Single-cell imaging based prognosis prediction identifies new breast cancer survival subtypes

Shashank Yadav[1], Shu Zhou[1], Bing He[1], Lana X Garmire[1, #]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, MI, 48105, USA

[#]Corresponding author

Email addresses:

SY: shyadav@med.umich.edu

SZ: shuzh@med.umich.edu

BH: hbing@med.umich.edu

LXG: lgarmire@med.umich.edu

ABSTRACT

Quantitative models that explicitly capture single-cell resolution cell-cell interaction features to predict patient survival at population scale are currently missing. Here, we describe the first quantitative model that extracted hundreds of features describing single-cell based cell-cell interactions and cellular phenotypes from a large, published cohort of cyto-images of breast cancer patients. We applied these features to a neural-network based Cox-nnet survival model and obtained high accuracy in predicting patient survival in test data (Concordance Index > 0.8). We identified seven survival subtypes using the top survival features, which present distinct profiles of epithelial, immune, fibroblast cells, and their interactions. We identified atypical subpopulations of TNBC patients with moderate prognosis (GATA3 over-expression) and Liminal A patients with poor prognosis (KRT6 and ACTA2 over-expression and CDH1 under-expression). These atypical subpopulations are validated in TCGA-BRCA and METABRIC datasets. This work outlines the roadmap to integrate single-cell data for survival prediction at population scale.

INTRODUCTION

Breast cancer became the most commonly diagnosed cancer in 2020, with an estimated 2.3 million new cases globally (Sung et al., 2021). Predicting survival in breast cancer can aid clinicians in making prompt prognostic decisions and deciding the direction of treatment. The relevance of prognosis in oncology is projected to grow in the future as new prognostic indicators allow for more precise treatment therapies (Mackillop, 2003). However, insufficient knowledge about the intercellular interaction among the tumor and tumor microenvironment is a significant roadblock in applying personalized therapies. The breast cancer tumor ecosystems in breast cancer consist of neoplastic epithelial cells forming the tumor core, as the 'tumor microenvironment' composed of several types of immune cells, fibroblasts, adipocytes, and mesenchymal cells (McAllister & Weinberg, 2010; Turashvili & Brogi, 2017). These diverse cell types alter molecular and cellular programs and present dynamic spatial heterogeneity as the disease progresses. Such temporal and spatial changes are responsible for differential responses to anti-cancer therapies and subsequent clinical outcomes. Hence, it is essential to develop comprehensive understanding of breast cancer heterogeneity by elucidating the

contribution of tumor, tumor microenvironment, and their intricate interactions (Dagogo-Jack & Shaw, 2018; Lüönd, Tiede, & Christofori, 2021).

Investigations at single-cell resolution (Casasent et al., 2018; Yan et al., 2021) have enabled detailed elucidation of tumor-microenvironment interactions lately. Various components of tumor-immune-stromal relationships have been identified, and tumor heterogeneity was also analyzed to distinguish breast cancer subtypes based on epithelial and immune cell populations (Azizi et al., 2018; Chung et al., 2017; Karaayvaz et al., 2018). Potential biomarkers have also been discovered for personalized cancer immunotherapy through Single-cell RNA sequencing (scRNA-seq) (Ding, Chen, & Shen, 2020). However, one major limitation of scRNA-seq is the loss of spatial information crucial in understanding tumor heterogeneity *in situ* (Saviano, Henderson, & Baumert, 2020). To address these issues, spatially resolved assays, such as single-cell transcriptomic and proteomic techniques have been developed to study the distribution of cells in cancers (Baharlou, Canete, Cunningham, Harman, & Patrick, 2019). Recently, Imaging Mass Cytometry (IMC) based approaches have recently quantified tumor heterogeneity with spatial context and identified novel breast cancer molecular subtypes in large population cohorts (Ali et al., 2020; Jackson et al., 2020). However, most of these studies that detect molecular subtypes are unsupervised, without explicitly fitting phenotypes such as survival in the learning process. Rather, survival is used as a *post hoc* metric to evaluate the subtypes (Poirion, Jing, Chaudhary, Huang, & Garmire, 2021). Moreover, these unsupervised approaches can not be directly used to predict new patients, significantly limiting the practical utility of the subtype findings.

