PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Deconvolution of cancer cell states by the XDec-SM method

Oscar D. Murillo[1,2☯], **Varduhi Petrosyan**[2☯], **Emily L. LaPlante**[2], **Lacey E. Dobrolecki**[3], **Michael T. Lewis**[3,4,5], **Aleksandar Milosavljevic**[2,4*]

1 Translational Science and Therapeutics Division, Fred Hutchinson Cancer Center, Seattle, Washington, United States of America, **2** Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America, **3** Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas, United States of America, **4** Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America, **5** Departments of Molecular and Cellular Biology and Radiology, Baylor College of Medicine, Houston, Texas, United States of America

☯ These authors contributed equally to this work.
* amilosav@bcm.edu

Citation: Murillo OD, Petrosyan V, LaPlante EL, Dobrolecki LE, Lewis MT, Milosavljevic A (2023) Deconvolution of cancer cell states by the XDec-SM method. PLoS Comput Biol 19(8): e1011365. https://doi.org/10.1371/journal.pcbi.1011365

Editor: Zhaolei Zhang, University of Toronto, CANADA

Received: January 6, 2023

Accepted: July 17, 2023

Published: August 14, 2023

Data Availability Statement: To empower the community to use this method, we made the XDec-SM code available online and as a R package under a free open-source license. All analysis code and source data is available online (https://github.com/BRL-BCM/XDec_SM). We also developed an interactive web service that allows users to deconvolute breast tumor RNA-seq profiles of interest and project them onto the cancer cell state map (https://brl-bcm.shinyapps.io/XDec_BRCA/).

Funding: his work was supported by a grant from the Common Fund of the National Institutes of

Abstract

Proper characterization of cancer cell states within the tumor microenvironment is a key to accurately identifying matching experimental models and the development of precision therapies. To reconstruct this information from bulk RNA-seq profiles, we developed the XDec Simplex Mapping (XDec-SM) reference-optional deconvolution method that maps tumors and the states of constituent cells onto a biologically interpretable low-dimensional space. The method identifies gene sets informative for deconvolution from relevant single-cell profiling data when such profiles are available. When applied to breast tumors in The Cancer Genome Atlas (TCGA), XDec-SM infers the identity of constituent cell types and their proportions. XDec-SM also infers cancer cells states within individual tumors that associate with DNA methylation patterns, driver somatic mutations, pathway activation and metabolic coupling between stromal and breast cancer cells. By projecting tumors, cancer cell lines, and PDX models onto the same map, we identify *in vitro* and *in vivo* models with matching cancer cell states. Map position is also predictive of therapy response, thus opening the prospects for precision therapy informed by experiments in model systems matched to tumors *in vivo* by cancer cell state.

Author summary

Complex tumor tissue may be characterized by the cellular composition and cell states of constituent cell types. The identification of cancer cell states within the tumor microenvironment is key to understanding tumor biology and an important step toward the development of precision therapies. Computational deconvolution of tumors may be applied in conjunction with physical separation methods such as single-cell RNA sequencing to map heterotypic interactions occurring within the tumor. Although these sequencing protocols allow for the transcriptional profiling of cells, these technologies present technical challenges, variabilities, increase cost and limit throughput. These challenges are addressed by

Health (NIH) (5U54 DA036134) (to A.M.), an NCI PDX Development and Trials Center grant U54CA224076 (to M.T.L), a CPRIT Core Facility Support Grant RP170691 (to M.T.L.), P30 Cancer Center Support Grant (NCI-CA125123), and Henry and Emma Meyer Chair in Molecular and Human Genetics (to A.M.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: O.D.M. is currently an employee of Illumina, Inc. M.T.L is a founder of and limited partner in StemMed Ltd., and a manager in StemMed Holdings L.L.C., its general partner, and is a founder of and equity stake holder in, Tvardi Therapeutics Inc. L.E.D. is a compensated employee of StemMed, Ltd. The remaining authors declare no competing interests.

the XDec-SM deconvolution method that provides insights into interactions between cancer cell states and the surrounding tumor microenvironment. Mapping reveals metabolic communication between stromal and cancer cells and elucidates other aspects of tumor biology including driver mutations, pathway activation, and methylation status. Map position is also predictive of therapy response, thus opening the prospects for precision therapy informed by experiments in model systems matched to tumors by cancer cell state. Taken together, this general approach bridges the cellular and tissue layers of biology in an innovative, conceptually simple way that can be applied to cancer and non-cancer tissues alike.

Introduction

Molecular profiling of breast tumors over the past two decades has further reinforced the understanding of breast cancer as a highly heterogeneous disease with a staggering complexity of molecular aberrations at both the tissue and cellular levels. Large data banks such as The Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium (ICGC) [2] characterize a vast array of tumors. However, the multi-omic characterization of these tumors is confounded by the bulk profiling analyses of heterogeneous tumor tissue. The averaging inherent in bulk profiling of tumor samples that consists of variable proportions of similar cell types precludes access to the state of cancer cells within tumors [3,4], which is essential for understanding effects of somatic mutations, identifying interactions between cancer cells and the tumor microenvironment and for predicting therapy response.

Single-cell profiling provides an alternative direct method to access to the states of constituent cell typess within individual tumors. Among the single-cell omic methods, the most widely used is single-cell RNA sequencing (scRNA-seq), [5–8]. While in principle providing the ultimate level of resolution, the scRNA-seq method has limitations: the depth of coverage is relatively sparse precluding precise quantitation; it is biased towards highly abundant cell types; the cost per sample is high [9–12]; and it is not readily applicable to the formalin-fixed paraffin-embedded (FFPE) samples which are routinely collected in practice. Computational deconvolution is highly synergistic with single-cell profiling, as it decreases cost and technical requirements dramatically, while benefiting from the information about the diversity of cell types gathered by scRNA-seq. Methods that combine deconvolution and single-cell profiling such as CIBERSORTx [13] and MuSiC [14] have demonstrated the synergy of the two approaches. However, these computational deconvolution methods are reference-based in that they explain bulk profiles as linear combinations of *a priori* defined profiles of constituent cell types that have previously been physically isolated, thus precluding data-driven discovery of recurrent states of cancer cells and their comparison to established tumor classification (e.g., Luminal, Basal, and HER2 breast cancer subtypes). Although methods such as EcoTyper [15] can identify cell states, these methods still rely on cell type references. One key limitation of the reference-based approach is that the physical separation that is needed for obtaining the references perturbs the constituent cell states, thus calling for deconvolution strategies that can infer cell states (beyond observable in single-cell references) from bulk profiling data.

Because of these current limitations, no deconvolution method has been shown to estimate cell states that correspond to subtypes of breast cancer or other solid tumors in the context of the complete tumor microenvironment. To address this knowledge gap, we developed XDec-SM, a new methodological framework and algorithm based on simplex mapping that combines bulk RNA-seq profiles of individual tumors and publicly accessible relevant scRNA-seq

information in a way that enables discovery of new recurrent states of cancer cells. The XDec-SM method is an extension of the EDec [16] deconvolution method that allows for the identification of unique cancer and stromal cell states. Unlike EDec which requires matched DNA methylation and gene expression profiling data, XDec-SM only requires bulk RNA-seq as an input and estimates the proportion and cell type specific expression of distinct cell types and states. XDec-SM is reference-optional as it may leverage information from references when available (e.g., scRNA-seq profiles). XDec-SM only uses references for the purpose of identifying informative loci (e.g., genes) for deconvolution and does not utilize quantitative information from references for any other purpose. Additionally, information about informative loci from an external source other than the references may also be used in conjunction with references identified in Stage 0 (e.g., literature-based signature). Moreover, XDec-SM predecessor EDec has previously been shown to compare well to other DNA methylation-based tumor deconvolution methods in reference-free mode [17].

Here we apply XDec-SM to breast tumors in the TCGA collection to construct the first simplex map of breast cancer, defined by the cellular makeup of individual tumor samples and the state of constituent cell fractions. Because the method is not biased by reference profiles, it produces data-driven estimates of constituent cell types and distinct cell states using both single cell information and previously published PAM50 subtype loci [2,16,18]. We determine associations between the cancer cell state, DNA methylation patterns, somatic driver mutations, activation of cancer promoting pathways and heterotypic interactions within the tumor microenvironment. The map enables precise modeling of tumors by placing the deconvoluted *in vivo* cancer cells, cancer cell lines, and PDX models within the same low-dimensional coordinate system and enables the use of experimentally tractable models to predict tumor-specific response to therapy. To empower future studies, we deploy XDec-SM as a R package. We also make it accessible online via a web user interface for placing individual breast tumors, PDX, and cell line models on the map based on their bulk RNA-seq profiles.

Results

XDec-SM deconvolution allows for the mapping of BRCA tumors onto a cancer cell state map

The XDec-SM deconvolution method (Methods) was applied to deconvolute the bulk RNA-seq breast cancer tumor profiles from the TCGA collection by leveraging publicly available single cell sequencing data [6] (Fig 1A). We first attempted to deconvolute the TCGA collection using just the PAM50 gene set. However, only three stable cell profiles were identified with this deconvolution. Moreover, each of these cell profiles is best correlated to an epithelial cell type and no other cell types were identified. This is not an unexpected result as the PAM50 genes have been developed to differentiate between breast cancer subtypes [18] and not between different cell types in the breast cancer microenvironment. Therefore, deconvoluting over just this set of genes does not allow for the identification of additional cancer cell profiles and non-cancer cell types. To improve our deconvolution, we also identified informative genes that were differentially expressed among distinct cell types in the tumor microenvironment. We then identified an informative gene set (n = 323) which was defined as the union of genes differentially expressed in the scRNA-seq data in different breast tumor cell types and the PAM50 genes (S1A Fig). The deconvolution algorithm identified nine distinct cell types within the tumors (S1B and S1C Fig) explaining 85.2% of the variance. Five distinct epithelial cell types, a Cancer Associated Fibroblast (CAF) as well as CAF adipocyte, macrophage, and a T-cell type were identified.

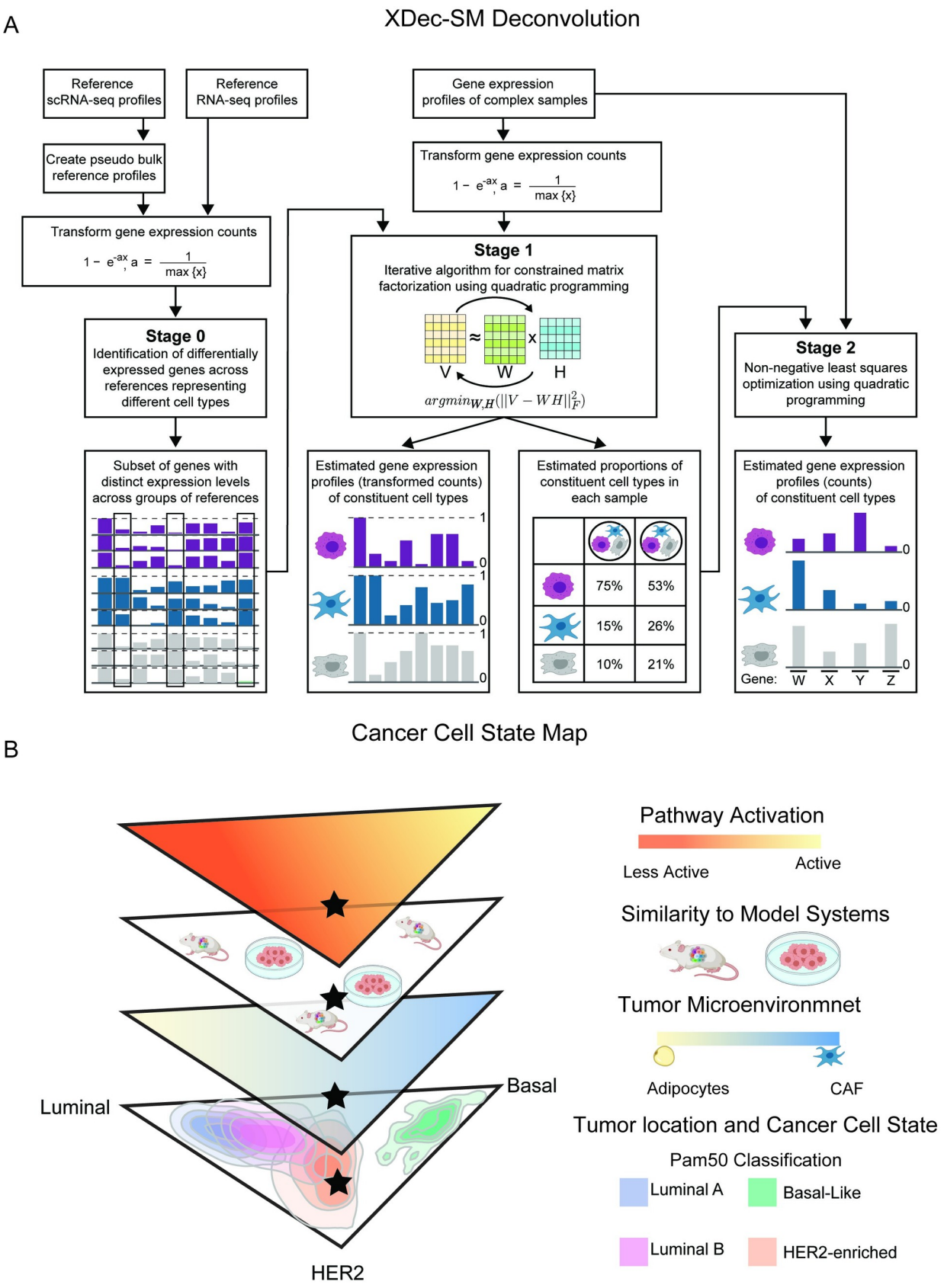
A
XDec-SM Deconvolution
Reference scRNA-seq profiles
Reference RNA-seq profiles
Create pseudo bulk reference profiles
Transform gene expression counts
Gene expression profiles of complex samples
Transform gene expression counts
Stage 0
Identification of differentially expressed genes across references representing different cell types
Stage 1
Iterative algorithm for constrained matrix factorization using quadratic programming
V
W
H
Stage 2
Non-negative least squares optimization using quadratic programming
Subset of genes with distinct expression levels across groups of references
Estimated gene expression profiles (transformed counts) of constituent cell types
Estimated proportions of constituent cell types in each sample
75%
53%
15%
26%
10%
21%
Estimated gene expression profiles (counts) of constituent cell types
Gene: W X Y Z
B
Cancer Cell State Map
Luminal
Basal
HER2
Pathway Activation
Less Active
Active
Similarity to Model Systems
Tumor Microenvironmnet
Adipocytes
CAF
Tumor location and Cancer Cell State
Pam50 Classification
Luminal A
Basal-Like
Luminal B
HER2-enriched

