *Cancer*, 100(12):2543–2548, 2004.
64. D. Malin, E. Strekalova, V. Petrovic, A. M. Deal, A. Al Ahmad, B. Adamo, C. Ryan Miller, A. Ugolkov, C. Livasy, K. Fritchie, E. Hamilton, K. Blackwell, J. Geradts, M. Ewend, L. Carey, E. V. Shusta, C. K. Anders, and V. L. Cryns. αB-Crystallin: A novel regulator of breast cancer metastasis to the brain. *Clinical Cancer Research*, 20(1):56–67, 2014.
65. E. Hirata, K. Ishibashi, S. Kohsaka, K. Shinjo, S. Kojima, Y. Kondo, H. Mano, S. Yano, E. Kiyokawa, and E. Sahai. The brain microenvironment induces DNMT1 suppression and indolence of metastatic cancer cells. *iScience*, 23(9):101480, 2020.
66. M. Montagner, R. Bhome, S. Hooper, P. Chakravarty, X. Qin, J. Sufi, A. Bhargava, C. D. H. Ratcliffe, Y. Naito, A. Pocaterra, C. J. Tape, and E. Sahai. Crosstalk with lung epithelial cells regulates sfrp2-mediated latency in breast cancer dissemination. *Nat. Cell Biol.*, 22 (3):289–296, 2020.
67. A. Bose, M.-T. Teh, I. C. Mackenzie, and A. Waseem. Keratin k15 as a biomarker of epidermal stem cells. *Int. J. Mol. Sci.*, 14(10):19385–19398, 2013.
68. P. Zhong, R. Shu, H. Wu, Z. Liu, X. Shen, and Y. Hu. Low KRT15 expression is associated with poor prognosis in patients with breast invasive carcinoma. *Exp. Ther. Med.*, 21(4): 305, 2021.
69. M. A. A. K. Folgueira, H. Brentani, M. L. H. Katayama, D. F. C. Patrão, D. M. Carraro, M. Mourão Netto, E. M. Barbosa, J. R. F. Caldeira, A. P. S. Abreu, E. C. Lyra, J. H. L. Kaiano, L. D. Mota, A. H. J. F. M. Campos, M. S. Maciel, M. Dellamano, O. L. S. D. Caballero, and M. M. Brentani. Gene expression profiling of clinical stages II and III breast cancer. *Braz. J. Med. Biol. Res.*, 39(8):1101–1113, 2006.
70. D. Cimino, L. Fuso, C. Sfiligoi, N. Biglia, R. Ponzone, F. Maggiorotto, G. Russo, L. Cicatiello, A. Weisz, D. Taverna, P. Sismondi, and M. De Bortoli. Identification of new genes associated with breast cancer progression by gene expression analysis of predefined sets of neoplastic tissues. *Int. J. Cancer*, 123(6):1327–1338, 2008.
71. X. Rao, J. Wang, H. M. Song, B. Deng, and J. G. Li. KRT15 overexpression predicts poor prognosis in colorectal cancer. *Neoplasma*, 67(02):410–414, 2020.
72. J.-B. Lin, Z. Feng, M.-L. Qiu, R.-G. Luo, X. Li, and B. Liu. KRT 15 as a prognostic biomarker is highly expressed in esophageal carcinoma. *Future Oncol.*, 16(25):1903–1909, 2020.
73. R. J. Norgard, J. R. Pitarresi, R. Maddipati, N. M. Aiello-Couzo, D. Balli, J. Li, T. Yamazoe, M. D. Wengyn, I. D. Millstein, I. W. Folkert, D. N. Rosario-Berrios, I.-K. Kim, J. B. Bassett, R. Payne, C. T. Berry, X. Feng, K. Sun, M. Cioffi, P. Chakraborty, M. K. Jolly, J. S. Gutkind, D. Lyden, B. D. Freedman, J. K. Foskett, A. K. Rustgi, and B. Z. Stanger. Calcium signaling induces a partial EMT. *EMBO Rep.*, 22(9):e51872, 2021.
74. E. Bulk, B. Sargin, U. Krug, A. Hascher, Y. Jun, M. Knop, C. Kerkhoff, V. Gerke, R. Liersch, R. M. Mesters, M. Hotfilder, A. Marra, S. Koschmieder, M. Dugas, W. E. Berdel, H. Serve, and C. Müller-Tidow. S100A2 induces metastasis in non-small cell lung cancer. *Clin. Cancer Res.*, 15(1):22–29, 2009.
75. W.-C. Tsai, S.-T. Tsai, Y.-T. Jin, and L.-W. Wu. Cyclooxygenase-2 is involved in S100A2-mediated tumor suppression in squamous cell carcinoma. *Mol. Cancer Res.*, 4(8):539–547, 2006.
76. A. R. Bresnick, D. J. Weber, and D. B. Zimmer. S100 proteins in cancer. *Nat. Rev. Cancer*, 15(2):96–109, 2015.
77. S. Naz, P. Ranganathan, P. Bodapati, A. H. Shastry, L. N. Mishra, and P. Kondaiah. Regulation of S100A2 expression by TGF-β-induced MEK/ERK signalling and its role in cell migration/invasion. *Biochem. J*, 447(1):81–91, 2012.
78. G. Huang, J. Zhang, G. Qing, D. Liu, X. Wang, Y. Chen, Y. Li, and S. Guo. Silencing relieves Epithelial-Mesenchymal transition in pulmonary fibrosis by inhibiting the Wnt/β-Catenin signaling pathway. *DNA Cell Biol.*, 40(1):18–25, 2021.
79. S. Naz, M. Bashir, P. Ranganathan, P. Bodapati, V. Santosh, and P. Kondaiah. Protumorigenic actions of S100A2 involve regulation of PI3/Akt signaling and functional interaction with smad3. *Carcinogenesis*, 35(1):14–23, 2014.
80. H. Wang, X. Hu, F. Yang, and H. Xiao. mir-325-3p promotes the proliferation, invasion, and EMT of breast cancer cells by directly targeting S100A2. *Oncol. Res.*, 28(7):731–744, 2021.
81. C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.
82. A. Prat, C. Fan, A. Fernández, K. A. Hoadley, R. Martinello, M. Vidal, M. Viladot, E. Pineda, A. Arance, M. Muñoz, L. Paré, M. C. U. Cheang, B. Adamo, and C. M. Perou. Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy. *BMC Med.*, 13:303, 2015.
83. I. Pastushenko and C. Blanpain. EMT transition states during tumor progression and metastasis. *Trends in Cell Biology*, 29(3):212–226, 2019.
84. M. Q. Reeves, E. Kandyba, S. Harris, R. Del Rosario, and A. Balmain. Multicolour lineage tracing reveals clonal dynamics of squamous carcinoma evolution from initiation to metastasis. *Nat. Cell Biol.*, 20(6):699–709, 2018.
85. K. J. Cheung, V. Padmanaban, V. Silvestri, K. Schipper, J. D. Cohen, A. N. Fairchild, M. A. Gorin, J. E. Verdone, K. J. Pienta, J. S. Bader, and A. J. Ewald. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc. Natl. Acad. Sci. U. S. A.*, 113(7):E854–63, 2016.
86. C. Ginestier, M. H. Hur, E. Charafe-Jauffret, F. Monville, J. Dutcher, M. Brown, J. Jacquemier, P. Viens, C. G. Kleer, S. Liu, A. Schott, D. Hayes, D. Birnbaum, M. S. Wicha, and G. Dontu. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell*, 1(5):555–567, 2007.
87. S. Liu, Y. Cong, D. Wang, Y. Sun, L. Deng, Y. Liu, R. Martin-Trevino, L. Shang, S. P. McDermott, M. D. Landis, S. Hong, A. Adams, R. D'Angelo, C. Ginestier, E. Charafe-Jauffret, S. G. Clouthier, D. Birnbaum, S. T. Wong, M. Zhan, J. C. Chang, and M. S. Wicha. Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Reports*, 2(1):78–91, 2014.
88. O. H. Ocaña, R. Córcoles, A. Fabra, G. Moreno-Bueno, H. Acloque, S. Vega, A. Barrallo-Gimeno, A. Cano, and M. A. Nieto. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer prrx1. *Cancer Cell*, 22(6):709–724, 2012.
89. J. H. Tsai, J. L. Donaher, D. A. Murphy, S. Chau, and J. Yang. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer Cell*, 22(6):725–736, 2012.
90. S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1098, 2013.
91. S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171–181, 2014.
92. T. T. M. Consortium, The Tabula Muris Consortium, O. Coordination, L. Coordination, Organ collection and processing, Library preparation and sequencing, C. D. Analysis, C. T. Annotation, Writing group, Supplemental text writing group, and Principal investigators. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727): 367–372, 2018.
93. S. Picelli, A. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, 24(12):2033–2040, 2014.
94. B. P. Hennig, L. Velten, I. Racke, C. S. Tu, M. Thoms, V. Rybin, H. Besir, K. Remans, and L. M. Steinmetz. Large-Scale Low-Cost NGS library preparation using a robust tn5 purification and tagmentation protocol. *G3*, 8(1):79–89, 2018.
95. F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, 2018.
96. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21, 2019.
97. M. S. Kowalczyk, I. Tirosh, D. Heckl, T. N. Rao, A. Dixit, B. J. Haas, R. K. Schneider, A. J. Wagers, B. L. Ebert, and A. Regev. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.*, 25(12): 1860–1872, 2015.
98. S. Yang, S. E. Corbett, Y. Koga, Z. Wang, W. E. Johnson, M. Yajima, and J. D. Campbell.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.*, 21 (1):57, 2020.