In this study, we asked whether quantifying single-cell level cell-cell interactions could provide meaningful insights for prognosis prediction for breast cancer patients. We computed single cell level imaging features that capture cell-cell interactions in breast cancer tissues from a previous single-cell imaging mass cytometry study on 259 breast cancer patients with survival data (Jackson et al., 2020). We applied a neural network-based method Cox-nnet previously developed in our group (Ching, Zhu, & Garmire, 2018) on these features to predict explicitly the patient survival outcome. Using the top survival features, we uncovered seven new single-cell interaction-based patient subtypes which show significantly better associations with survival compared to the current mainstream molecular subtypes in breast cancer. We characterized these survival subtypes with distinct profiles on cellular phenotypes as well as cell-cell interactions. Using the new survival subtyping classification, we identified two atypical patient subpopulations: a subgroup of triple negative breast cancer patients with moderate prognosis and a subgroup of Luminal A patients with

poor prognosis. We further utilized transfer learning approach and validated the presence of such unconventional subgroups and their biomarkers in the TCGA and METABRIC breast cancer datasets

RESULTS

Feature engineering from the breast cancer single-cell images

The breast cancer cohort contains 259 patients with survival data, as reported earlier (Jackson et al., 2020). The summary of the cohort's patients, including clinical variables such as tumor grades, clinical features (ER, PR, HER2), and clinicopathological classes, are shown in **Table S1.** We first extracted 27 pre-defined cellular-phenotype (CP) features based on image mass cytometry data (**Fig. 1A, Table S2**). These features describe epithelial, immune, and stromal cell phenotypes at the single-cell resolution in the original study (**Methods**). We next used graphs ("network of cells") to represent the cells and cellular communities with the spatial arrangement as shown in the imaging data. We then calculated phenograph neighborhood information based on the cellular phenotypes in the tissue (**Methods**). This process results in additional 378 cell-cell interaction features that can be broadly divided into two subsets. The first subset contains 273 features related to the tumor microenvironment interaction (TMI) features (**Fig. 1A**). The TMI features are computed from pairwise interactions among the three types of cells: immune, stromal, and epithelial cells. Specifically, they represent immune-immune (21 features), immune-stromal (42 features), immune-epithelial (84 features), stromal-stromal (28 features), and stromal-epithelial (98 features) interactions. The second subset contains 105 tumor core interaction (TCI) features, which are exclusively epithelial-epithelial interactions (**Fig. 1A**). The distributions of these 378 features in patients together with clinically defined breast cancer subtypes are illustrated in **Fig. 1B.** Anchoring on the TCI features, hierarchical clustering demonstrates broad but distinct cellular heterogeneity among patients, far more complex than that defined by the clinical subtypes. Compared to TCI features, TMI features describing immune-epithelial and stromal-epithelial interactions have similar but less distinct heterogeneity. On the other hand, all CP features show far less global correlations with those TCI patterns presented in epithelial-epithelial interactions.