Fig 1. Cancer cell state map built with XDec-SM deconvolution. (A) The XDec-SM algorithm contains three stages (Stage 0, 1, 2). Stage 0 is generalized to utilize a set of reference RNA-seq or scRNA-seq profiles. Reference scRNA-seq profiles with user provided cell-type classifications are utilized to create pseudo bulk reference profiles by summing every five profiles ordered by total gene coverage. The reference RNA-seq profiles or pseudo bulk profiles are transformed to a 0–1 range to identify differentially expressed genes representing different cell types. The subset of informative genes with distinct expression levels across the groups of references are utilized in Stage 1. In Stage 1, using the gene expression profiles of complex samples transformed to 0–1, XDec-SM estimates the gene expression profiles (transformed counts) of constituent cell types and the proportions of constituent cell types in each sample. XDec-SM uses an iterative algorithm for constrained matrix factorization using quadratic programming. The estimated proportions of constituent cell types in each sample are used in Stage 2. In Stage 2, using the untransformed gene expression profiles of the same complex samples and the proportions estimated in Stage 1, XDec-SM estimates the gene expression profiles (counts) of constituent cell types. Stage 2 uses non-negative least squares optimization using quadratic programming. (B) Cancer cell state map of breast cancer. The cancer cell state map of breast cancer acts as a framework to determine a tumor's cancer cell state (layer 1 bottom). The map is also indicative of a tumor's microenvironment (layer 2), similarity to model organisms including murine models and cell lines (layer 3), and the activation of cancer associated pathways (layer 4 top). Figure components created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1011365.g001

Unlike other reference-based methods, XDec-SM utilized a reference-optional, data driven approach to define constituent cell types and states within tumors and is not constrained by the availability of references profiles. In an improvement over the EDec [16] algorithm, both a normal adipocyte stromal profile and a CAF stromal profile were identified (S1D Fig). Of the five epithelial profiles identified by XDec-SM, three were identified as Luminal, Basal, and HER2 cancer epithelial profiles by observing their high abundance in TCGA tumors with respective Luminal, Basal, and HER2 subtype classification (S1D Fig and Methods). XDec-SM identified a common Luminal epithelial profile for TCGA tumors that were classified as Luminal A and Luminal B. Moreover, an additional epithelial profile was identified that was classified as a normal epithelial profile as it was highly abundant in control breast tissue samples. To further validate these designations of deconvoluted profiles, Stage 2 XDec-SM deconvolution was performed, and the cell type specific expression of marker genes for the three subtypes was confirmed (S2 Fig and Methods).

While the majority of cancer cells within a specific TCGA tumor was typically of the same type (e.g., Luminal) as its PAM50 classification based on its bulk profile, other subtypes (e.g., Basal, HER2) were also detected albeit in smaller proportions. As we discuss in detail below, this is only partially due to the presence of cancer cells of different type; the fact that the cancer cells can be modeled as a linear combination of pure subtypes also reflects intermediate cancer cell states that resemble to various degrees (as indicated by relative proportion estimates) the three subtypes. In other words, the "relative proportions" of the three subtypes also reflects relative similarities to the "pure" deconvoluted cancer cell states of the three subtypes. Mindful that the deconvoluted proportions of the three subtypes may to a significant degree reflect relative similarity, we placed the cancer cell fractions of each TCGA tumor within a two-dimensional simplex map (Fig 1B). This three-dimensional simplex map acts as a framework onto which breast cancer tumors can be projected, and the position of a tumor in the cancer cell state is associated with the tumor microenvironment, similarity to model organism, and other tumor biology such as the activation of tumor promoting pathways (Fig 1B).

XDec-SM compares favorably to other deconvolution algorithms

After deconvoluting TCGA breast cancer samples, we next compared XDec-SM against other widely used computational deconvolution methods (Methods). To determine how XDec-SM performed in comparison with other widely used deconvolution methods, we deconvoluted breast cancer samples from TCGA with XDec-SM as well as CIBERSORTX [13], Kassandra [19], EPIC [20], and xCell [21] (Fig 2). Each of these deconvolution tools identifies a different number and subset of cell types when deconvoluting the TCGA breast cancer collection. XDec-SM identified nine data-driven cell types including five epithelial cell types, a CAF cell type, and adipocyte cell type, a macrophage cell type and a T-cell cell type. Other

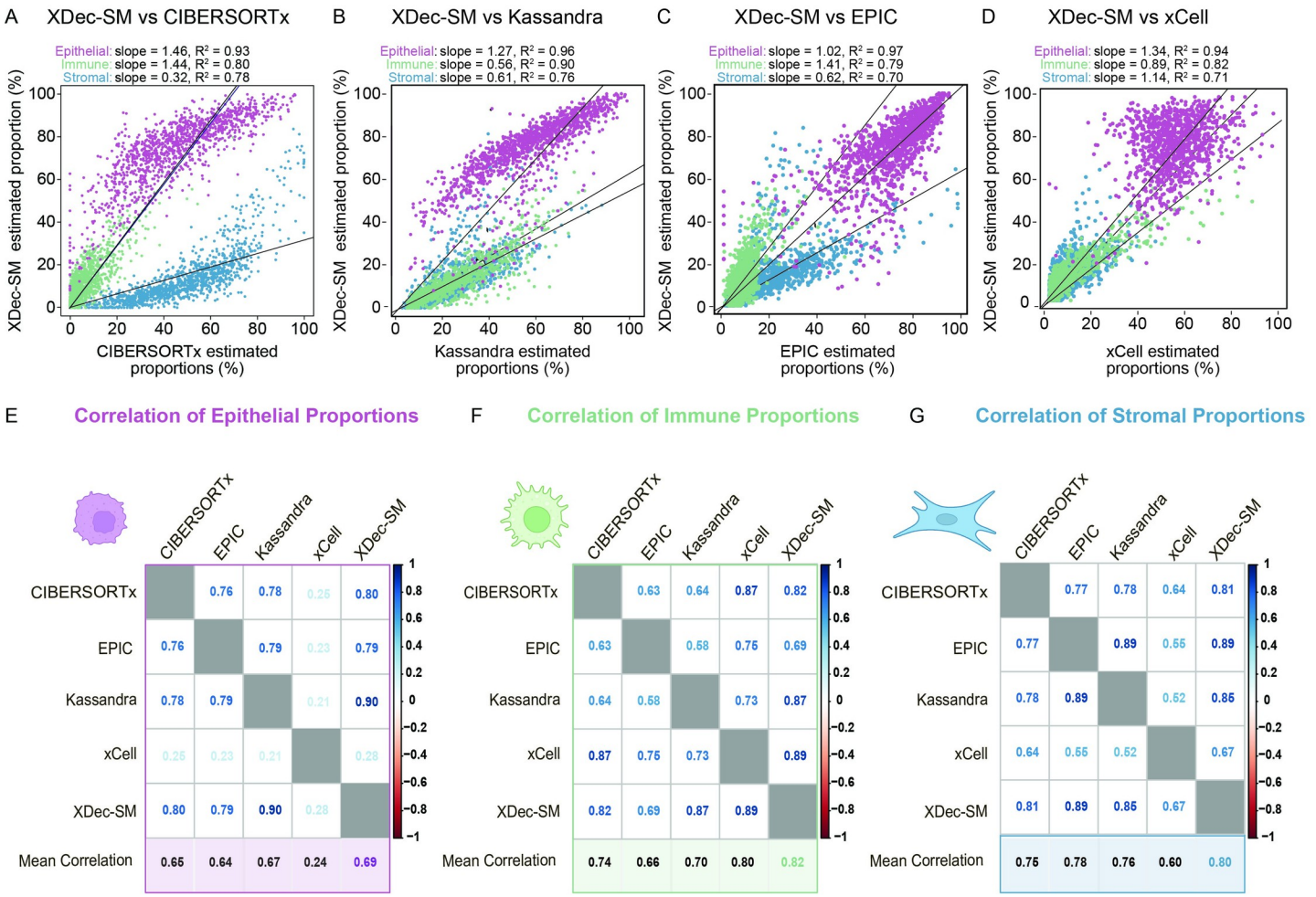

Fig 2. Comparison of XDec-SM estimated proportions to other deconvolution methods. (A) Scatterplot between CIBERSORTx (x-axis) and XDec-SM (y-axis) estimated per-sample proportions of each constituent cell type. For XDec-SM, the five epithelial profiles are summed to represent epithelial, the two stromal profiles are summed to represent stromal, and the T cell and Macrophage are summed to represent immune. For CIBERSORTx, the three immune (B cell, T cell, macrophage) profiles are summed to represent immune and the two stromal (stroma, endothelial) profiles are summed to represent stromal. (B) Scatterplot comparing the estimated per-sample proportions between Kassandra (x-axis) and XDec-SM (y-axis). For XDec-SM the general cell type proportions were summarized as noted above. For Kassandra, the summed immune stromal and cancer proportions were obtained from https://science.bostongene.com/kassandra/. (C) Scatterplot comparing the proportions estimated with EPIC (x-axis) and XDec-SM (y-axis). For EPIC, the following cell types were summed for the immune compartment: B-cells, T-cells (CD4 and CD8), macrophages, and NK cells. CAFs and endothelial cells were summed for the stromal compartment, and the uncharacterized cell type was categorized as epithelial. (D) Scatterplot comparing the proportions estimated by xCell (x-axis) and XDec-SM (y-axis). Finally, for xCell, the proportions were estimated with TIMER. (E) Correlation of the epithelial proportions across all CIBERSORTx, EPIC, Kassandra, xCell, and XDec-SM. The color of the correlation value in each cell is representative of the strength of the correlation (scale on the y-axis). (F) Correlation of the immune proportions across all CIBERSORTx, EPIC, Kassandra, xCell, and XDec-SM. The color of the correlation value in each cell is representative of the strength of the correlation (scale on the y-axis). (G) Correlation of the stromal proportions across all CIBERSORTx, EPIC, Kassandra, xCell, and XDec-SM. The color of the correlation value in each cell is representative of the strength of the correlation (scale on the y-axis).

https://doi.org/10.1371/journal.pcbi.1011365.g002

deconvolution tools identify a widely varying number of cell types. For example, Kassandra and xCell both identified a large number of cell types (20 and 64 respectively) of which a vast majority were immune cell types. CIBERSORTx and EPIC identified fewer and more generalized cell types (6 and 9 cell types respectively). Despite the variability, all the deconvolution tools inferred epithelial, immune, and stromal proportions, which we used as a "common denominator" for comparing their performance. There is a fairly high correlation of XDec-SM with each of these methods (Fig 2A–2D). To further investigate the cell type specific correlation between the proportions predicted with each of these methods, we then correlated the

predicted proportions of epithelial cells (Fig 2E), the predicted proportion of immune cells (Fig 2F), and the predicted proportion of stromal cells (Fig 2G). Across all cell types XDec-SM has the highest correlation with all other methods (epithelial mean correlation = 0.69; immune mean correlation = 0.82; stromal mean correlation = 0.80).

Due to the widespread use of CIBERSORTx [13], we then investigated the differences in the proportions estimated by CIBERSORTx and XDec-SM (Fig 3A). CIBERSORTx is also equipped to employ scRNA-seq reference profiles and outputs both cell type specific profiles and per-sample proportions. Despite relatively high correlations (epithelial $R^2 = 0.93$; immune $R^2 = 0.80$; stromal $R^2 = 0.78$, Fig 3A), systematic discrepancies between the two methods could be observed (Fig 3A).

To interpret the discrepancies between XDec-SM and CIBERSORTx, we proceeded to independently estimate cell type proportions by InfiniumPurify [22] and TIMER [23]. We observed higher concordance between cancer cell fraction estimates of InfiniumPurify and XDec (slope = 1.02, $R^2 = 0.97$, Fig 3B), than CIBERSORTx (slope = 0.72, $R^2 = 0.89$, Fig 3E). We also observed higher concordance between immune fraction estimates of TIMER and XDec-SM (slope = 0.11, $R^2 = 0.94$, Fig 3C) than CIBERSORTx (slope = 0.06, $R^2 = 0.74$, Fig 3F). Taken together, higher concordances of XDec-SM with both InfiniumPurify and TIMER suggest that constituent cell proportions predicted by XDec-SM are accurate.