99. M. L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
100. trevis. GitHub - trevismd/statannotations: add statistical significance annotations on seaborn plots. further development of statannot, with bugfixes, new features, and a different API. `https://github.com/trevismd/statannotations`. Accessed: 2022-4-25.
101. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(85):2825–2830, 2011.
102. G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, and A. Sergushichev. Fast gene set enrichment analysis. *bioRxiv*, 2021.
103. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, 2005.
104. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6): 417–425, 2015.
105. C. Davidson-Pilon. lifelines: survival analysis in python. *J. Open Source Softw.*, 4(40): 1317, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Animal experiments

Fresh primary breast tumor samples were obtained from the Cooperative Human Tissue Network (CHTN) in accordance with the Institutional Review Boards' approval. Tissues were received as de-identified samples and all subjects provided written informed consent. Medical reports were obtained without personally identifiable information. The UCSF Institutional Animal Care and Use Committee (IACUC) reviewed and approved all animal experiments. Tumor tissues were cut into 1 mm thick chunks and orthotopically transplanted into cleared mammary fat pats of 4-week-old NOD-SCID gamma mice to generate novel PDX models (J53353, J2036, and J55454, Supplementary Table 1). Established PDX lines were transplanted in the same way and as previously described (25, 28). Once palpable, tumors were measured 2×/week using a caliper to monitor growth kinetics. Tumor volume was calculated using following formula: $\frac{\pi}{6} \cdot \text{height}^{1.5} \cdot \text{width}^{1.5}$. Unless otherwise noted, all PDX animals were euthanized at the endpoint, when the primary tumor reached 2.5 cm in diameter. In resection experiments, tumors were surgically removed at 1.0–2.0 cm in diameter. Resected animals were allowed to grow metastases until endpoint was reached (2.5 cm diameter of recurrent tumor). At endpoint, primary tumor and metastatic lungs were harvested, cut in small chunks and cryopreserved using Recovery Cell Culture Freezing Medium (Thermo Fisher, 12648010) and stored in liquid nitrogen until further analysis.

Histology and tissue staining

For each PDX animal, after dissection, the middle and postcaval lobes of the right lung were fixed in 4 % PFA overnight and processed for paraffin embedding. For histological analysis, tissue sections were stained with haematoxylin and eosin using standard protocols. Tissue slides were scanned (Zeiss Axio ScanZ.1) and images were analyzed using QuPath. Metastatic foci were easily identified by a larger nuclei/cytoplasm ratio. Micrometastases were defined as < 10 tumor cells, intermediate metastatic foci 10–100 cells and macrometastases >100 cells. Number and area of metastatic foci and total tissue area were determined.

Lysis plate preparation

Lysis plates were prepared by dispensing 0.4 µL lysis buffer (0.5 U Recombinant RNase Inhibitor (Takara Bio, 2313B), 0.0625 % Triton™ X-100 (Sigma, 93443-100ML), 3.125 mM dNTP mix (Thermo Fisher, R0193), 3.125 µM Oligo-dT30VN (IDT, 5′-AAGCAGTGGTATCAACGCAGAGTACT30VN-3′) and 1:600,000 ERCC RNA spike-in mix (Thermo Fisher, 4456740)) into 384-well hard-shell PCR plates (Biorad HSP3901) using a Dragonfly liquid handler (STP Labtech). All plates were then spun down for 1 min at 3220×g and snap-frozen on dry ice. Plates were stored at -80 °C until used for sorting.

Sample preparation and FACS sorting

Primary tumor and metastatic lung tissues were processed, stained, MULTI-Seq labeled and FACS sorted as previously described (36). In brief, tissues were thawed, dissociated in digestion media containing 50 µg/ml Liberase TL (Sigma-Aldrich) and $2{\cdot}10^4$ U/ml DNase I (Sigma-Aldrich) in DMEM/F12 (Gibco) using standard GentleMacs (37C_m_LDK_1, 37_m_TDK1) protocols. Washed and filtered single-cell suspension were stained with viability dye (Zombie NIR, 1:500, BioLegend, no. 423105), blocked with Fc-block (1:200, Tonbo, 70-0161-U500), and with LIN (anti-mouse TER119-FITC, Thermo Fisher, 11-5921-82; anti-mouse CD31-FITC, Thermo Fisher, 11-0311-85; anti-mouse CD45-BV450, Tonbo, 75-0451-U100; anti-mouse MHC-I-APC, eBioscience, 17-5999-82) and anti-human CD298 (PE, BioLegend, 341704). For the Smart-Seq2 experiments, live, $\text{LIN}^-/\text{hCD298}^+$ primary tumor and metastatic cells were sorted directly into cooled lysis plates and snap-frozen until library preparation. If multiple plates were sorted from one PDX model, each plate contained half primary tumor and metastatic cells to avoid plate-specific batch effects. For the MULTI-Seq experiments, MULTI-Seq LMO barcode anchor and co-anchor were used at a final concentration of 2.5 µM directly after antibody staining before FACS sorting as described previously (36). For one experiment (PDX1) we used sets of three unique MULTI-seq barcodes/sample. After sorting, enriched live, $\text{LIN}^-/\text{hCD298}^+$ cells were pooled and loaded into 10x microfluidics lanes at an average loading concentration of about 30,000 cells/lane.