Survival prediction using single-cell phenotype features

A major goal of our study is to identify single-cell level features associated with patient survival and further evaluate the relative contributions of these features towards patient survival. To this end, we used the recently developed Cox-nnet neural-network-based survival prediction models from our group (Ching, Zhu, & Garmire, 2018), which had shown advantages in single or multiple data modalities (Zhan et al., 2021) (Wang, Jing, He, & Garmire, 2021), compared to the conventional Cox-PH method (Cox, 1972). Here as a comparison to Cox-nnet models, we built a Cox-PH model using the clinical information including ER status, PR status, HER2 status, and tumor grade (Set I), as the baseline model. We constructed a series of Cox-nnet (version 2) models, including one-stage Cox-nnet models (Fig. 2A) on the CP features (Set II), TMI features (Set III), TCI features (Set IV), and their combinations (Sets V, VI, VII, and VIII). We also constructed two-stage Cox-nnet models (Fig. 2B) on the feature combinations (Set IX, X, XI, and XII). Two-stage Cox-nnet models are complex models that use simpler Cox-nnet models as individual building blocks (Zhan et al., 2021). In the first-stage of training, each Cox-nnet model is built to fit a specific set of data (CP, TMI or TCI features) to predict survival. The hidden nodes in the first-stage Cox-nnet model are then combined as the input features to train a second stage Cox-nnet model. For each Cox-nnet model, we used L2 regularization to reduce overfitting.

To rank the relative contributions of different features, we estimated the relative importance of the features in each of the CP, TMI, and TCI feature sets (**Table S3-5)**. Among CP features, immune cells including type 2 macrophages (CD68+/vimentin low) and T & B immune cell clusters have the highest relative importance scores (1.000 and 0.692 respectively) compared to those (0.507-0.631) of different subtypes of epithelial cells (**Table S3**), highlighting their significance in patient prognosis. In TMI features, two types of immune-epithelial interactions (T and B cells - $CK^{low}HR^{low}$ epithelial cells, $Macrophage_1$ - $CK^{low}HR^{hi}p53^{+}$) and a type of fibroblast - epithelial interaction (small elongated fibroblast - $CK7^{+}CK^{+}$ epithelial cells) and are top 3 dominant features (**Table S4**). Among the TCI features, the interactions among most proliferative epithelial cells are the strongest, as expected (**Table S5**).

Identifying survival subtypes among breast cancer patients

Our next goal was to identify patient subpopulations associated with survival, using the top survival features selected by importance scores from the best model using Set XII (the combination of CP, TMI, and TCI features). We performed Non-negative Matrix Factorization (NMF) based consensus clustering on the top fifty features according to the importance scores (**Methods**). As a result, we identified seven optimal patient subpopulation clusters, indexed by 1-7 from the best to the worst survival risks (**Fig. 3A).** We confirmed that seven clusters are the optimum value, based on

Cophenetic score and Silhouette coefficient, two metrics of clustering accuracy (**Fig. 3B**). The Kaplan-Meier plot for the overall survival of the seven survival subtypes yields a much higher C-index of 0.80 and a more significant log-rank p-value $p = 8.6e^{-0.6}$(**Fig. 3C**), compared to the C-index of 0.63 and the log-rank p-value of 0.001 from the stratification of the four molecular subtypes (**Fig. 3D**). These results demonstrate that the single-cell level CP, TMI, and TCI features yield more informative subtypes that better reveal the heterogeneity in survivorship among patients.

We further describe the survival subtypes based on their top features as well as their relationships with other clinical information (**Fig. 3E, Table S6).** As expected, the better survival subtypes tend to have more ER+ and PR+ cases, and the worst survival subtypes tend to have higher tumor grade (Grade III) and HER2- cases. Some emerging patterns on the molecular features are clearly present in the survival subtypes, supported by the single cell IMC images (**Supplementary Fig. 1**). The best survival subtype 1 is enriched with a subtype of epithelial cells with high levels of cytokeratin (CK) 7 and 19 but low hormone receptors (CK7, CK19, low HR). It also has high levels of interaction scores between this epithelial cell subtype and several subtypes of fibroblast cells, highlighting the importance of the interaction between these two cell types for patient outcome. The next best subtype 2 also has enriched scores on a subtype of epithelial cells that express pan-cytokeratins and hormone receptors (pan-CK, HR), as well as their interactions with certain fibroblast cells. On the other hand, the subtype 7 with the worst survival is characterized by high degrees of interactions among proliferative epithelial cells, as well as interactions between macrophages/T-cells expressinging high vimentin and the proliferative epithelial cells. The second worst survival subtype 6 has a high level of epithelial cells lacking CK and HR expression, as well as strong interactions between fibroblast/B-cells and these epithelial cells. In summary, the survival subtypes show distinct profiles of epithelial cell subtypes and interactions between the epithelial cells and adjacent immune and fibroblast cell subtypes.