Because XDec-SM was developed using the EDec method as the starting point and focuses on the state of gene activation (vs. gene-specific transcription levels), we next asked if XDec-SM recapitulates the results of epigenomic methylation-based deconvolution. Despite XDec-SM and EDec exploiting different data types–RNA-seq and DNA methylation respectively, both methods focus on the gene activation state (constrained to the 0–1 scale). Therefore, we hypothesize that these methods will be concordant. Indeed, we observed high concordance between XDec-SM and EDec across all three cell class proportions (epithelial, slope = 0.99, $R^2 = 0.98$; immune, slope = 0.74, $R^2 = 0.88$; stromal, slope = 1.05, $R^2 = 0.84$, Fig 3D). CIBERSORTx shows lower concordance with EDec (epithelial, slope = 0.62, $R^2 = 0.89$; immune, slope = 0.41, $R^2 = 0.72$; stromal, slope = 2.59, $R^2 = 0.66$, Fig 3G).

Cancer cell state mapping provides insights into intra-tumoral heterogeneity

Unlike the categorical PAM50 classification of BRCA tumors, the XDec-SM mapping approach places cancer cells on the spectrum between the pure Luminal, Basal and HER2 subtypes (Fig 4A and 4B). As expected, the tumors classified as Basal by PAM50 were clustered primarily in the Basal corner of the simplex map. However, we observed that the tumors classified as Luminal or HER2 by PAM50 were found on a spectrum between the Luminal and HER2 vertices.

There are three possible scenarios that could account for the variation in the location of the tumors within the simplex map: (1) The cancer cells within any specific tumor are relatively pure representatives of Luminal, Basal and HER2 subtypes and that the map position reflects their proportions within the tumor; (2) The cancer cells within a tumor are mostly homogeneous and the position reflects their similarity to the three subtypes; (3) A hybrid between the first two scenarios. To begin to address this question, we leveraged a publicly available single cell RNA-seq dataset (GSE75688) that includes both single cell for breast cancer patients, as well as pooled single cell sequencing and bulk sequencing [24]. The single cell and bulk RNA-seq was deconvoluted with XDec, and the distribution of the cancer cell profiles was visualized with the simplex map (S3A–S3H Fig). Although these results are preliminary, we identify interesting trends in the distribution of the cancer cell profiles of samples generated from the same sample. The single cell and pooled profiles from the same sample did not cluster tightly

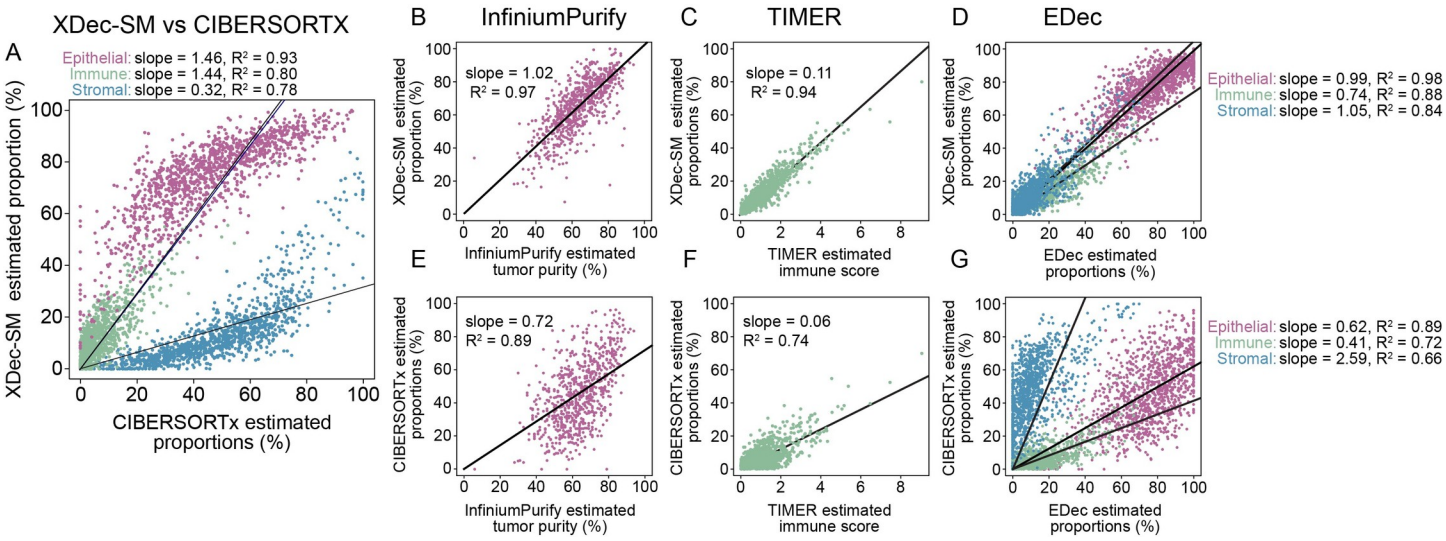

Fig 3. Comparison of XDec-SM and CIBERSORTx. (A) Scatterplot between CIBERSORTx (x-axis) and XDec-SM (y-axis). (B) Scatterplot between InfiniumPurify (x-axis) estimated tumor purity and XDec-SM (y-axis) estimated per-sample proportion of epithelial. For XDec-SM, the five epithelial profiles are summed to represent epithelial. (C) Scatterplot between TIMER (x-axis) estimated immune score and XDec-SM (y-axis) estimated per-sample proportion of epithelial. For XDec-SM, the two immune profiles are summed to represent immune and compared to the sum of all immune subtype scores estimated by TIMER. (D) Scatterplot between EDec (x-axis) and XDec-SM (y-axis) estimated per-sample proportions of each constituent cell type. For EDec, the six epithelial profiles are summed to represent epithelial. For XDec-SM, the five epithelial profiles are summed to represent epithelial, the two stromal profiles are summed to represent stromal, and the T cell and Macrophage are summed to represent immune. (E) Scatterplot between InfiniumPurify (x-axis) estimated tumor purity and CIBERSORTx (y-axis) estimated per-sample proportion of epithelial. (F) Scatterplot between TIMER (x-axis) estimated immune score and CIBERSORTx (y-axis) estimated per-sample proportion of epithelial. For CIBERSORTx, the three immune profiles are summed to represent a general immune and compared to the sum of all immune subtype scores estimated by TIMER. (G) Scatterplot between EDec (x-axis) and CIBERSORTx (y-axis) estimated per-sample proportions of each constituent cell type. For EDec, the six epithelial profiles are summed to represent epithelial. For CIBERSORTx, the two immune (B cell, T cell, macrophage) profiles are summed to represent immune and the two stromal (stroma, endothelial) profiles are summed to represent stromal.

https://doi.org/10.1371/journal.pcbi.1011365.g003

(S3A Fig), suggesting intra-tumoral heterogeneity of cancer cell states. Interestingly, most of the heterogeneity occurred along the Luminal-Basal or Luminal-HER2 axes (S3A–S3H Fig). This suggests that intra-tumoral heterogeneity partially contributes to the intermediate epithelial profiles seen in bulk tumors in S3A Fig. This preliminary conclusion is supported by previous scRNA seq studies which have shown the wide range of heterogeneity in breast cancer tumors [25]. While pooling single cells has advantages in both library preparation and cost as well as the reduction of technical noise [26] these results suggest that pooling samples into pseudo-bulks may not accurately recapitulate the biology of all of the single cell samples. Due to the large degree of intratumoral cancer cell heterogeneity [27] these results may also argue against the use of pooled single cell data as proxies for the entire tumor in future cancer studies.

The results in this study remain preliminary and may be confounded by the presence of pooled samples. Additionally, the wide range of intra-tumoral heterogeneity observed in this study could be related to the spatial relationship of the single cells within the tumor. However, this single cell deconvolution analysis has shown that XDec-SM may be a useful tool for the future interpretation of single cell data and could aid in the interpretation of how representative pseudo-bulks are of individual single cell samples.

Mapping reveals heterotypic interactions between cancer cells and the tumor microenvironment

We next investigated interactions between stromal and cancer cells by correlating the transition from adipocyte to CAF stroma with changes within the cancer cells. CAFs have been previously shown to be associated with progression and response to therapy [26–33]. Along the

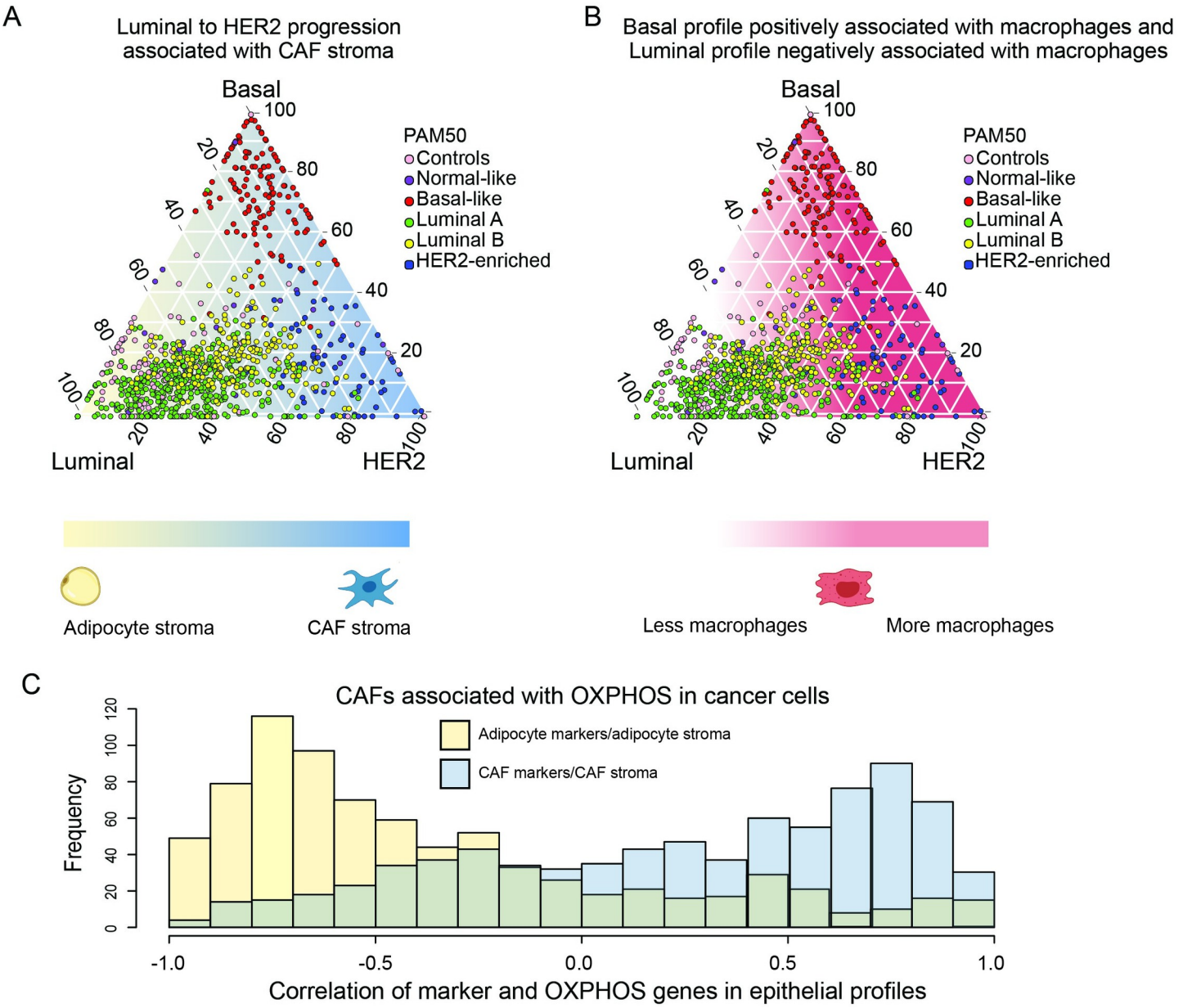

Fig 4. Cancer cell state map reveals intra-tumoral heterogeneity and tumor microenvironment involvement. (A) Map of the deconvoluted TCGA breast cancer sample proportions of epithelial 1 (Basal), epithelial 3 (HER2) and epithelial 4 (Luminal). Proportions of all three profiles are normalized to equal 100. Dot color indicates the TCGA defined PAM50 subtype. The background color indicates the enrichment of the adipocyte stromal profile vs the CAF stromal profile. Blue indicates a higher proportion of CAFs and yellow represents a higher proportion of adipocytes. (B) Map of the deconvoluted TCGA breast cancer sample. In this figure the background color indicates the proportion of macrophages present in these tumor samples. (C) Histogram showing the correlation between adipocyte markers in the adipocyte stromal profile (yellow) and OXPHOS genes in the epithelial profile, and CAF markers in the CAF stromal profile (blue) and OXPHOS genes in the epithelial profile (blue).

https://doi.org/10.1371/journal.pcbi.1011365.g004

Luminal-HER2 axis we observed an increase in the proportion of CAFs (correlation = 0.19) as the proportion of the HER2 profile increased (Fig 4A). We hypothesized that as the cancer cells shift towards the HER2 phenotype, metabolite exchange between cancer cells and the stroma contributes to a more malignant phenotype [28,32,34]. There are several molecular mechanisms of metabolite exchange between the epithelial and stromal cells, including exosomes [35] and other soluble factors [36,37].