cDNA synthesis and library preparation

cDNA synthesis was performed using the Smart-seq2 protocol (90–92). Briefly, 384-well plates containing single-cell lysates were thawed on ice followed by first-strand synthesis. 0.6 µL of reaction mix (16.7 U/µl SMARTScribe Reverse Transcriptase (Takara Bio, 639538), 1.67 U/µl Recombinant RNase Inhibitor (Takara Bio, 2313B), 1.67X First-Strand Buffer (Takara Bio, 639538), 1.67 µM TSO (Exiqon, 5′-AAGCAGTGGTATCAACGCAGACTACATrGrG+G-3′), 8.33 mM DTT (Bioworld, 40420001-1), 1.67 M Betaine (Sigma, B0300-5VL), and 10 mM MgCl2 (Sigma, M1028-10X1ML)) was added to each well using a Dragonfly liquid handler (STP Labtech). Reverse transcription was carried out by incubating wells on a ProFlex 2x384 thermal-cycler (Thermo Fisher) at 42 °C for 90 min and stopped by heating at 70 °C for 5 min. Subsequently, 1.5 µL of PCR mix (1.67X KAPA HiFi HotStart ReadyMix (Kapa Biosystems, KK2602), 0.17 µM IS PCR primer (IDT, 5′-AAGCAGTGGTATCAACGCAGAGT-3′), and 0.038 U/µl Lambda Exonuclease (NEB, M0262L)) was added to each well with

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a Dragonfly liquid handler (STP Labtech), and second strand synthesis was performed on a ProFlex 2x384 thermal-cycler by using the following program: 1. 37 °C for 30 min, 2. 95°C for 3 min, 3. 23 cycles of 98 °C for 20 s, 67 °C for 15 s, and 72 °C for 4 min, and 4. 72 °C for 5 min. The amplified product was diluted 1:10 with 10 mM Tris-HCl (Thermo Fisher, 15568025). 0.6 µL of diluted product was transferred to a new 384-well plate using the Viaflow 384 channel pipette (Integra). Illumina sequencing libraries were prepared using a library preparation protocol modified from previously reported tagmentation-based protocols (93, 94). Briefly, tagmentation was carried out by mixing each well with 1 uL of 1.6x Homebrew Tn5 Tagmentation Buffer and 0.2 uL of homebrew Tn5 enzyme, then incubated at 55 °C for 3 min. The reaction was stopped by adding 0.4 µl 0.1 % sodium dodecyl sulfate (Fisher Scientific, BP166-500) and centrifuging at room temperature at 3,220g for 5 min. Indexing PCR reactions were performed by adding 0.4 µL of 5 µM i5 indexing primer, 0.4 µL of 5 µM i7 indexing primer, and 1.2 µL of Nextera NPM mix (Illumina). All reagents were dispensed with the Mosquito liquid handlers (STP Labtech). PCR amplification was carried out on a ProFlex 2x384 thermal cycler using the following program: 1. 72 °C for 3 min, 2. 95°C for 30 s, 3. 12 cycles of 98 °C for 10 s, 67 °C for 30 s, and 72 °C for 1 min, and 4. 72 °C for 5 min.

Library sequencing

Following library preparation, wells of each library plate were pooled using a Mosquito liquid handler (STP Labtech). Pooling was followed by two purifications using 0.7x AMPure beads (Fisher, A63881). Library quality was assessed using high sensitivity capillary electrophoresis on a Tapestation (Agilent), and libraries were quantified by qPCR (Kapa Biosystems, KK4923) on a CFX96 Touch Real-Time PCR Detection System (Biorad). Plate pools were normalized to 2 nM and equal volumes from library plates were mixed together to make the sequencing sample pool. Sequencing libraries from 384-well plates Libraries were sequenced on the NextSeq or NovaSeq 6000 Sequencing System (Illumina) using 2×100 bp paired-end reads and 2×12 bp index reads. NextSeq runs used high output kits, whereas NovaSeq runs used 300-cycle kit (Illumina, 20012860). PhiX control library was spiked in at ∼1 %.

Sequencing libraries from MULTI-seq

For the MULTI-Seq dataset, gene expression library preparation was performed using the v2 10x library kit with modifications as described previously to generate MULTI-seq libraries (36).

Data extraction

For Smart-Seq2, sequences from the NovaSeq or NextSeq were de-multiplexed using bcl2fastq v.2.19.0.316. Reads were aligned to the gencode V30 genome using STAR v.2.5.2b with parameters TK. Gene counts were produced using HTSEQ v.0.6.1p1 with default parameters, except 'stranded' was set to 'false', and 'mode' was set to 'intersection-nonempty'. For MULTI-Seq, sequences from the microfluidic droplet platform were de-multiplexed and aligned using CellRanger v.5.0.1, available from 10x Genomics with default parameters.

MULTI-Seq demultiplexing

MULTI-seq barcode FASTQs were converted to barcode UMI count matrices using the 'MULTIseq.preProcess' and 'MULTIseq.align' functions in the deMULTIplex R package (36) with default parameters. Notably, 'PDX3' FASTQs were randomly down-sampled to 10^8 total reads prior to UMI count matrix conversion in order to minimize computation time. Next, since cells labeled with the same MULTI-seq barcodes were split across multiple 10x Genomics microfluidics lanes in each experiment, MULTI-seq UMI count matrices from each lane were concatenated (PDX1 and PDX3 matrices were concatenated separately) to maximize classification performance. Using these concatenated matrices, samples were then classified into sample groups using the deMULTIplex workflow desired previously (with semi-supervised negative-cell reclassification) (36). Notably, since samples in the PDX1 experiment were encoded by sets of three unique MULTI-seq barcodes, classification was performed on each cell's median barcode count for each set. Moreover, cells with the top and bottom 5 % of MULTI-seq barcode counts were masked during the initial classification workflow in the PDX1 data, and were reintroduced as 'negatives' during semi-supervised negative cell reclassification. Analogous barcode count-merging and outlier-masking were not necessary for the PDX3 data, which was classified successfully using the default deMULTIplex workflow.

Data pre-processing

For Smart-seq 2 data, gene count tables were combined with the metadata variables using the Scanpy Python package version 1.8.1 (95). We removed the genes that were not expressed in at least 5 cells. Cells with less than 5,000 counts and 500 detected genes were removed. Additionally, we removed cells with more than 50 % mitochondrial genes and 20 % ERCC reads. The data were then normalized using size factor normalization such that every cell has 10,000 counts and log-transformed. We selected the top 2,000 genes with the highest standardized variance as the highly variable genes by using VST method from Seurat V3 (96), which is also implemented in Scanpy. Cell-cycle regression was performed after calculating the score of S and G2M phases for each cell (97). The data was then scaled to a maximum value of 10. We then computed principal component

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(PC) analysis, neighborhood graph and clustered the data using Louvain and Leiden methods. The data was visualized using UMAP projection.

For MULTI-seq data, for each 10x lane, we first removed the cells with less than 2,500 UMI and 250 genes and more than 50 % mitochondria reads by using the Scanpy Python package version 1.8.1. In addition, in order to filter out reads from ambient RNA, we ran DecontX (98) separately for each 10x lane by using default parameters. Next, we re-filtered the dataset from every 10x run when cells did not contain a minimum number of genes (250), minimum of counts/UMIs (2,500), and/or having more than 50 % mitochondria reads. The data were then further processed as described above for the Smart-Seq2 dataset. Cells were sample assigned using the MULTI-Seq demultiplexing result, thereby removing doublets but including unassigned 'negative' cells. In order to recover MULTI-Seq-unassigned 'negative' cells, we used DBSCAN clustering. Based on the results of the Smart-Seq2 data, cells from different PDX tumors would cluster distinct from each other in transcriptional space. Negative cells in one DBSCAN cluster were assigned as the same tumor sample as the majority of MULTI-Seq classified cells in that cluster. After completed cell assignment, all 10x runs were combined to one MULTI-Seq dataset and removed genes that were not expressed in at least 20 cells. We then performed normalization, log-transformation, finding highly variable genes, cell cycle regression, principal component analysis, UMAP dimension reduction, and Louvain and Leiden clustering as described for the Smart-Seq data.