Characterization of the new survival subtypes

We next directly compared the enrichment of different types of immune, fibroblast and epithelial cells, as well as their interaction scores in the seven survival subtypes (**Fig 4**). Subtypes 5, 6, and 7 show quite distinct patterns from subtype 1-4. In particular, subtype 5 has a high level of hypoxic epithelial cells (**Fig 4E**), as well as high levels of interactions between these hypoxic epithelial cells and macrophage$_2$, T/B cells and vimentin expressing fibroblasts (**Fig. 4I, 4K and 4O**). Subtype 6 has the second highest level of proliferative epithelial cells (**Fig. 4F**) but the lowest levels of vimentin expression fibroblast cells (**Fig. 4C**), and accordingly the second highest scores in interactions between T/B cell and

proliferative epithelial cells (**Fig. 4L**). Subtype 7 stands out with the highest level of proliferative epithelial cells (**Fig. 4F**) and second highest level of hypoxic epithelial cells (**Fig. 4E**). As a result, subtype 7 has the highest scores in the most number of TMI-tumor interaction categories **(Fig. 4J-L, 4P-R**). In summary, the subtypes 5-7 with some of the worst survival outcomes are highly enriched with hypoxic (subtype 5) or proliferative epithelial cells (subtype 6 and 7), as well as the corresponding interactions between immune cells and these epithelial cells; however, they have much lower interactions between fibroblast and immune cells in the tumor microenvironment (**Fig. 4G-H, 4M-N**).

Next, we investigated the higher order of correlation relationship among the cell-cell interactions for each survival subtype. We calculated the correlations among all the CP, TMI and TCI features and studie the pairs of cell-cell interaction with correlations greater than 0.5 (**Fig. 4S**). For subtype 1, a high correlation exists between epithelial cell-small, elongated fibroblast interaction and epithelial cell-small circular fibroblast interaction, indicating that they share similar modes of interactions. Similarly, subtype 3 shows a high correlation between macrophage-small elongated fibroblast interaction and macrophage-small circular fibroblast interaction**.** In subtype 7, proliferative epithelial-macrophage/T-Cell (vimentin-expressing) interaction is highly correlated with interactions between vimentin-expressing macrophage and endothelial cells**,** further confirming the detrimental and synergistic effect of certain immune cells on proliferative tumor.

Discovery of subpopulations of TNBC and luminal A subtypes with atypical survival outcome

To uncover the novel insights of the new survival subtyping, we compared the classification results using molecular subtyping vs. the new survival subtyping approach (**Fig 5A).** Each molecular subtype is split into multiple survival subtypes, demonstrating the wide-spread heterogeneity in terms of survival. Two atypical subpopulations within liminal A and TNBC molecular subtypes are well noticed. Unlike the conventional concept that liminal A molecular subtype has good prognosis, a significant proportion (50%) of luminal A breast cancers have poor survival, distributing in subtypes 4-7 where subtype 4 is the predominant cluster (78%). On the other hand, TNBC the molecular subtype regarded as having the worst prognosis, has a large proportion (55%) of atypical subgroups with moderate survival in clusters 1-3, where cluster 3 is the major cluster (75%). We investigated the signatures of these atypical subpopulations by differential expression analysis. The atypical “good survival” subpopulation in TNBC has over-expression of GATA3 (**Fig. 5B).** On the other hand, the atypical “poor survival” subpopulation in the luminal A subtype has over-expression of KRT7 and ACTA2, and under-expression of CDH1 (**Fig. 5E**).