Previous studies reported the reverse Warburg effect model in which glycolytic stroma (CAFs) feed lactate and promote oxidative phosphorylation (OXPHOS) in cancer epithelial cells [18,38,39]. These early results were indirect and not based on the deconvolution of the transcriptomic states of stromal cells within individual tumors. To correlate the gene expression changes in stromal and epithelial cells directly, we performed Stage 2 deconvolution of the epithelial fractions (Methods). As predicted, we observed a negative correlation between adipocyte gene expression markers in the stromal profile and the expression of epithelial OXPHOS genes. Moreover, the expression of OXPHOS genes in the epithelial fraction correlated positively with stromal CAF markers (Fig 4C). Taken together these correlations suggest metabolic coupling between stromal and cancer epithelial cells consistent with the negative Warburg effect that involves lactate transfer from stroma to cancer epithelial cells.

We also investigated the relationship between cancer cell state, and the immune cell component. While we did not find a correlation between T-cell proportion and cancer cell state, we observed an increase in the proportion of macrophages for the Basal epithelial profile (correlation = 0.22, p-value = 7.12e-12), and a decrease in the proportion of macrophages for the Luminal epithelial profile (correlation = -0.25, p-value = 1.28e-14) (Fig 4B). The increase of Tumor Associated Macrophages (TAMs) was previously associated with Basal breast cancers, non-luminal breast cancers, and predicts poor survival [40–42]. Thus, the state of the cancer cell is associated with both the stromal and immune profiles of the tumor.

Matching tumors *in vivo* to PDX models and cell lines

One key challenge in using model systems, such as murine PDX models and cell lines, is to determine which models best represent specific patient tumors. Variability in cell type composition precludes matching by the intrinsic state of the cancer cell based on bulk profiling patient-derived tumor samples.

To explore the correspondence based on the transcriptomic state of constituent cancer cells, we projected primary TCGA tumors (left), PDX models (middle), and BRCA cell lines from the Cancer Cell Line Encyclopedia (CCLE) (right) onto the same map (Fig 5A). The density of the distributions indicates that the TCGA tumors are primarily Luminal, the PDX models largely Basal, while the CCLE cell lines show a moderate bias toward the HER2 phenotype. The map therefore provides information about model-specific biases toward specific tumor subtypes while enabling selection of models that best match specific tumors.

Cancer cell state map refines PAM50 classification based on cancer cell-intrinsic states

The PAM50 classification system has been widely utilized for identifying the "intrinsic" subtype of breast tumors. The classification is actionable, with endocrine therapy recommended as the first-line treatment for the tumors classified as Luminal by PAM50 [43] In a significant fraction of cases, however, PAM50 classification of tumors does not match HER2 IHC status, another clinically established marker [44]. Notably, PAM50 classification was originally developed based on microarray-based gene expression profiling. In contrast, the map is constructed based on the more informative RNA-seq profiling information. Moreover, PAM50 classification is confounded by the sample-to-sample variation in cell type composition. In contrast, XDec-SM deconvolution provides information about the transcriptomic state intrinsic to the cancer cell fraction of the tumor. We therefore asked how the two classifications compared in primary TCGA breast cancer tumors.

For each tumor we obtained PAM50 classification assigned in the TCGA database and determined XDec classification by identifying the cancer profile (Luminal, HER2, Basal) with

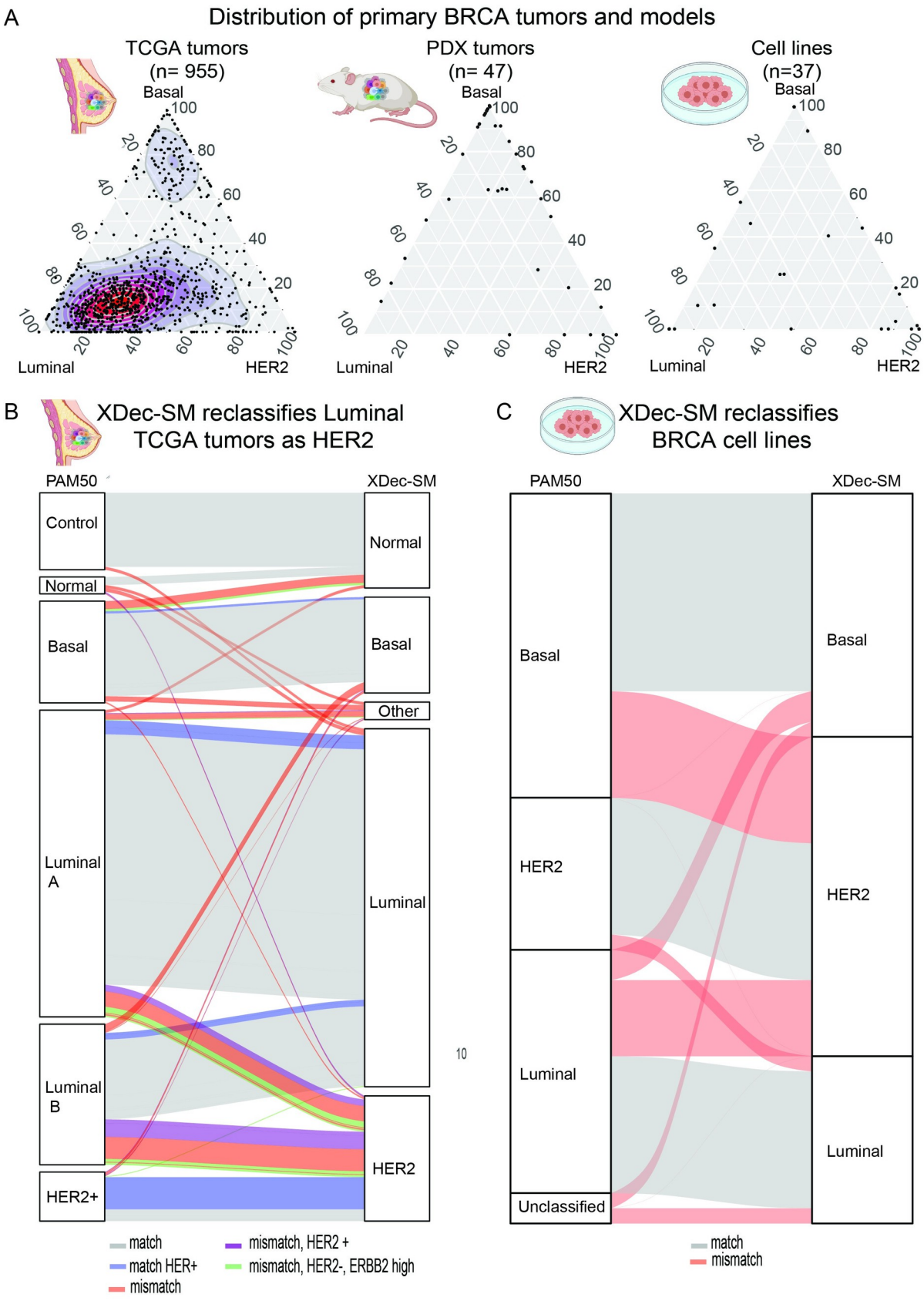
A
Distribution of primary BRCA tumors and models
TCGA tumors
(n= 955)
PDX tumors
(n= 47)
Cell lines
(n=37)
Basal
Luminal
HER2
B
XDec-SM reclassifies Luminal
TCGA tumors as HER2
PAM50
XDec-SM
Control
Normal
Basal
Luminal A
Luminal B
HER2+
Other
Luminal
HER2
match
match HER+
mismatch
mismatch, HER2 +
mismatch, HER2-, ERBB2 high
C
XDec-SM reclassifies
BRCA cell lines
Unclassified
match
mismatch

Fig 5. Distribution of samples and classification of cell lines by PAM50 and XDec-SM. (A) Ternary map of sample position for TCGA (left) PDX (center) and cell lines (right). The density is represented as less dense (blue) to denser (red). The distribution of different sample types varies with PDXs primarily representing Basal tumors and TCGA and cell lines being more widely spread in the simplex. (B) Alluvial plot of TCGA breast cancer sample PAM50 and XDec-SM classification. PAM50 subtypes are provided by TCGA sample metadata. XDec-SM classification indicates the classification based on the maximum proportion of the deconvoluted epithelial profiles. Gray line indicates matching subtype, blue line indicates matching subtype and HER2+ based on the metadata provided by TCGA. Red line indicates mismatching subtype and purple indicates mismatch and HER2+. Green line indicates a mismatch that is HER2— but has a high activation of the ErBB2 pathway. (C) Alluvial plot of XDec-SM clustering classification vs highest proportion for CCLE cell lines. The grey lines indicate samples that were consistently classified by both methods. The red lines indicate samples that are differently classified by both methods. Figure components created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1011365.g005

the highest proportion (closest corner in the map). Most tumors had the same classification (Fig 5B, grey line). As expected, HER2+ tumors (Fig 5B, blue line) were primarily classified as HER2 by both methods. The largest discrepancies involved tumors classified as Luminal by PAM50 and HER2 by XDec-SM (Fig 5B, red line). A subset of those tumors was HER2 positive (Fig 5B, purple), or had a high ErBB2 signaling network activation score (Fig 5B, green line), concordant with XDec-SM classification, suggesting that the state at least some of the cases reflect ErBB2 pathway activation. We then proceeded to ask if we could also use the XDec-SM classification method to characterize breast cancer cell lines.

Cancer cell lines are an important model system, and datasets such as the CCLE have characterized large numbers of them. Because cell lines often evolve *in vitro* and differ from primary tumors, there is long-standing interest in identifying the cell lines that best model specific types of breast tumors [45]. PAM50 classification has been utilized for both cell lines and tumors, thus helping match the tumors and cell lines by type. However, there are several drawbacks to applying PAM50 to cell lines. Primarily, PAM50 was developed on bulk array data, and contains stromal genes that are not intrinsically represented in cell lines. We focused on characterizing the state of cell lines by placing them on the same cancer cell state map as deconvoluted primary tumors.

To characterize commonly used breast cancer cell lines, we obtained the expression data from 51 cell lines profiled by the CCLE [46]. The PAM50 annotations for the cell lines were obtained from the study by Jiang et al [45]. We classified the cell lines by their map position, by assigning them to their closest vertex (Fig 5C). As with the primary TCGA breast cancer tumors, for the majority of cell lines the PAM50 and XDec-SM classification were concordant (n = 31). However, a fraction of cell lines that were classified as Luminal or Basal by PAM50 were reclassified as HER2 by the XDec-SM classification method (n = 12). Two cell lines that were classified as Luminal by PAM50 were reclassified as Basal by XDec-SM and one cell line that was classified as HER2 by PAM50 was reclassified as Luminal. As the XDec-SM classification method is based on cancer cell state, we were also able to classify cell lines which were previously unclassified by the PAM50 method (n = 2) (Fig 5C).

Cancer cell state associates with driver mutations, key pathway activations, and DNA methylation patterns

We next asked how the positions on the map correlate with mutations in key tumor suppressors, oncogenes, and with the activation of pathways driving cancer progression. As indicated in Fig 6A, mutations in PIK3CA localize in the Luminal corner, while TP53 mutations localize in the Basal corner consistent with the high subtype-specific prevalence of mutations in these genes. In contrast TTN mutations are not correlated with a specific subtype, and instead show the same pattern of distribution as TCGA (Figs 5A and 6A).

We then used the cancer cell state map as a framework to visualize pathway activation by plotting the 25 tumors with the highest pathway activation score for EGFR, estrogen, and hypoxia. EGFR pathway activation is more activated in the Basal subtype, with the estrogen

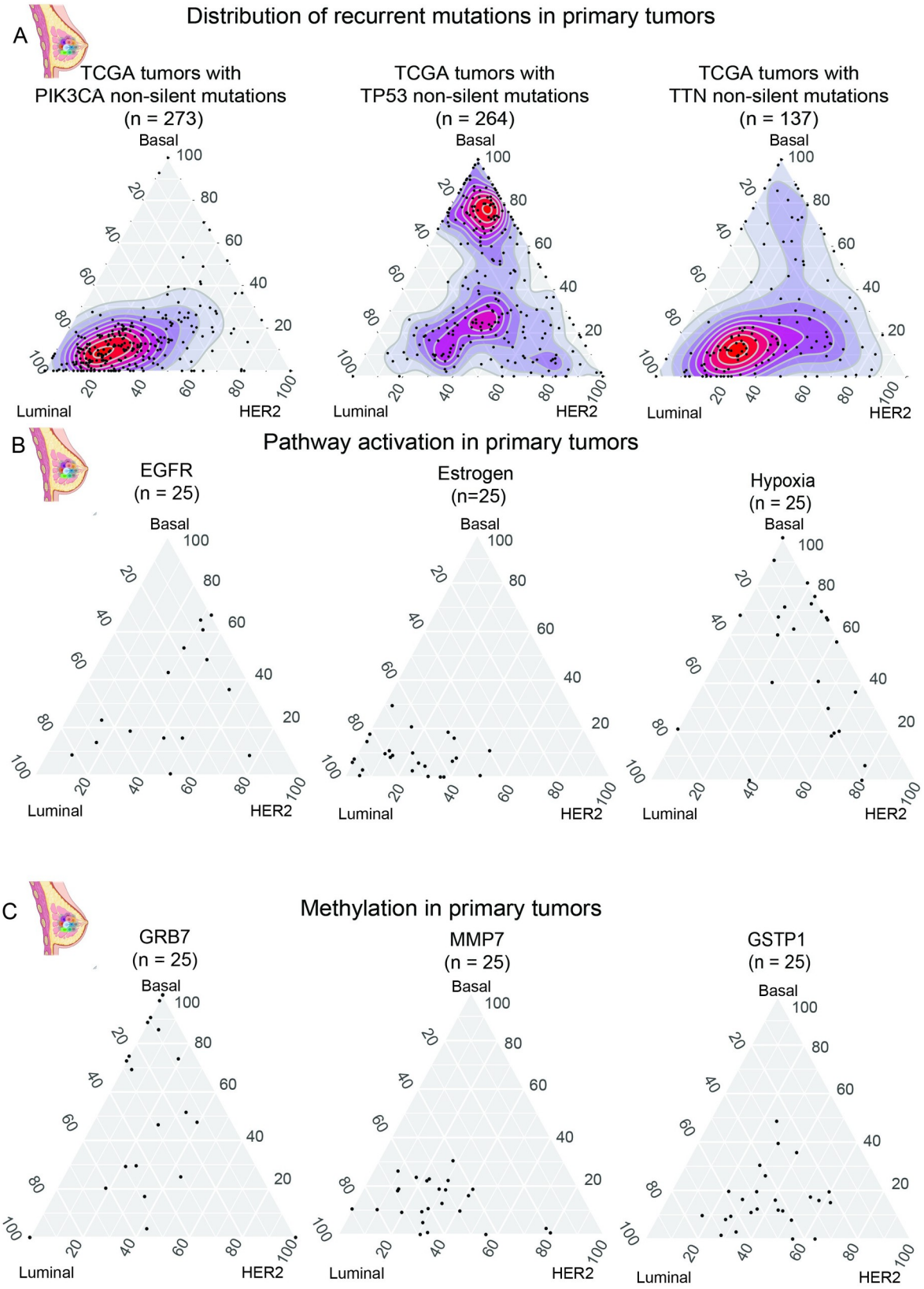
A
Distribution of recurrent mutations in primary tumors
TCGA tumors with PIK3CA non-silent mutations (n = 273)
TCGA tumors with TP53 non-silent mutations (n = 264)
TCGA tumors with TTN non-silent mutations (n = 137)
B
Pathway activation in primary tumors
EGFR (n = 25)
Estrogen (n=25)
Hypoxia (n = 25)
C
Methylation in primary tumors
GRB7 (n = 25)
MMP7 (n = 25)
GSTP1 (n = 25)
Basal
Luminal
HER2