EMP scoring and classification

We used GSVA scoring with its default parameters to assign each cell an epithelial (E-score) and mesenchymal score (M-score) using epithelial and mesenchymal marker genes (43). The EMP-score for each cell is calculated by the sum of E-score and M-score for that cell. Cells with an EMP-score > 0.2 were classified as mesenchymal-like cells, cells with an EMP-score < -0.2 were classified as epithelial-like cells, and cells with an EMP-score between -0.2 and 0.2 were classified as EMP intermediate cells.

Cell Phase Proportion Statistical Test

Cells were assigned in different cell cycle phases based on the cell cycle score calculated previously. Then cell phase proportions in each tumor were calculated in different groups in each category, such as EMP cell stage, sort, and metastatic potential group. Finally, Wilcoxon rank test was performed for comparing group to group in each category in each cell phase. The statistical tests were generated by using Seaborn (99) and Statannot (100) packages in Python.

ROC Curve and AUC Value

The cells were first ordered by their PC2 value either in the whole SS2 dataset or in individual tumor models. The true and false-positive rates were calculated based on the cell's label (primary tumor cell or metastatic cell) and PC2 value by using the "roc_curve" function from Scikit-learn (101). In addition, "roc_auc_score" from Scikit-learn was used to calculate the AUC value by using the true and false-positive rates.

Differentially expressed genes

Identifying DEGs in primary tumor and metastatic cells. We performed differential expression analysis between primary tumor and metastatic cells in the entire dataset using the Seurat function FindMarkers using MAST (33) and the tumor model as the latent variable. In addition, we identified DEGs between primary tumor and metastatic cells for each tumor sample separately using the same Seurat function without setting the latent variable. Genes with p-values < 0.05 and $\log_2$ fold change > 0.5 were kept for further analysis. After filtering, we combined the DEGs from tumors within the same metastatic potential group. We included DEGs that are shared between at least two tumors in the same metastatic potential group.

Identifying gene signatures associated with metastatic potential. To identify genes in the primary tumor that are associated with metastatic potential we first removed metastatic cells from the data. Then, we used the Seurat function FindMarkeres using the MAST test and tumor model as the latent variable for identifying DEGs between one individual tumor and all tumors in the other metastatic groups. Genes were filtered based on p-values < 0.05 and $\log_2$ fold change > 0.5. After filtering, we combined the up-regulated gene lists from tumors within the same metastatic potential group. Signature genes related to metastatic potential were determined as genes that are shared between at least two tumors in the same metastatic potential group.

Identifying EMP marker genes. We identified EMP marker genes for each EMP category (epithelial-like, EMP intermediate, mesenchymal-like) using the Seurat function FindMarkers using the MAST test and tumor model as the latent variable. Genes were filtered based on p-values < 0.05 and $\log_2$ fold change > 0.5.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene set enrichment analysis

To identify pathways that were enriched in primary tumor or metastatic cells, we used the fgsea package (102) with Hallmark and GO gene sets from MSigDB (103, 104). We examined pathways that were significantly enriched in at least four tumor models. Enriched pathways in highly and poorly metastatic signatures were identified with the online tool of MSigDB.

Survival Analysis

For survival analysis, we used the KM-plotter (35) website's breast cancer gene chip mRNA dataset. The mean expression of the signature genes were calculated for each sample in the dataset. The patient samples were separated based on the median of the mean expression in low and high expressing samples. Visualization was done using Lifelines Python package (105). The metastatic potential gene lists resulted from the overlap genes between the MULTI-seq and Smart-Seq2 datasets from each metastatic potential group. For EMP signature gene lists, the epithelial signature gene list and mesenchymal signature gene list were the overlapping genes from the top 100 differentially expressed genes in the MULTI-Seq and Smart-Seq2 datasets, and the intermediate EMP marker gene signature included the overlapping genes found in both MULTI-seq and Smart-Seq2 datasets. In addition, METABRIC (81) dataset was obtained from cBioPortal. We used "all_sample_Zscores" to create Kaplan-Meier survival plots (105) and perform logrank tests for each breast cancer subtype using the mean expression of the intermediate EMP marker genes.

Supplemental Tables

- Table S1. PDX info
- Table S2. SS2 met vs. primary DEGs filtered
  (Tabs: global, HCI005, H3404, H4272, HCI009, HCI011, HCI001, H5097, J2036, J53353, HCI010)
- Table S3. Overlaped DEGs primary tumor vs. metastasis per metastatic potential group
- Table S4. MULTI primary tumor 1 vs. rest DEGs filtered
  (Tabs: HCI002_MULTI, J55454_MULTI, HCI005_MULTI, H4272_MULTI, HCI011_MULTI, HCI001_MULTI, H5097_MULTI, J2036_MULTI, J53353_MULTI, HCI010_MULTI)
- Table S5. SS2 primary tumor 1 vs. rest DEGs filtered
  (Tabs: J55454_SS2, H5471_SS2, HCI005_SS2, H3404_SS2, H4272_SS2, HCI009_SS2, HCI011_SS2, HCI001_SS2, H5097_SS2, J2036_SS2, J53353_SS2, HCI010_SS2)
- Table S6. Low, moderate, high met. signature
- Table S7. SS2 EMP DEGs filtered
  (Tabs: Epithelial-like, Mesenchymal, EMP intermediate)
- Table S8. MULTI EMP DEGs filtered
  (Tabs: Epithelial-like, Mesenchymal, EMP intermediate)

Data and materials availability

Raw sequencing files are available at the NCBI BioProject number PRJNA847563. Raw and processed data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE210283. Processed data are available as h5ad files on figshare (https://figshare.com/s/328942c0b8dc9aa69be1 and https://figshare.com/s/b53f327a8b612a7b2eeb). Code is available on github https://github.com/czbiohub/scBC.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