We next tested if such atypical subpopulations can be validated in general. We used the UNION-COM (Cao, Bai, Hong, & Wan, 2020) method to perform label-transfer learning based on the pseudo-bulk protein expression in this single cell dataset, and applied the model to the transcripts of the same genes in TCGA-BRCA and METABRIC data. The results show that TCGA-BRCA and METABRIC datasets indeed have such "good survival" subpopulations in TNBC (**Fig. 5C-D)** and "poor survival" subpopulations in luminal A patients **(Fig. 5F-G).** The atypical subpopulations in both validation cohorts have significantly different survival curves (log-rank p-value < 0.05) compared to their counterparts. Moreover, GATA3 is also consistently overexpressed in the atypical "good survival" subpopulations in both the TCGA-BRCA and METABRIC datasets (**Fig. 5C-D**). Similarly, the matching "poor survival" liminal A subpopulations in TCGA-BRCA and METABRIC datasets show the same patterns of over-expression of KRT7 and ACTA2 and under-expression of CDH1 (**Fig. 5F-G**).

DISCUSSION

Breast cancer is a highly heterogeneous disease and molecular subtypes based on ER, PR and HER2 statuses are currently the mainstream classification system. In this study, we leverage the strengths of detailed single-cell pathology images and neural-network based prognosis modeling and define a new class of survival related subtypes for breast cancer. We argue that cellular phenotypes and their interactions provide additional valuable information to predict survival, leading to higher clinical impacts.

The uniqueness of this study lies in several aspects. From the analytical aspect, we explicitly computed hundreds of features describing cell-cell interactions (TMI and TCI) on single-cell imaging data. We used these features as inputs for a novel neural-network based survival prediction method called 2-stage Cox-nnet, which highly accurately fits patient survival and is much better than clinical data based Cox-PH regression. From the biomedical aspect, we have defined seven new survival subtypes with distinct profiles of epithelial, immune, and fibroblast cells as well as their interaction patterns. These new survival subtypes re-classify the molecular subtypes, each of which is highly heterogeneous in terms of patients prognosis. Using multiple population cohorts, we verified that there exist "good survival" subpopulations within TNBC patients and "poor survival" subpopulations in luminal A patients, and that the molecular signatures of these atypical subpopulations are robustly consistent.

These survival subpopulations are well characterized by protein markers and cellular phenotypes. The relatively good survival subtypes 1-4 are enriched with CK or HR expression epithelial cells as well as interactions between fibroblasts and other immune cell types, which are lacking in the poorer survival subtypes 5-7. The best survival subtype 1 has a high expression level of luminal CK. Corresponding to this observation, CKs were reported to be associated with better overall survival in breast cancer before (Lu, Yakirevich, Wang, Resnick, & Wang, 2019). Subtype 2, which also has good survival, has high levels of CK8/18 and HR, which was also reported to be associated with good overall survival (Menz et al., 2021). Subtype 5 has a dominant hypoxia phenotype, demonstrated by the high presence of hypoxia epithelial cells. Hypoxia is associated with resistance against therapies and poor outcomes (Jögi, Ehinger, Hartman, & Alkner, 2019; Mimeault & Batra, 2013). Despite most of the patients in subtype 6 are luminal A subtype, patients in this subgroup show the lowest level of Vimentin^{hi} Fibroblasts cells (high vimentin expression, low smooth muscle actin (SMA) expression, and low fibronectin expression). It also has the 2nd highest level of proliferative epithelial cells, both of which may contribute to poor survival(Pellegrino et al. 2021). Subtype 7 has the highest level of proliferative epithelial cells marked by KI-67 expression, associated with poor survival (Kanyılmaz et al., 2019; Yerushalmi, Woods, Ravdin, Hayes, & Gelmon, 2010). It also has some of the strongest interactions with cells in EMT, suggesting the highest degree of tumor infiltration by immune cells (Liu, Li, Jiang, & Wang, 2018).