Fig 6. Distribution of mutations, pathway activation, and methylation profiles in cancer cell state map. (A) Distribution of common breast cancer mutations including PIK3CA (left), TP53 (middle), and TTN (right). The density is represented as less dense (blue) to denser (red). (B) The activation of the EGFR, Estrogen, and hypoxia pathways are also projected on the ternary graph. For each pathway, the tumors with the highest activation score are plotted. (C) Methylation patterns of GRB7, MMP7, and GSTP1. The 25 tumors with the highest average methylation scores across all loci in each gene are plotted on the cancer cell state map. Figure components created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1011365.g006

pathway being active in Luminal and HER2 subtypes (Fig 6B). Basal tumors show highest activation for Hypoxia, consistent with inactivating TP53 mutations protecting the Basal cancer cells from the apoptotic effects of the TP53 pathway.

The cancer cell state was also used to visualize the methylation profiles of the TCGA breast cancer tumors (Fig 5C). As with the pathway analysis, the tumors that had the highest level of methylation (n = 25) were plotted on the map for GRB7, MMP7, and GSTP1. GRB7 is the most methylated in the Basal corner. Likewise, MMP7 and GSTP1 which have been associated with luminal A and luminal B subtypes respectively are found to be the most methylated along the Luminal to HER2 axis. In summary, these mutational, pathways activation, and methylation patterns recapitulate known breast cancer biology, suggesting that the position in the map should be highly informative for the state and behavior of the cancer cell.

Cancer cell state in PDX models is predictive of chemotherapy response

We next asked if the transcriptomic state of cancer cells in PDX models was predictive for chemotherapy response. Forty- four PDX models from the BCM collection with known response to carboplatin and docetaxel were deconvoluted. Information about these PDX models can be accessed through the PDX portal (https://pdxportal.research.bcm.edu/). The Basal, Luminal, and HER2 proportions of these PDXs were used to build generalized linear models (GLMs) of response to both therapies.

The GLM models built on the BCM PDX cohort were used to predict the chemotherapy response of an independent dataset consisting of 22 PDXs from the Rosalind & Morris Goodman Cancer Research Centre (RMGCRC) (GSE142767) [47]. The response to both agents in the RMGCRC cohort was annotated with RECIST criteria as complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD). While the training and test datasets did not use the same platinum and taxane compounds, they were matched by compound class (carboplatin GLM used to predict cisplatin response and docetaxel GLM used to predict paclitaxel response). With only the three predictor variables (Luminal, Basal, HER2 proportion), we found that the GLM models had high predictive power for the platinum agents (ROC complete response vs all else (CR) = 0.73, ROC complete and partial response vs all else (CRPR) = 0.80) and to a lesser degree the taxanes (ROC complete response vs all else = 0.58 (CR), ROC complete and partial response vs all else (CRPR) = 0.67) (Fig 7A). Taken together, these results suggest that the state of cancer cells in PDX models is predictive of response to both platinum and taxane therapies. A similar approach could be utilized to predict the response of patients to neoadjuvant chemotherapy.

Cancer cell states in patient tumors predicts response to tamoxifen

As the state of the cancer cell is predictive of response to chemotherapy, we then asked if the XDec-SM classification itself was predictive of response. In the previous section titled "*Cancer cell state map refines PAM50 classification based on cancer cell-intrinsic states*" we found that the largest discrepancy between the PAM50 and XDec-SM classification of TGCA tumors was that a subset of tumors that were classified as Luminal by PAM50 were classified as HER2 by

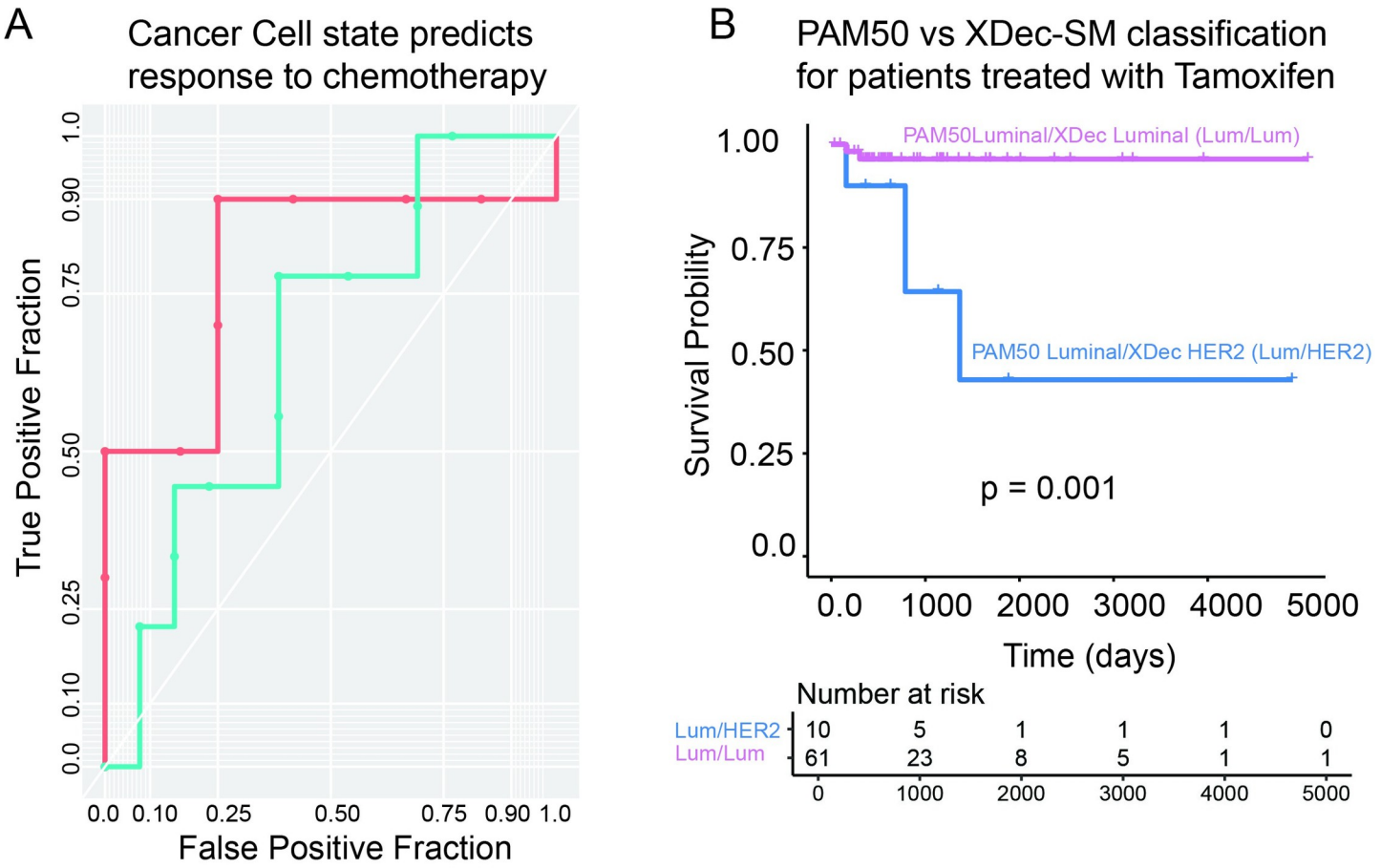

Fig 7. Cancer cell state map is predictive of therapy response. (A) ROC curve of the response to cisplatin and paclitaxel in the Goodman PDX dataset as predicted using a GLM model built on the epithelial proportions and the carboplatin and docetaxel response in the BCM PDX dataset. CRPR vs all else for platinum agents is in red (with an AUC of 0.80) and CRPR vs all else for taxanes is in blue (with an AUC of 0.67). (B) Kaplan-Meier Curve that indicates the difference in survival between samples classified as Luminal by PAM50 and XDec-SM (pink) and samples classified as Luminal by PAM50 and HER2 by XDec-SM (blue). The patient cohort size is denoted below the graph.

https://doi.org/10.1371/journal.pcbi.1011365.g007

XDec-SM. Additionally, some of tumors that were reclassified as HER2 by XDec-SM showed ErBB2 pathway activation which is a negative predictor of tamoxifen response.

To examine whether cancer cell state indeed identifies cancer cells that respond poorly to tamoxifen, we next focused on 72 TCGA BRCA tumors that were classified as Luminal by PAM50 [48] (using recount package treated with tamoxifen and had survival information (NCI GDC annotations). Of those, 62 were classified as Luminal and 10 as HER2 by XDec-SM. The two groups showed survival differences, with patients classified as HER2 by XDec-SM alone having significantly worse survival (Fig 7B, p = 0.001). Notably, neither HER2 status alone (S4A Fig), nor ErbB2 score (S4B Fig), could identify subsets of patients with significantly worse survival (p = 0.4 and 0.77 respectively). This suggests that cancer state alone could identify a subset of poor responders to tamoxifen therapy among those that were classified as Luminal by PAM50.

Discussion

To construct the cancer cell state map of breast cancer, we developed the XDec-SM deconvolution algorithm. Unlike other RNA-seq based deconvolution methods that are based on

reference profiles known *a priori* from profiling of previously isolated cell populations, XDec-SM is "reference-optional" as it starts by selecting informative genes, which may optionally be derived from reference profiles, such as those derived from single-cell profiling. Because references themselves are not used, the method enables data-driven discovery of novel transcriptomic states of cells within tumors and other complex tissues. Remarkably, the algorithm produces data-driven estimates of constituent cell types that correspond to PAM50 [18,49] (and the current molecular classification of breast cancer into Luminal, HER2 and Basal subtypes. Unlike other similar methods [50] XDec-SM allows us to map tumors onto a low dimensional (2D) simplex map. Importantly, this simplex map is biologically interpretable, with each vertex anchored at a known cancer cell subtype. By mapping breast cancer tumors onto this simplex map, we refine the PAM50 classification by placing tumors on the spectrum between established PAM50 subtypes. XDec-SM is a flexible and generic method that is not constrained to the deconvolution of breast cancer and future studies could use XDec-SM to deconvolute other cancer types to determine data-derived cancer cell states. In fact, XDec-CHI a closely related algorithm [51]) has been used to deconvolute a wide variety of cancers including: gliomas, glioblastomas, head and neck cancer, pancreatic adenocarcinomas, and breast cancer.

Additionally, unlike other RNA-seq based deconvolution methods that focus on explaining variance in levels of RNA, XDec-SM in Stage 1 scales gene expression levels within the 0–1 range, thus factoring out absolute levels of gene transcription and focusing on the information about the epigenetically regulated on/off states of gene transcription. The epigenetic state of the cell is programmable during development and intrinsic factors such as hormonal signaling, paracrine signaling and other heterotypic interactions between cells within complex tumor tissue leave an "epigenetic footprint". The epigenetic state of the cell can in principle be estimated by measuring either "classical" epigenetic marks such as DNA methylation or gene expression levels. However, molecular array or sequencing-based profiling of bulk tumor tissue is confounded by the sample-to-sample variability in cell type composition, precluding access to the intrinsic epigenetic states of constituent cell types. By deconvoluting the cell state with XDec-SM, we can capture the cancer cell states in the sample and obtain a snapshot of the "epigenetic footprint" of the tumor projected onto the cancer cell state map. In this manner, XDec-SM extends the previously proposed Epigenomic Deconvolution method based on bulk DNA methylation profiles demonstrating that the epigenetic heterogeneity of tumors can be concordantly assessed from bulk RNA-seq or DNA methylation.