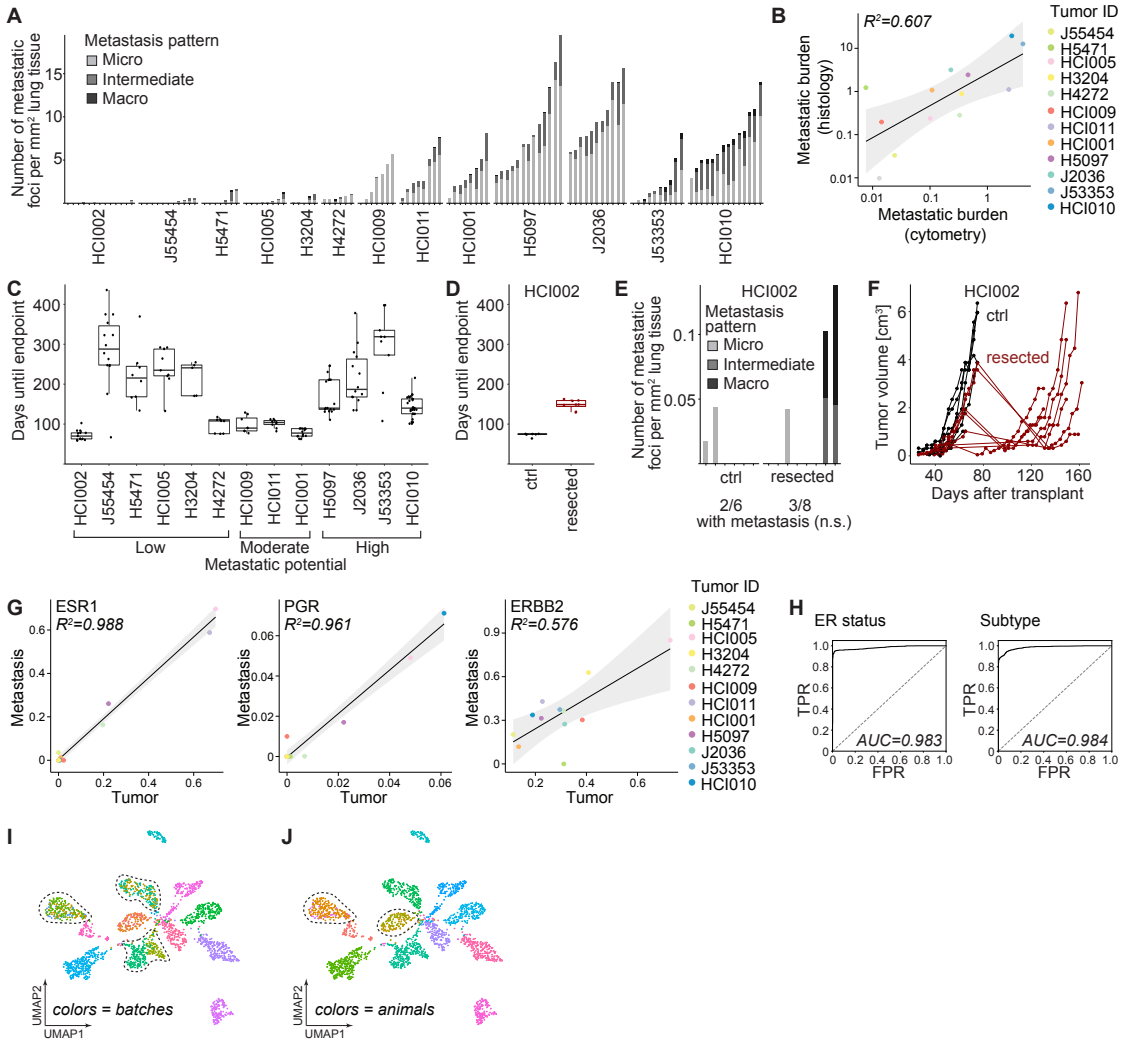


Figure S1. Biological characteristics of the PDX models.

(**A**) Bar chart shows the number of metastatic foci per mm^2 lung tissue area for individual animals ordered by the metastatic potential of the tumor models determined by histology. Each tick mark represents one animal. The size of metastatic foci is colored in shades of gray (micrometastasis < 10 cells, intermediate 10–100 cells and macrometastasis > 100 cells). (**B**) Scatter plot shows the correlation of mean metastatic burden assessed by histology (proportion of metastatic tissue area to total lung tissue area) and flow cytometry (proportion of metastatic cells to total live cells) colored by individual tumor models. Linear regression with 95 % confidence intervals and Pearson correlation coefficient are shown. (**C**) Boxplot shows median days until endpoint (2.5 cm diameter of primary tumor) after tumor transplantation per tumor model ordered by the metastatic potential as determined in Figure 1B. (**D**) Boxplot shows median days until endpoint (2.5 cm diameter of the primary tumor or recurrent tumor) after tumor transplantation comparing HCI002 (black) and after HCI002 resection (red). (**E**) Bar chart shows the number of metastatic foci per mm^2 lung tissue area for individual animals of HCI002 and resected HCI002 at endpoint. The size of metastatic foci is colored in shades of gray (micrometastasis < 10 cells, intermediate 10–100 cells and macrometastasis > 100 cells). Each tick mark represents one animal showing 2/6 (control) and 3/8 animals (resected) that developed metastases (p-value=0.872, Chi-Square test). (**F**) Spider plot shows tumor growth (volume in cm^3) for each animal transplanted with HCI002 (black) or resected with subsequent recurrent tumor (red). (**G**) Scatterplots show the correlation of the mean expression of the indicated receptors in primary tumor and metastatic cells colored by individual tumor models. Linear regressions with 95 % confidence intervals and Pearson correlation coefficients are shown. (**H**) ROC curves with the corresponding area under the curve (AUC) show PC1 categorized based on ER status (left) and BC subtype (right). (**I**) UMAP projection of single-cell transcriptomes color-coded by batch (individual plates). Dashed lines highlight clusters of cells from the same tumor model measured in multiple batches (technical replicates). (**J**) UMAP projection of single-cell transcriptomes color-coded by individual animals. Dashed lines highlight clusters of cells from the same tumor model retrieved from multiple animals (biological replicates).

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

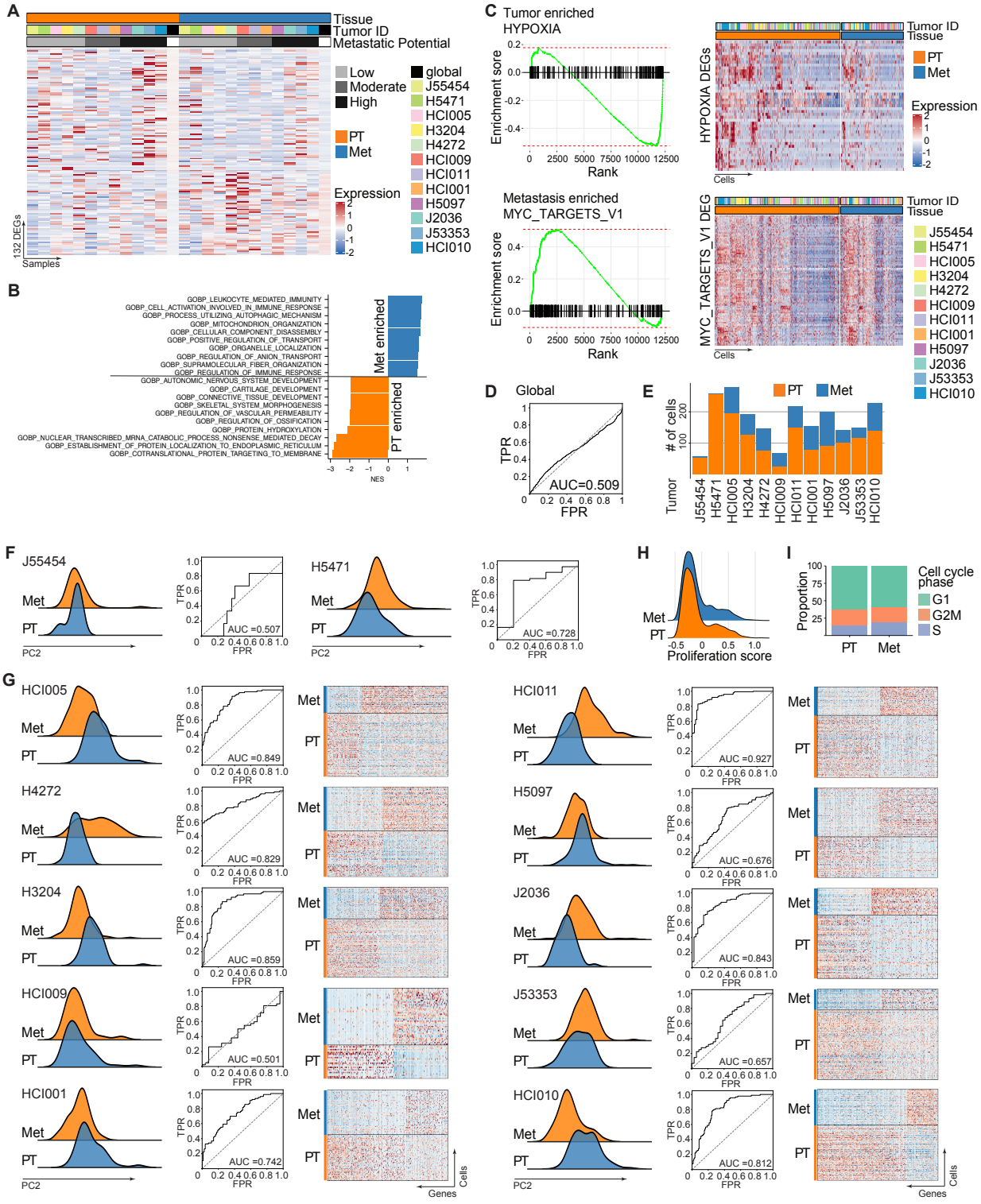
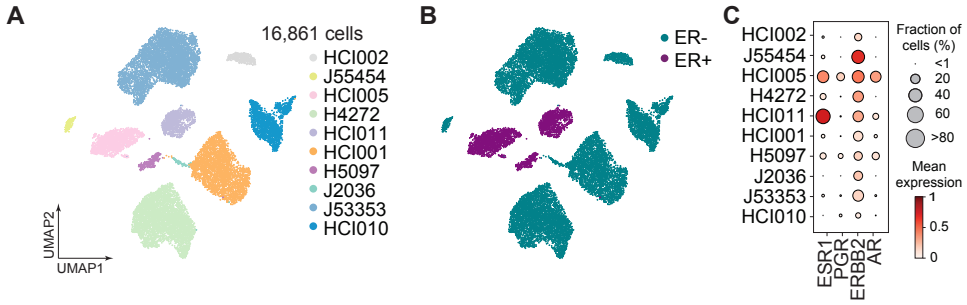


Figure S2. Differential gene expression between primary tumor and matched metastatic cells.
(**A**) Heatmap shows mean expression per tumor model of DEGs between primary tumors and metastases. Annotations indicate tissue, tumor model, and metastatic potential. (**B**) Bar chart shows pathways enriched in primary tumors (negative NES, orange) and metastases (positive NES, blue) using GO biological pathways from MSigDB. (**C**) Enrichment plots show hypoxia as the top enriched pathway in primary tumors (top) and MYC targets as the top enriched pathway in metastases (bottom). Heatmaps show expression in single cells of DEGs associated with either hypoxia (top) or MYC targets (bottom). Annotations show tissue and tumor model. (**D**) ROC curve using PC2 coordinates to classify cells into either primary tumor or metastatic cells of all tumors grouped together (global) with depicted AUC. (**E**) Bar chart shows the number of primary tumor (orange) and metastatic cells (blue) for each tumor model. (**F**) Ridge plots show normalized cell counts along PC2 color-coded by primary tumor and metastasis for tumor models J55454 and H5471 without a sufficient number of metastatic cells and corresponding ROC curves (same as in (D)). (**G**) Ridge plots show normalized cell counts along PC2 color-coded by primary tumor and metastasis for individual tumor models with a sufficient number of metastatic cells and corresponding ROC curves of PC2 (same as in (D)). Clear separations in PC2 are reflected by AUC > 0.7 by ROC curve analysis. Heatmaps show the expression of DEGs between primary tumor and metastatic cells. (**H**) Ridge plot shows proliferation score for primary tumors and metastases. (**I**) Bar chart shows the proportion of cells in G1/G2M/S cell cycle phase for primary tumors and metastases. Not significant using Wilcoxon rank test.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



D Pathways enriched in low metastatic tumors

Pathway (HALLMARK)	p-value	FDR q-value	Genes
TNFA_SIGNALING_VIA_NFKB	6.17E-11	3.09E-09	FOS, CCND1, AREG, ZFP36, TAP1, SOD2, NFKBIA, ATF3, SQSTM1, DUSP1
ESTROGEN_RESPONSE_LATE	1.48E-09	2.46E-08	FOS, CCND1, AREG, ZFP36, KRT19, SLC9A3R1, BLVRB, PLAAT3, CD9
INTERFERON_GAMMA_RESPONSE	1.48E-09	2.46E-08	B2M, UBE2L6, PSMB8, IFI35, HLA-A, HLA-B, TAP1, SOD2, NFKBIA
INTERFERON_ALPHA_RESPONSE	4.00E-09	5.00E-08	B2M, UBE2L6, PSMB8, IFI35, CD47, HLA-C, TAP1
ESTROGEN_RESPONSE_EARLY	3.13E-08	3.13E-07	FOS, CCND1, AREG, KRT19, SLC9A3R1, BLVRB, PLAAT3, ELF3

Pathway (GOBP)	p-value	FDR q-value	Genes
REGULATION_OF_CELL_POPULATION_PROLIFERATION	2.63E-12	1.67E-08	HLA-E, HLA-A, B2M, ZFP36, NFKBIA, SOD2, ROMO1, LGALS3, CD47, IFI35, CDK6, DUSP1, ATF3, SLC9A3R1, CD9, NUPR1, MALAT1, CCND1, S100A11, PLAAT4, AREG, FTH1, RARRES1 CAPN1
ANTIGEN_PROCESSING_AND_PRESENTATION_OF_ENDOGENOUS_PEPTIDE_ANTIGEN	4.36E-12	1.67E-08	HLA-E, HLA-A, B2M, HLA-B, HLA-C, TAP1
ANTIGEN_PROCESSING_AND_PRESENTATION_OF_ENDOGENOUS_ANTIGEN	3.65E-11	8.66E-08	HLA-E, HLA-A, B2M, HLA-B, HLA-C, TAP1
BIOLOGICAL_PROCESS_INVOLVED_IN_INTERSPECIES_INTERACTION_BETWEEN_ORGA	4.52E-11	8.66E-08	HLA-E, HLA-A, B2M, ZFP36, NFKBIA, SOD2, ROMO1, LGALS3, CD47, IFI35, CDK6, HLA-B, HLA-C, OPTN, FOS, S100A14, WFDC2, PLAAT3, OCIAD2, C15orf48, IFI6, CHMP2A
ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN_VIA_MHC_CLAS	7.48E-11	1.15E-07	HLA-E, HLA-A, B2M, HLA-B, HLA-C, TAP1

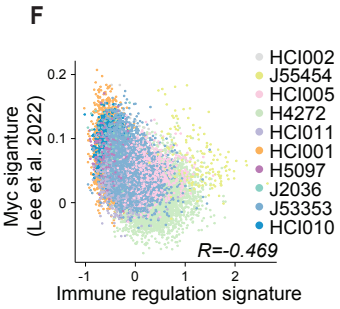
Pathway (Reactome)	p-value	FDR q-value	Genes
INNATE_IMMUNE_SYSTEM	1.66E-16	2.68E-13	HLA-A, HLA-B, HLA-C, B2M, LAMTOR2, LAMTOR1, LAMP2, LGALS3, DSP, CAPN1, FTH1, CD59, CD47, MGST1, S100A11, CREG1, HLA-E, PSMB8, UBC, PSMB3, FOS, UBE2L6, NFKBIA, ATP6V0E1
NEUTROPHIL_DEGRANULATION	4.01E-14	3.24E-11	HLA-A, HLA-B, HLA-C, B2M, LAMTOR2, LAMTOR1, LAMP2, LGALS3, DSP, CAPN1, FTH1, CD59, CD47, MGST1, S100A11, CREG1
CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	1.54E-12	8.31E-10	HLA-A, HLA-B, HLA-C, B2M, HLA-E, PSMB8, UBC, PSMB3, FOS, UBE2L6, NFKBIA, SOD2, IFI6, IFI35, SQSTM1, CCND1, IL32
ANTIGEN_PROCESSING_CROSS_PRESENTATION	4.97E-12	2.01E-09	HLA-A, HLA-B, HLA-C, B2M, HLA-E, PSMB8, UBC, PSMB3, TAP1
ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_	2.81E-11	8.88E-09	HLA-A, HLA-B, HLA-C, B2M, HLA-E, TAP1