While XDec-SM benefits from information about which genes are informative for deconvolution gleaned from publicly accessible scRNA-seq profiles of tumor cells, it is also designed to infer new recurrent states of cancer cells not obvious from the analysis of scRNA-seq data. This method is reference-optional and can leverage information from both informative cell type specific loci and PAM50 literature-based gene signatures. The cancer cell state map is synergistic and complementary with these scRNA-seq methods, as it starts from bulk RNA-seq profiles and thus provides an embedding based on vastly more samples, including those without single-cell data covering a larger diversity of tumors albeit at lower resolution. Because bulk RNA-sequencing data is available for many more samples and does not involve perturbation of the cell state associated with physical isolation, XDec-SM approach is uniquely positioned to provide a view of the full spectrum of recurrent states of cancer cells *in vivo*.

We show that the cancer cell state map of breast cancer not only extends established tumor classification, but also provides additional information about tumor biology, model systems, and drug response. Additionally, through a preliminary analysis we demonstrate that the cancer cell state map could be useful as a framework for interpreting scRNA-seq profiles of individual tumors, through the identification of cancer cell state heterogeneity. Future application of XDec-SM will include deconvoluting single cell data for a wide range of cancer types.

Our findings demonstrate that cancer cell state map helps visualize and quantitatively analyze (by map position) heterotypic interactions within tumors. By correlating the states of stromal and cancer cell fractions, we recapitulate metabolic coupling between them, consistent with the previously reported reverse Warburg effect [16]. There are several mechanisms through which metabolite exchange between the epithelial and stromal cells can occur, including soluble factors [36,37] and exosomes [35]. More broadly, our findings reinforce the role of cancer-associated fibroblasts in cancer [26–33]. Additionally, the map sheds light on the association of tumor-associated macrophages with the Basal subtype. Because some of the heterotypic interactions may be mediated by exosomes and other highly specific factors that may be found in bodily fluids, the cell-cell communication discovered by cell state mapping may open the doors toward the development of liquid biopsy biomarkers.

Additionally, the precise characterization of the state of the cancer cell fraction within individual tumors improves the matching of individual tumors with specific experimentally accessible PDX and cell line models. The map places breast cancer cell lines on the epigenetic spectrum between established breast cancer subtypes, identifying the subsets that can serve as subtype-specific models, versus those that do not clearly model any established subtype of breast cancer. This information is critical for understanding the translational potential of *in vitro* therapeutic drug screens [46] as each breast cancer subtype may have a distinct response [52]. We demonstrate the utility of this low dimensional to map to predict therapy response in experimentally tractable models and tumors *in vivo*. The map position predicts response to standard chemotherapies based on PDX models and targeted therapies based on cell line models. Cancer cell state mapping identifies a subset of tumors classified as Luminal by PAM50 [18,49] that are HER2-like and are resistant to tamoxifen therapy, a first-line therapy for this predominant subtype of breast cancer. While these preliminary results are based on a small number of samples, they demonstrate the potential utility of the application of cell state mapping in precision medicine.

We note that the XDec-SM algorithm is very generic and is not limited to deconvolution of solid tissues. Specifically, XDec-SM has previously been applied to deconvolute thousands of extracellular RNA-seq profiles of human body fluids in the exRNA Atlas and to construct the first reference map of extracellular RNA in human body fluids [53]. Analogously to the scRNA-seq profiles, the selection of informative features was informed by publicly available RNA-seq profiles from experimentally isolated exRNA carriers (vesicle, lipoprotein, RNA-binding protein).

To empower the community to use this method, we made the XDec-SM code available online and as a R package under a free open-source license. All analysis code and source data is available online (https://github.com/BRL-BCM/XDec_SM). We also developed an interactive web service that allows users to deconvolute breast tumor RNA-seq profiles of interest and project them onto the cancer cell state map (https://brl-bcm.shinyapps.io/XDec_BRCA/).

Limitations of the study

The power of XDec-SM method is yet to be fully characterized and it may be hard to characterize under realistic assumptions. For example, while several states of breast cancer cells were identified, is not clear how many additional profiles would be required to identify the diverse states within stromal cell types. The granularity of the XDec-SM deconvolution is also dependent of the availability of single cell reference profiles and the amount of sample-to-sample variability in cell type proportions and the variability in the proportions of cells in specific states. While the proliferation of single cell studies and bulk profiling will help increase the resolution of the map over time, because of the interplay of relevant parameters, it would be hard

to estimate, under realistic assumptions. the number and types of profiles required for such high-resolution map.

Methods

Material availability

This study did not generate new unique reagents.

XDec-SM deconvolution algorithm

XDec-SM is a variation of the EDec algorithm, adjusted for RNA-seq data. EDec is an iterative algorithm for constrained matrix factorization using quadratic programming for DNA methylation-based deconvolution [16]. The key difference is that in Stage 1 of XDec-SM expression values are transformed by a single-parameter negative exponential function into the 0–1 range (the same range that applies to original DNA methylation). XDec-SM has originally been applied to deconvolute extracellular RNA-seq profiles of human bodily fluids [53]. Like EDec, XDec-SM has three stages: Stage 0, 1, 2 (Fig 1A).

XDec-SM relies on a set of informative features that are differentially abundant across each component modeled in the system (Fig 1A, Stage 0). The set of informative genes may be obtained by RNA sequencing of cell lines or by single-cell RNA sequencing. In case of exRNA deconvolution, this information is provided by the RNA sequencing of the cargo of specific carriers of extracellular RNA in human body fluids (various vesicles, lipoprotein particles, and RNA-binding proteins). In any case these profiles are used for the sole purpose of identifying sets of informative loci and otherwise do not bias the output of deconvolution.

Stage 1 involves iterative matrix factorization to estimate the gene expression profiles of constituent cell types and the per-sample proportions of constituent cell types in individual tumors. As mentioned above, for the purpose of this computation, expression counts are transformed to the [0–1] range using the following negative exponential function:

$$y = 1 - e^{-ax}, a = \frac{1}{max\ \{x\}}$$

For each application of the XDec-SM Stage 1, we identify the most stable k (number of cell types) that best models the data using random initializations. Using 3 replicates of 80% of the input data, deconvolution is performed modeling 3 to 10 cell types using the informative features from Stage 0. The estimated proportions from the three replicates are compared for each cell type model and the number of cell types is determined once the correlations are no longer significant. The most stable cell number is then selected as the estimated cell number. This cell number selection was done using the estimate_stability function which was developed and tested in the original EDec method [16].

In Stage 2 (Fig 1A, Stage 2), XDec-SM uses the per-sample proportions estimated in Stage 1 to estimate the untransformed gene expression profile of each constituent cell type utilizing constrained least-squares fit using quadratic programming. This gene expression profile represents the average gene expression across the set of samples.

Cell line RNA sequencing data processing

This set represented 7 cell types: epithelial (GSM1695870, GSM1695871, GSM1695872, GSM1695873, GSM1695874, GSM1695875, GSM1695876, GSM1695877, GSM1695878, GSM1695879, GSM1401648, GSM1401649, GSM1401650, GSM1401651, GSM1401653, GSM1401654, GSM1401665, GSM1401668, GSM1401672, GSM1401673, GSM1401674,

GSM1401675), B cells (GSM1576391, GSM1576419), dendritic cells (GSM1576395, GSM1576423), monocytes (GSM1576399, GSM1576427), CD56+ natural killer cells (GSM1576407, GSM1576435), T cell (GSM1576415, GSM1576443) and stroma (GSM2430225, GSM2430226, GSM2430227, GSM2839374). Transcript fastq files were downloaded from GEO and then aligned and quantified using STAR (Version 2.4) and RSEM (Version 1.3). The reference genome was assembled using the standard rsem-prepare-reference function using the hg19 genome assembly. RSEM was run using standard parameters (rsem-calculate-expression–star -p 6 –star-gzipped-read-file–paired-end) to produce RSEM gene counts for each of the cell line profiles. The resulting 36 profiles all met high read quality standards using fastqc with standard parameters. RSEM gene expression profiles for all 36 profiles were transformed using Equation 1 to constrain the values to a [0–1] range. Genes with low coverage (= < 0.01) and those not identified within the TCGA breast cancer profiles were removed from further analysis.

scRNA sequencing data processing

Single cell gene expression RSEM counts (GSE118389_counts_rsem) were downloaded from GEO (GSE118389). Metadata provided by the study determined the cell type identity [6]. For XDec-SM, endothelial and B cell profiles were removed from further analysis due to low number of profiles. Initially all single cell profiles per cell type are ranked based on total read coverage. Profiles with low coverage across all genes are moved (sum of gene counts < 100,00) are removed. Every five profiles are then summed to create pseudo bulks [54]. Gene reads for each pseudo bulk are then normalized to match the highest coverage pseudo bulks across all cell types. All pseudo bulks are transformed using Equation 1 to a 0–1 range. This results in pseudo bulks for four cell types: epithelial (n = 128), stromal (n = 18), T cell (n = 10), macrophages (n = 12).

TCGA data processing

The TCGA BRCA dataset (IlluminaHiSeq_RNASeqV2) and clinical phenotype metadata (BRCA_clinicalMatrix) were downloaded from UCSC Xena (https://xena.ucsc.edu). Only samples with matching gene expression profiles and metadata (n = 1,218) were kept for further analysis. PAM50_mRNA_nature2012 metadata information was used for the subtype determination. Samples with "-11" were designated as healthy controls. TCGA samples were quantile normalized and transformed using Equation 1 to constrain gene expression values to a 0–1 range. Genes with low gene expression coverage across all samples (average < 0.01) were removed.

XDec-SM application to simulated dataset

To test the XDec-SM methodology, we first simulated 100 mixtures based on publicly available, experimentally isolated gene expression profiles (twelve breast cancer epithelial, ten normal epithelial, ten immune, and four stromal, all listed below). The gene expression profiles for each cell type were randomly chosen and transformed into the [0–1] range. These profiles were then mixed in varying proportions, and random noise was added to simulate heterogenous tumor data (described below).

Stage 0 was performed to identify differentially expressed genes (n = 226) across the reference profiles of the 4 cell types. XDec-SM Stage 1 was applied to the simulated mixture samples and 4 constituent cell types were modeled. The estimated expression profiles are correlated to the reference expression profiles over the informative gene set to ascertain the corresponding identity of each cell type. XDec-SM can accurately deconvolute the gene expression profile of

each cell type when compared to the reference profile with the greatest correlation (cancer epithelial, $R^2 = 0.88$; normal epithelial, $R^2 = 0.97$; immune, $R^2 = 0.96$; stromal, $R^2 = 0.98$). Additionally, XDec-SM accurately estimates the per-sample proportion of all four cell types (cancer epithelial, $R^2 = 0.98$; normal epithelial, $R^2 = 0.97$; immune, $R^2 = 0.96$; stromal, $R^2 = 0.96$).

To generate the simulated mixture samples, a set of 36 cell line expression profiles collected from GEO (https://www.ncbi.nlm.nih.gov/geo/) were utilized. This set represented 4 major cell types: normal epithelial (GSM1695870, GSM1695871, GSM1695872, GSM1695873, GSM1695874, GSM1695875, GSM1695876, GSM1695877, GSM1695878, GSM1695879), cancer epithelial (GSM1401648, GSM1401649, GSM1401650, GSM1401651, GSM1401653, GSM1401654, GSM1401665, GSM1401668, GSM1401672, GSM1401673, GSM1401674, GSM1401675), immune (GSM1576391, GSM1576395, GSM1576399, GSM1576407, GSM1576415, GSM1576419, GSM1576423, GSM1576427, GSM1576435, GSM1576443), and stroma (GSM2430225, GSM2430226, GSM2430227, GSM2839374). The cancer epithelial cell lines include a diverse set of breast cancer cell lines of varying subtypes (ZR-75-1, BT-474, MDA-MB-468, HCC38, MCF-7, T-47D, HCC1954, HCC1937, HCC1187, BT-20, HCC1569, SUM-102). Immune cell lines include two profiles for each subtype: B cells, dendritic cells, monocytes, CD56+ natural killer cells, and T cells. Stromal cell lines represent cancer associated fibroblasts (CAFs). Transcript fastq files were downloaded from GEO and aligned and quantified using STAR (Version 2.4) and RSEM (Version 1.3). The reference genome was assembled using the standard rsem-prepare-reference function using the hg19 genome assembly. RSEM was run using standard parameters (rsem-calculate-expression—star -p 6—star-gzipped-read-file—paired-end) to produce RSEM gene counts for each of the cell line profiles. The resulting 36 profiles all met high read quality standards using fastqc with standard parameters (https://github.com/s-andrews/FastQC). RSEM gene expression profiles for all 36 profiles were transformed as described above to constrain the values to the [0–1] range. Genes with low coverage ($= < 0.01$) are removed from further analysis.

An independent gene expression random variable was used to generate noisy versions of the expression profiles. The random variable was selected as 10% of the maximum variance for the cancerous epithelial profiles and 5% for the normal cell types (normal epithelial, stromal, immune). Once the noisy profiles were generated, each simulated mixture contains a randomly selected profile for each of the 4 cell types. The simulated mixtures are a linear combination of the noisy expression profiles and a set of proportions for each of the 4 cell types. High purity mixtures included greater than 60% cancer epithelial, impure samples include 40% cancer epithelial, low purity included 10% cancer epithelial and over 50% normal epithelial, and control samples included 0% cancer and 70% normal epithelial. The remaining proportions are divided by the stromal and immune cell types.

Using the 4 cell type classes (cancerous epithelial, normal epithelial, immune, stroma), XDec-SM Stage 0 was performed to identify informative genes. We performed t-test across the 4 cell type classes comparing the transformed gene expression values between each group of reference profiles of the same cell type to the remaining expression profiles. For those genes with a significant differential expression ($p.value < 0.0001$), the 25 most upregulated and the 25 most downregulated genes were selected to represent each cell type. To further separate cancerous epithelial and stroma and normal epithelial and stroma, a direct t-test comparison was performed comparing the each of the two groups. For those genes with a significant differential expression ($p.value < 0.0001$), the 25 most upregulated and the 25 most downregulated genes were selected. If any genes appeared multiple times, they were excluded from the list of informative genes resulting in 226 genes.

XDec-SM Stage 1 was performed using the 100 simulated mixtures as input (4 normal epithelial, 4 immune and 4 stromal profiles were included in the input datasets for stability).

Deconvolution was performed using the 226 informative genes, a preset number of cells (n = 4), max iterations = 2000, and residual sum of squares differential stop = 1e-10.

XDec-SM application to TCGA dataset with informative genes from cell line profiles

Using the 7 cell type classes (epithelial, B cells, dendritic cells, monocytes, CD56+ natural killer cells, stroma), XDec-SM Stage 0 was performed to identify informative genes. We performed t-test across the 7 cell type classes comparing the transformed gene expression values between each group of reference profiles of the same cell type to the remaining expression profiles. For those genes with a significant differential expression (p.value < 0.0001), the 25 most upregulated and the 25 most downregulated genes were selected to represent each cell type. To further separate epithelial and stroma, a direct t-test comparison was performed comparing the two groups. For those genes with a significant differential expression (p.value < 0.00001), the 75 most upregulated and the 75 most downregulated genes were selected. If any genes appeared multiple times, they were excluded from the list of informative genes resulting in 391 genes. The PAM50 gene set was also included to define subtype specific expression for a total of 440 informative genes (49 out of the 50 PAM50 genes were identified in the TCGA dataset).

XDec-SM Stage 1 was performed using the 1,215 TCGA bulk tumor samples as input (4 normal epithelial, 5 immune and 4 stromal profiles were included in the input datasets for stability). Deconvolution was performed using the 440 informative genes, max iterations = 2000, and residual sum of squares differential stop = 1e-10. Stability criteria was used to identify a stable number of cells in the model. Using 3 replicates of 80% of the input data, deconvolution is performed modeling 3 to 10 cell types. The estimated proportions from the three replicates are compared for each cell type model and the number of cell types is determined once the correlations are no longer significant. The most stable cell number (which results in the most stable model) is selected as the data-driven cell number. The cell number estimation was done using the estimate_stability function which was developed and tested in the original EDec method [16]. Stability for this model resulted in 6 cell types.

XDec-SM application to TCGA dataset with informative genes from single-cell profiles

XDec-SM Stage 0 was performed to identify differentially expressed genes across the pseudo bulk profiles of the 4 cell types (epithelial, stromal, T cells, macrophages), XDec-SM Stage 0 was performed to identify informative genes. We performed t-test across the 4 cell type classes comparing the transformed gene expression values between each group of pseudo profiles of the same cell type to the remaining expression profiles. For those genes with a significant differential expression (p.value < 0.05), the 50 most upregulated and the 50 most downregulated genes were selected to represent each cell type. To further separate T cells and macrophages, a direct t-test comparison was performed comparing the two groups. For those genes with a significant differential expression (p.value < 0.05), the 25 most upregulated and the 25 most downregulated genes were selected. If any genes appeared multiple times, they were excluded from the list of informative genes resulting in 274 genes. Additionally, we included the PAM50 gene set to identify any subtype specific signatures for a total of 323 informative genes (274 plus 49 out of the 50 PAM50 genes).

XDec-SM Stage 1 was performed using the 1,215 TCGA bulk tumor samples as input. Deconvolution was performed using the 323 informative genes, max iterations = 2000, and residual sum of squares differential stop = 1e-10. Stability criteria was used to identify a stable number of cells in the model. Using 3 replicates of 80% of the input data, deconvolution is

performed modeling 3 to 12 cell types. The estimated proportions from the three replicates are compared for each cell type model and the number of cell types is determined once the correlations are no longer significant. Stability for this model resulted in 9 cell types.

XDec-SM Stage 2 was performed for three distinct analyses. First, TCGA breast cancer samples were subset into five groups based on the predominant epithelial proportion (epithelial 1 (Basal) [n = 133], epithelial 2 (Normal Control) [n = 107], epithelial 3 (HER2) [n = 176], epithelial 4 (Luminal) [n = 510], epithelial 5 (Other) [n = 30]). Proportions of the nine deconvoluted profiles were combined into three main cell types: epithelial (epithelial 1 through 5), stromal (stromal 1 and stromal 2), and immune (T cell and macrophage).

Second, XDec-SM Stage 2 was performed on each of the PAM50 breast cancer subtypes. This resulted in 6 cohorts: Controls [n = 95], Normal-like [n = 24], Basal-like [n = 142], Luminal A [n = 422], Luminal B [n = 194], HER2 [n = 67]. The proportions of the two stromal cell profiles (stromal 1 and stromal 2) were not combined in this analysis but the epithelial (epithelial 1 through 5) and immune (T cell, macrophage) cell profiles were combined respectively.

Lastly, XDec-SM Stage 2 was performed on each of the novel breast cancer subtypes with methylation data. This resulted in 4 groups: control [n = 73], Basal [n = 85], HER2 [n = 102], Luminal [n = 321]. The proportions of epithelial, immune, and stromal components were combined in this analysis.

Deconvolution methods

CIBERSORTx. Single cell gene expression RSEM counts (GSE118389_counts_rsem) were downloaded from GEO (GSE118389). Metadata provided by the study determined the cell type identity [6]. Using the CIBERSORTx web interface (https://cibersortx.stanford.edu), the "Create Signature Matrix" module (default parameters) was used to create the signature matrix for the 6 cell types: epithelial, stroma, endothelial, B cell, T cell, and macrophage. The TCGA BRCA dataset (IlluminaHiSeq_RNASeqV2) was downloaded from UCSC Xena (https://xena.ucsc.edu) and was uploaded as the "Mixture file". The "Impute Cell Fractions" module was run using default parameters using the previously defined Signature Matrix. For comparison to other methods, the three immune cell fractions (B cell, T cell, macrophage) are summed to represent immune and the two stromal cell fractions (stroma, endothelial) are summed to represent stromal.

Kassandra. Proportions for the Kassandra [19] deconvolution of TCGA breast cancer samples were obtained from BostonGene (https://science.bostongene.com/).

Epic. The proportions for the EPIC [20] deconvolution were also obtained from BostonGene for TCGA breast cancer samples. To obtain a general immune profile the following cell types were summed: B-cells, T-cells, macrophages and Natural Killer (NK) cells. CAFs and endothelial cells proportions were summed to constitute the stromal proportion. Additionally the epithelial cell type was defined as uncharacterized in this deconvolution [55].

xCell. xCell [21] proportions were obtained using the TIMER tool [23] and downloaded from the (http://timer.cistrome.org). General stromal, immune, and epithelial proportions were generated by TIMER.

InfiniumPurify. The TCGA BRCA proportions were previously estimated [56] and downloaded from Zenodo (https://zenodo.org/record/253193#.Xr1_xy-z2uV). Only TCGA samples with both XDec-SM proportions and InfiniumPurify tumor purity were compared.

Timer. The TCGA BRCA proportions were previously estimated [23] and downloaded (http://timer.cistrome.org). Only TCGA samples with both XDec-SM proportions and TIMER immune scores were compared. TIMER estimated the immune score for 6 immune cell subtypes (B_cell, CD4_Tcell, CD8_Tcell, Neutrophil, Macrophage, Dendritic) and all scores were summed for comparison.

Epigenomic deconvolution (EDec). The TCGA BRCA proportions were previously estimated [16] (and were downloaded from Genboree (http://genboree.org/theCommons/projects/edec). Only TCGA samples with both XDec-SM and EDec estimated proportions were compared. The six epithelial profiles (cancerous epithelial 1–5, normal epithelial) are summed to represent the epithelial cell type proportions.

Mapping of samples onto the cancer cell state map

After the deconvolution of all samples (TCGA, PDXs, cell line, single cell profiles), only samples with a combined epithelial proportion of > 0.7 and an epithelial Basal + epithelial HER2 + epithelial Luminal proportion of > 0.1 were selected to ensure that these samples were well modeled by the three epithelial profiles. The samples were then placed on a cancer cell state map representing their cancer cell state as a combination of the epithelial Basal, epithelial, HER2 and epithelial Luminal profiles. To visualize the simplex, the proportions of the sum of these three profiles were normalized to equal 1. The ggtern package was utilized to visualize the map [57].

Association of tumor microenvironment and cancer cell profiles

After the deconvolution of TCGA samples with XDec-SM, a correlation analysis was done to determine the relationship between the proportion of CAFs and the proportion of the epithelial HER2 profile. Likewise, a correlation was done between macrophage proportion and the proportion of the epithelial Basal and epithelial Luminal profiles.

Classification of TCGA and cell lines

For the TCGA XDec-SM classification, the cancer cell profile (epithelial Luminal, epithelial Basal, epithelial HER2) with the highest proportion was used to reclassify the sample. The TCGA annotations were used to determine the PAM50 classification of the TCGA tumors.

For the cell line XDec-SM classification, the BRCA cell lines from the Cancer Cell Line Encyclopedia were self-self-clustered with the epithelial profiles obtained from the deconvolution of TCGA over the PAM50 genes. Four distinct groups of cell lines were identified that matched the *in vivo* epithelial profiles. The PAM50 classifications were obtained from previous literature [45].

Tamoxifen survival analysis

The recount package was used to determine the treatment of TCGA models. 72 tumors were identified that were treated with tamoxifen and classified as Luminal by PAM50. We then compared the tumors that were also classified as Luminal by XDec-SM vs those that were classified as HER2 by XDec-SM using the survival and survminer packages in R.

PDX datasets

BCM cohort: A set of 50 TNBC PDX models were obtained from the BCM PDX collection, which are available on the PDX portal (https://pdxportal.research.bcm.edu/). These models were treated with four cycles of human equivalent docetaxel (20mg/kg, IP), carboplatin (50mg/kg, IP) and the control models were untreated for four weeks, and response was measured quantitatively as the change in tumor volume from baseline. Deep RNA-seq was also obtained for all models in this collection (~200M reads/sample). The murine vs human reads were separated by Xenome and the deconvolution was done on the human reads. Of the 50 models, 44 models that didn't model the same patient had a combined epithelial proportion

of > 0.7 and an epithelial Basal + epithelial HER2 + epithelial Luminal proportion of > 0.1 and were used for the response predictions.

RMGCRC cohort: The RNA sequencing and response to therapy for 30 PDX models from the RMGCRC were obtained from GEO (GSE142767). The response for these models was classified according to the RECIST criteria. Of these models, 22 had a combined epithelial proportion of > 0.7 and an epithelial Basal + epithelial HER2 + epithelial Luminal proportion of > 0.1 and were used for the response predictions.

Response to chemotherapy

PDX: A GLM model was built on the BCM dataset for carboplatin and docetaxel using all models from the epithelial Basal, epithelial Luminal, and epithelial HER2 proportions. These GLMs were then applied to the RMGCRC dataset for drug treatments in the same class (cisplatin and carboplatin, paclitaxel, and docetaxel) and an ROC was calculated for qualitative response.

Supporting information

S1 Fig. XDec-SM Deconvolution of RNA-seq profiles of TCGA breast tumor samples utilizing scRNA-seq references, related to Figs 1–3. (A) Heatmap representing the transformed gene expression counts of the 274 informative genes across the pseudo bulk reference expression profiles generated from the scRNA-seq gene expression profiles. (B) Heatmap representing the correlation between the XDec-SM estimated expression profiles (n = 9) and the pseudo bulk reference expression profiles. Red boxes are placed over the highest correlation. XDec-SM estimates five epithelial profiles, two stromal profiles, one T cell profile, and one macrophage profile. (C) Heatmap representing the per-sample proportion of the nine constituent cell types in the TCGA BRCA dataset. Top color bar represents the PAM50 expression subtypes. Only samples with identified subtypes are included in the heatmap. (D) Boxplot representing the per-sample proportion of the nine constituent cell types in the TCGA BRCA dataset separated by the PAM50 classification subtypes.
(PDF)

S2 Fig. Cell type specific gene expression. The cell type specific gene expression of marker genes across different breast cancer subtypes. ESR1 and FOXA1 are most highly expressed in the epithelial compartment. ADIPOQ and FABP4 are most highly expressed in the stromal adipocyte compartment. FN1 and COL1A1 are most highly expressed in the stromal CAF compartment. CD3G and CD8A are most highly expressed in the immune compartment.
(PDF)

S3 Fig. Distribution of scRNA-seq profiles for individual tumors, related to Fig 4. (A-H) Single cell RNA-seq and bulk RNA seq for the same tumor plotted on the cancer cell state map. to visualize the single cell and pooled data for this sample. The pooled sample is composed of ~1X105 cells, while the single cell samples are individual cells.
(PDF)

S4 Fig. Survival curves for patients treated with Tamoxifen, related to Fig 7. (A) Survival curve showing no significant survival difference between patients classified as Luminal by PAM50 and HER2 negative (blue) and Luminal by PAM50 and HER2 positive (pink). (B) Survival curve showing no significant survival difference between patients classified as Luminal by PAM50 and with high ERBB2 pathway activation (blue) and Luminal by PAM50 and without high ERBB2 pathway activation (pink).
(PDF)

Author Contributions

Conceptualization: Oscar D. Murillo, Varduhi Petrosyan, Michael T. Lewis, Aleksandar Milosavljevic.