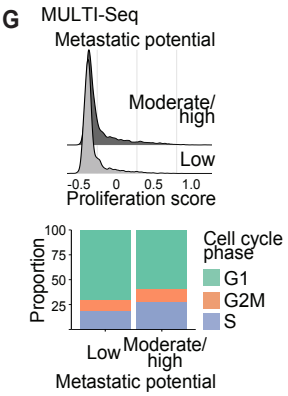
E Pathways enriched in high metastatic tumors

Pathway (HALLMARK)	p-value	FDR q-value	Genes
OXIDATIVE_PHOSPHORYLATION	5.57E-11	2.78E-09	CYC1, MDH2, GPX4, ECHS1, MRPL15, NDUFB7, UQCRC1, LDHB, IDH2, UQCRFS1
MYC_TARGETS_V1	1.35E-09	3.37E-08	CYC1, MDH2, RACK1, PRDX4, SRM, FBL, RPS2, RPS3, RPS6, RPL14
ADIPOGENESIS	5.43E-07	6.79E-06	CYC1, MDH2, GPX4, ECHS1, MRPL15, NDUFB7, UQCRC1
EPITHELIAL_MESENCHYMAL_TRANSITION	5.43E-07	6.79E-06	TPM2, BGN, GJA1, EMP3, CTHRC1, PCOLCE, VIM
MYOGENESIS	8.79E-06	7.33E-05	TPM2, SPHK1, CLU, CFD, APOD, IGFBP7

Pathway (Reactome)	p-value	FDR q-value	Genes
TRANSLATION	3.95E-14	2.87E-10	RPS2, RPS6, RPS3, RPS3A, RPS20, RPL7, RPL13, RPL32, RPL37, RPL14, RPN2, TRAM1, EIF3E, EIF3H, EEF1D, MRPL15
SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	7.50E-14	2.87E-10	RPS2, RPS6, RPS3, RPS3A, RPS20, RPL7, RPL13, RPL32, RPL37, RPL14, RPN2, TRAM1
EUKARYOTIC_TRANSLATION_INITIATION	5.16E-13	1.32E-09	RPS2, RPS6, RPS3, RPS3A, RPS20, RPL7, RPL13, RPL32, RPL37, RPL14, EIF3E, EIF3H
EUKARYOTIC_TRANSLATION_ELONGATION	2.38E-11	3.95E-08	RPS2, RPS6, RPS3, RPS3A, RPS20, RPL7, RPL13, RPL32, RPL37, RPL14, EEF1D
CELLULAR_RESPONSES_TO_STIMULI	2.58E-11	3.95E-08	RPS2, RPS6, RPS3, RPS3A, RPS20, RPL7, RPL13, RPL32, RPL37, RPL14, TUBA1B, CYBA, HMGA1, CDKN2A, IGFBP7, UBE2S, PRDX2, BLVRB, MYDGF,

Pathway (GOBP)	p-value	FDR q-value	Genes
PEPTIDE_METABOLIC_PROCESS	3.95E-14	2.87E-10	PKM, RPS3, RPL37, RPS20, RPS6, RPS3A, RPS2, RPL7, RPL14, RPL13, RPL32, RACK1, VIM, EEF1D, EIF3E, EIF3H, CLU, GPX4, TMED10, ETHE1
CYTOPLASMIC_TRANSLATION	7.50E-14	2.87E-10	PKM, RPS3, RPL37, RPS20, RPS6, RPS3A, RPS2, RPL7, RPL14, RPL13, RPL32
CELLULAR_AMIDE_METABOLIC_PROCESS	5.16E-13	1.32E-09	PKM, RPS3, RPL37, RPS20, RPS6, RPS3A, RPS2, RPL7, RPL14, RPL13, RPL32, RACK1, VIM, EEF1D, EIF3E, EIF3H, CLU, GPX4, TMED10, ETHE1, SPHK1
PEPTIDE_BIOSYNTHETIC_PROCESS	2.38E-11	3.95E-08	PKM, RPS3, RPL37, RPS20, RPS6, RPS3A, RPS2, RPL7, RPL14, RPL13, RPL32, RACK1, VIM, EEF1D, EIF3E, EIF3H
AMIDE_BIOSYNTHETIC_PROCESS	2.58E-11	3.95E-08	PKM, RPS3, RPL37, RPS20, RPS6, RPS3A, RPS2, RPL7, RPL14, RPL13, RPL32, RACK1, VIM, EEF1D, EIF3E, EIF3H





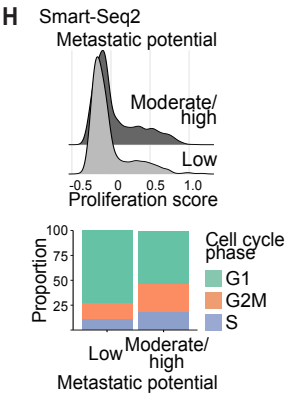


Figure S3. Characteristics of the metastatic signatures.

(**A**) UMAP projection of single-cell transcriptomes color-coded by individual tumor models. (**B**) UMAP projection of single-cell transcriptomes color-coded by ER status. (**C**) Bubble plot shows the expression of receptors per tumor model. The size of dots indicates the fraction of cells expressing and the red color indicates gene expression. (**D**) Pathway enrichment of DEGs shared between poorly metastatic tumors. (**E**) Pathway enrichment of DEGs shared between highly metastatic tumors. (**F**) Scatterplot shows the correlation of MYC (42) and immune regulation signature expression colored by tumor model. Pearson correlation coefficient is shown. (**G**) Ridge plot shows proliferation score for primary tumor of low and moderate/high metastatic potential. The bar chart shows proportion of cells in different cell cycle phases for primary tumors of low and moderate/high metastatic potential. Showing MULTI-Seq dataset. Proportion changes are not significant using Wilcoxon rank test. (**H**) Same as in (G) for the Smart-Seq2 dataset. Not significant using Wilcoxon rank test.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