Data curation: Oscar D. Murillo, Varduhi Petrosyan, Emily L. LaPlante.

Formal analysis: Oscar D. Murillo, Varduhi Petrosyan, Aleksandar Milosavljevic.

Funding acquisition: Michael T. Lewis, Aleksandar Milosavljevic.

Investigation: Oscar D. Murillo, Varduhi Petrosyan, Emily L. LaPlante, Aleksandar Milosavljevic.

Methodology: Oscar D. Murillo, Varduhi Petrosyan, Emily L. LaPlante, Lacey E. Dobrolecki, Aleksandar Milosavljevic.

Project administration: Aleksandar Milosavljevic.

Resources: Lacey E. Dobrolecki, Michael T. Lewis, Aleksandar Milosavljevic.

Software: Oscar D. Murillo, Varduhi Petrosyan, Emily L. LaPlante.

Supervision: Aleksandar Milosavljevic.

Validation: Oscar D. Murillo, Varduhi Petrosyan.

Visualization: Oscar D. Murillo, Varduhi Petrosyan, Aleksandar Milosavljevic.

Writing – original draft: Oscar D. Murillo, Aleksandar Milosavljevic.

References

1. Liu Z, Li M, Jiang Z, Wang X. A Comprehensive Immunologic Portrait of Triple-Negative Breast Cancer. Transl Oncol. 2018; 11: 311–329. https://doi.org/10.1016/j.tranon.2018.01.011 PMID: 29413765
2. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020; 578: 82–93. https://doi.org/10.1038/s41586-020-1969-6 PMID: 32025007
3. Beca F, Polyak K. Novel Biomarkers in the Continuum of Breast Cancer. Adv Exp Med Biol. 2016; 882: 169–189. https://doi.org/10.1007/978-3-319-22909-6_7
4. Place AE, Huh SJ, Polyak K. The microenvironment in breast cancer progression: biology and implications for treatment. Breast Cancer Res. 2011; 13: 227. https://doi.org/10.1186/bcr2912 PMID: 22078026
5. González-Silva L, Quevedo L, Varela I. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. Trends Cancer. 2020; 6: 13–19. https://doi.org/10.1016/j.trecan.2019.11.010 PMID: 31952776
6. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun. 2018; 9: 3588. https://doi.org/10.1038/s41467-018-06052-0 PMID: 30181541
7. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344: 1396–1401. https://doi.org/10.1126/science.1254257 PMID: 24925914
8. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017; 171: 1611–1624.e24. https://doi.org/10.1016/j.cell.2017.10.044 PMID: 29198524
9. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers Genetics. 2019; 10: 317. https://doi.org/10.3389/fgene.2019.00317 PMID: 31024627
10. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017; 9: 75. https://doi.org/10.1186/s13073-017-0467-4 PMID: 28821273
11. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016; 17: 29. https://doi.org/10.1186/s13059-016-0888-1 PMID: 26887813

12. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. Mol Cell. 2015; 58: 610–620. https://doi.org/10.1016/j.molcel.2015.04.005 PMID: 26000846

13. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019; 37: 773–782. https://doi.org/10.1038/s41587-019-0114-2 PMID: 31061481

14. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019; 10: 380. https://doi.org/10.1038/s41467-018-08023-x PMID: 30670690

15. Luca BA, Steen CB, Matusiak M, Azizi A, Varma S, Zhu C, et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell. 2021; 184: 5482–5496.e28. https://doi.org/10.1016/j.cell.2021.09.014 PMID: 34597583

16. Onuchic V, Hartmaier RJ, Boone DN, Samuels ML, Patel RY, White WM, et al. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. Cell Reports. 2016; 17: 2075–2086. https://doi.org/10.1016/j.celrep.2016.10.057 PMID: 27851969

17. Decamps C, Privé F, Bacher R, Jost D, Waguet A, Achard S, et al. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. Bmc Bioinformatics. 2020; 21: 16. https://doi.org/10.1186/s12859-019-3307-2 PMID: 31931698

18. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009; 27: 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370 PMID: 19204204

19. Zaitsev A, Chelushkin M, Dyikanov D, Cheremushkin I, Shpak B, Nomie K, et al. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. Cancer Cell. 2022; 40: 879–894.e16. https://doi.org/10.1016/j.ccell.2022.07.006 PMID: 35944503

20. Racle J, Gfeller D. EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data. Methods Mol Biology Clifton N J. 2020; 2120: 233–248. https://doi.org/10.1007/978-1-0716-0327-7_17 PMID: 32124324

21. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017; 18: 220. https://doi.org/10.1186/s13059-017-1349-1 PMID: 29141660

22. Zhang N, Wu H-J, Zhang W, Wang J, Wu H, Zheng X. Predicting tumor purity from methylation microarray data. Bioinformatics. 2015; 31: 3401–3405. https://doi.org/10.1093/bioinformatics/btv370 PMID: 26112293

23. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res. 2017; 77: e108–e110. https://doi.org/10.1158/0008-5472.CAN-17-0307 PMID: 29092952

24. Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017; 8: 15081. https://doi.org/10.1038/ncomms15081 PMID: 28474673

25. Kim C, Gao R, Sei E, Brandt R, Hatschek T, Crosetto N, et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell. 2018; 173: 879–893.e13. https://doi.org/10.1016/j.cell.2018.03.041 PMID: 29681456

26. Brechbuhl HM, Finlay-Schultz J, Yamamoto TM, Gillen AE, Cittelly DM, Tan A-C, et al. Fibroblast Subtypes Regulate Responsiveness of Luminal Breast Cancer to Estrogen. Clin Cancer Res. 2017; 23: 1710–1721. https://doi.org/10.1158/1078-0432.CCR-15-2851 PMID: 27702820

27. Cazet AS, Hui MN, Elsworth BL, Wu SZ, Roden D, Chan C-L, et al. Targeting stromal remodeling and cancer stem cell plasticity overcomes chemoresistance in triple negative breast cancer. Nat Commun. 2018; 9: 2897. https://doi.org/10.1038/s41467-018-05220-6 PMID: 30042390

28. Eiro N, Gonzalez LO, Fraile M, Cid S, Schneider J, Vizoso FJ. Breast Cancer Tumor Stroma: Cellular Components, Phenotypic Heterogeneity, Intercellular Communication, Prognostic Implications and Therapeutic Opportunities. Cancers. 2019; 11: 664. https://doi.org/10.3390/cancers11050664 PMID: 31086100

29. Hu G, Xu F, Zhong K, Wang S, Huang L, Chen W. Activated Tumor-infiltrating Fibroblasts Predict Worse Prognosis in Breast Cancer Patients. J Cancer. 2018; 9: 3736–3742. https://doi.org/10.7150/jca.28054 PMID: 30405845

30. Mao Y, Keller ET, Garfield DH, Shen K, Wang J. Stromal cells in tumor microenvironment and breast cancer. Cancer Metast Rev. 2013; 32: 303–315. https://doi.org/10.1007/s10555-012-9415-3 PMID: 23114846

31. Orimo A, Gupta PB, Sgroi DC, Arenzana-Seisdedos F, Delaunay T, Naeem R, et al. Stromal Fibroblasts Present in Invasive Human Breast Carcinomas Promote Tumor Growth and Angiogenesis through Elevated SDF-1/CXCL12 Secretion. Cell. 2005; 121: 335–348. https://doi.org/10.1016/j.cell.2005.02.034 PMID: 15882617

32. Plava J, Cihova M, Burikova M, Matuskova M, Kucerova L, Miklikova S. Recent advances in understanding tumor stroma-mediated chemoresistance in breast cancer. Mol Cancer. 2019; 18: 67. https://doi.org/10.1186/s12943-019-0960-z PMID: 30927930

33. Shiga K, Hara M, Nagasaki T, Sato T, Takahashi H, Takeyama H. Cancer-Associated Fibroblasts: Their Characteristics and Their Roles in Tumor Growth. Cancers. 2015; 7: 2443–58. https://doi.org/10.3390/cancers7040902 PMID: 26690480

34. McCuaig R, Wu F, Dunn J, Rao S, Dahlstrom JE. The biological and clinical significance of stromal-epithelial interactions in breast cancer. Pathology. 2017; 49: 133–140. https://doi.org/10.1016/j.pathol.2016.10.009 PMID: 28040198

35. Boyiadzis M, Whiteside TL. Information transfer by exosomes: A new frontier in hematologic malignancies. Blood Rev. 2015; 29: 281–290. https://doi.org/10.1016/j.blre.2015.01.004 PMID: 25686749

36. Harper J, Sainson RCA. Regulation of the anti-tumour immune response by cancer-associated fibroblasts. Semin Cancer Biol. 2014; 25: 69–77. https://doi.org/10.1016/j.semcancer.2013.12.005 PMID: 24406209

37. Popivanova BK, Kostadinova FI, Furuichi K, Shamekh MM, Kondo T, Wada T, et al. Blockade of a Chemokine, CCL2, Reduces Chronic Colitis-Associated Carcinogenesis in Mice. Cancer Res. 2009; 69: 7884–7892. https://doi.org/10.1158/0008-5472.CAN-09-1451 PMID: 19773434

38. Martinez-Outschoorn UE, Lisanti MP, Sotgia F. Catabolic cancer-associated fibroblasts transfer energy and biomass to anabolic cancer cells, fueling tumor growth. Semin Cancer Biol. 2014; 25: 47–60. https://doi.org/10.1016/j.semcancer.2014.01.005 PMID: 24486645

39. Martinez-Outschoorn UE, Sotgia F, Lisanti MP. Caveolae and signalling in cancer. Nat Rev Cancer. 2015; 15: 225–237. https://doi.org/10.1038/nrc3915 PMID: 25801618

40. Gwak JM, Jang MH, Kim DI, Seo AN, Park SY. Prognostic Value of Tumor-Associated Macrophages According to Histologic Locations and Hormone Receptor Status in Breast Cancer. Plos One. 2015; 10: e0125728. https://doi.org/10.1371/journal.pone.0125728 PMID: 25884955

41. Klingen TA, Chen Y, Aas H, Wik E, Akslen LA. Tumor-associated macrophages are strongly related to vascular invasion, non-luminal subtypes, and interval breast cancer. Hum Pathol. 2017; 69: 72–80. https://doi.org/10.1016/j.humpath.2017.09.001 PMID: 28923419

42. Prasmickaite L, Tenstad EM, Pettersen S, Jabeen S, Egeland EV, Nord S, et al. Basal-like breast cancer engages tumor-supportive macrophages via secreted factors induced by extracellular S100A4. Mol Oncol. 2018; 12: 1540–1558. https://doi.org/10.1002/1878-0261.12319 PMID: 29741811

43. Marti C, Sánchez-Méndez JI. Neoadjuvant endocrine therapy for luminal breast cancer treatment: a first-choice alternative in times of crisis, such as the COVID-19 pandemic. Ecancermedicalscience. 2020; 14: 1027. https://doi.org/10.3332/ecancer.2020.1027 PMID: 32368252

44. Raj-Kumar P-K, Liu J, Hooke JA, Kovatich AJ, Kvecher L, Shriver CD, et al. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. Sci Rep-uk. 2019; 9: 7956. https://doi.org/10.1038/s41598-019-44339-4 PMID: 31138829

45. Jiang G, Zhang S, Yazdanparast A, Li M, Pawar AV, Liu Y, et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. Bmc Genomics. 2016; 17: 525. https://doi.org/10.1186/s12864-016-2911-z PMID: 27556158

46. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019; 569: 503–508. https://doi.org/10.1038/s41586-019-1186-3 PMID: 31068700

47. Savage P, Pacis A, Kuasne H, Liu L, Lai D, Wan A, et al. Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. Commun Biology. 2020; 3: 310. https://doi.org/10.1038/s42003-020-1042-x PMID: 32546838

48. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017; 35: 319–321. https://doi.org/10.1038/nbt.3838 PMID: 28398307

49. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. Clin Cancer Res. 2010; 16: 5222–5232. https://doi.org/10.1158/1078-0432.CCR-10-1282 PMID: 20837693

50. Rolong A, Chen B, Lau KS. Deciphering the cancer microenvironment from bulk data with EcoTyper. Cell. 2021; 184: 5306–5308. https://doi.org/10.1016/j.cell.2021.09.028 PMID: 34653367

51. LaPlante EL, Liu D, Petrosyan V, Yao Q, Milosavljevic A. XDec-CHI reveals immunosuppressive interactions in pancreatic ductal adenocarcinoma. iScience. 2022; 25: 105249. https://doi.org/10.1016/j.isci.2022.105249 PMID: 36274954

52. Holliday DL, Speirs V. Choosing the right cell line for breast cancer research. Breast Cancer Res. 2011; 13: 215. https://doi.org/10.1186/bcr2889 PMID: 21884641

53. Murillo OD, Thistlethwaite W, Rozowsky J, Subramanian SL, Lucero R, Shah N, et al. exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids. Cell. 2019; 177: 463–477.e15. https://doi.org/10.1016/j.cell.2019.02.018 PMID: 30951672

54. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016; 17: 75. https://doi.org/10.1186/s13059-016-0947-7 PMID: 27122128

55. Racle J, Jonge K de, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017; 6: e26476. https://doi.org/10.7554/elife.26476 PMID: 29130882

56. Qin Y, Feng H, Chen M, Wu H, Zheng X. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. Genes Dis. 2018; 5: 43–45. https://doi.org/10.1016/j.gendis.2018.02.003 PMID: 30258934

57. Hamilton NE, Ferry M. ggtern: Ternary Diagrams Using ggplot2. J Stat Softw. 2018;87. https://doi.org/10.18637/jss.v087.c03