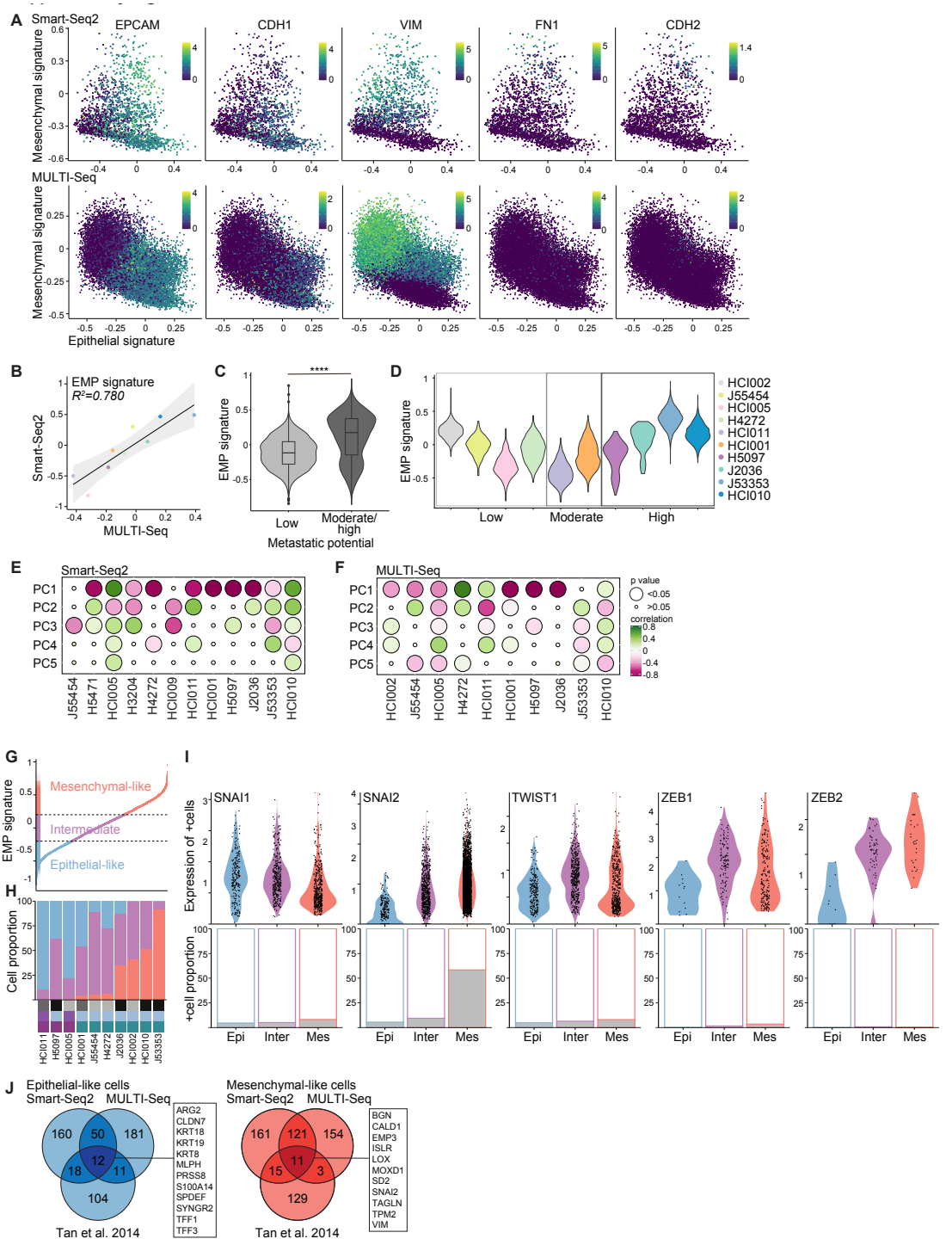


Figure S4. EMP is a key feature of tumor heterogeneity.
(**A**) Scatter plots show mesenchymal against epithelial signatures for individual cells colored by the expression of indicated epithelial (EPCAM, CDH1) and mesenchymal markers (VIM, FN1, CDH2). The upper panels show Smart-Seq2 and the lower panels show MULTI-Seq datasets. The color scale indicates the magnitude of gene expression. (**B**) Scatter plot shows the correlation of the mean EMP signature expression per tumor model between Smart-Seq2 and MULTI-Seq datasets. Linear regression with 95 % confidence intervals and Pearson correlation coefficient are shown. (**C**) Violin plot shows EMP signature expression of tumor models with low and intermediate/high metastatic potential using the MULTI-Seq dataset. Boxplot showing median, significance p<0.001 by Wilcox test. (**D**) Violin plot shows EMP signature expression per tumor model ordered by metastatic potential using the MULTI-Seq dataset. (**E**) Bubble plot shows the correlation of EMP signature with PCs 1-5 using Smart-Seq2 dataset. The color indicates positive (green) or negative (purple) correlation coefficient, larger circle indicates significant p-value < 0.05, small circle indicates no significant p-value > 0.05. (**F**) same as in (E) for the MULTI-Seq dataset. (**G**) Cells ranked by EMP signature defining three cell states: epithelial-like (blue), intermediate EMP (purple) and mesenchymal-like cells (red) using the MULTI-Seq dataset. (**H**) Bar chart shows the proportion of the three different EMP cell states per tumor model ranked by the increasing proportion of mesenchymal-like cells. Gray-scale boxes indicate the metastatic potential. Other annotations indicate ER status and BC subtype as in Figure 4F. Showing MULTI-Seq dataset. (**I**) Violin plots (top) show expression of EMT-associated TFs in expressing cells grouped by EMP cell states (Epi = epithelial-like, Inter = Intermediate EMP, Mes = mesenchymal-like cells). Bar charts (bottom) show the fraction of expressing cells in gray. Showing MULTI-Seq dataset. (**J**) Venn diagrams show overlaps of epithelial (blue, left panel) and mesenchymal markers (red, right panel) for Smart-Seq2, MULTI-Seq and Tan et al. 2014. Highlighted are genes shared between all three sets.

bioRxiv preprint doi: https://doi.org/10.1101/2022.08.04.502697; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

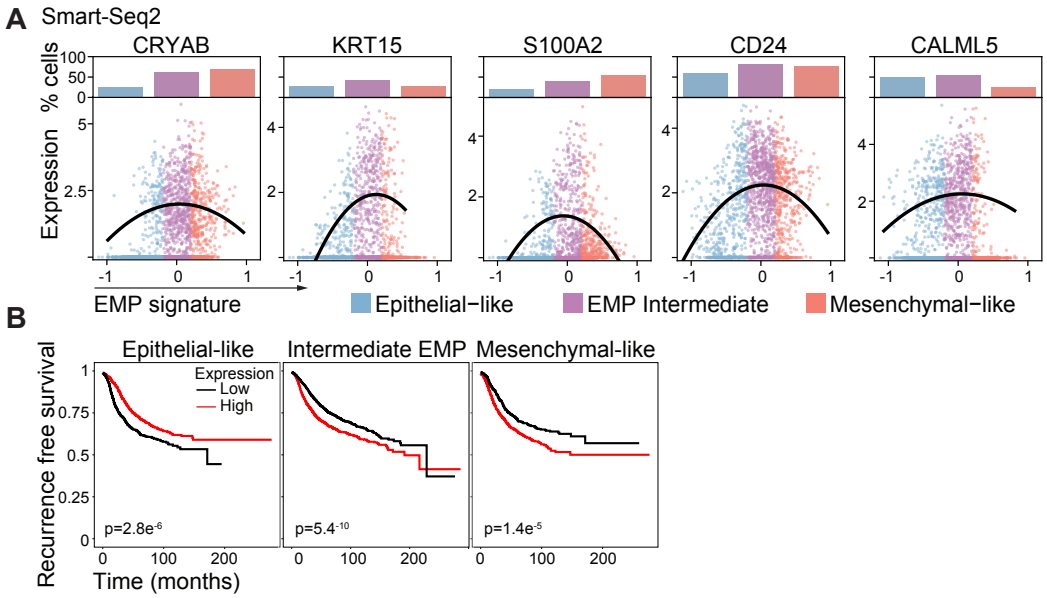


Figure S5. Intermediate EMP cell markers were correlated with patient outcome.
(**A**) Scatter plots show the expression of indicated genes ordered by increasing EMP signature expression. Dots show expression for individual cells, lines show smoothed expression of expressing cells. Bar charts on top show the proportion of positive expressing cells for the three EMP cell states (blue=epithelial-like, purple=intermediate EMP, red=mesenchymal-like cells). Showing the Smart-Seq2 dataset. (**B**) Recurrence-free survival of BC patients using the mean expression of the overlapped genes for each EMP cell state (generated with KM-plotter (35)).