The highly active TMI observed in hypoxic subtype 5 and proliferative subtype 7 are quite interesting. Subtype 5 has significantly lower proportions of immune cells (eg. macrophages) but the strongest interactions between macrophages and the hypoxic epithelial cells, the signature epithelial cells of this subtype. It also has the highest interaction between Vimentin^{hi} fibroblasts and hypoxic epithelial cells. The apparent paradox between low macrophage concentration vs high interaction with epithelial cells may be explained by the inflammatory cytokines produced by macrophages at the hypoxic site, which decreases tumoricidal activity (Saccani et al., 2006). Subtype 7 has the highest proportion of proliferative epithelial cells and 75% of the patients are TNBC. This subtype has the strongest interactions between various cells in tumor microenvironment (vimentin expression T/B cells, fibroblasts, and macrophages2) and proliferative epithelial cells, as well interactions between vimentin-expressing macrophages and hypoxic cells. A high degree of immune infiltration was reported in TNBC patients with exhausted T Cell populations (H. Zhang, Qin, Yu, Han, & Zhu, 2021), consistent with our results. This subtype also shows the highest interaction between hypoxic epithelial cells and proliferative epithelial cells. The hypoxic condition may be a result of the increased aggressiveness

of proliferating epithelial cells(Keith and Celeste Simon 2007), consistent with prior observation of a close relationship between proliferation-hypoxia in other solid cancers such as squamous cell carcinoma (Kennedy et al. 1997).

Contrary to the general perception that liminal A cancers have good prognosis and TNBC cancers have poor prognosis, the new subtyping system enabled the discovery of atypical survival subpopulations. A "good survival" subpopulation among TNBC patients (mostly from survival subtype 3) are identified with over-expression of GATA3; and a "poor survival" subtype among luminal A patients are found, with over expression of KRT7 and ACTA2 but under expression of CDH1. These results are robustly verified in TCGA-BRCA and METABRIC datasets. GATA3 is a zinc finger transcription factor involved in cell type differentiation and proliferation. Corresponding to our observation, higher levels of GATA3 were also shown to be correlated with better survival in breast cancer patients (Yoon et al. 2010; Mehra et al. 2005). KRT7 encodes cytokeratin-7 and ACTA2 encodes smooth muscle alpha-2 actin, both are cytoskeleton components. KRT7 plays a role in cell migration and EMT pathways and is associated with poor survival subpopulations for ovarian cancer (An et al., 2021; Communal et al., 2021). Similarly, ACTA2 was reported to be a marker for poor prognosis in lung cancer (Lee et al. 2013). CDH1 is a transmembrane glycoprotein that is primarily responsible for cell adhesion, and its downregulation in tumors led to increased invasiveness in breast cancer and lung cancer (Oka et al., 1993). The work here paved the foundation for improving subtyping of TNBC and luminal A cancers using above mentioned biomarkers. Moreover, therapeutics targeting these molecules may be effective at improving TNBC and luminal A cancer patients' survival.

In summary, the imaging cytometry data at the single cell resolution has enabled an unprecedented opportunity to explicitly study cells in the tumor and tumor microenvironment, as well as their interactions. By taking the advantages of CP, TMI and TCI features, novel survival subtypes are identified in breast cancers, each with distinct profiles and more molecular and survival homogeneity. Moreover, this new subtyping system allows to identify good survival subpopulations in TNBC and bad survival subpopulations in luminal A cancers, with robust biomarkers in multiple population cohorts. The work lays down the foundation to integrate single-cell resolution information for survival prediction at large population scale. It has far-reaching impacts to transform breast cancer prognosis prediction and also likely treatment options in the future.

METHODS

Dataset and extracted features

The dataset analyzed in the study is obtained from a previously published study containing 259 patients (Jackson et al., 2020). The dataset contains the image mass cytometry (IMC) data, phenograph neighborhood information, clinical features (Tumor Grade, ER Status, PR Status and HER2 Status), and patient prognosis outcome (overall survival time, disease-free survival time, and alive/dead status). The previous work defined 27 cellular phenotypes that described the histopathological landscape of breast cancer. In this study, we used the clinical features, cellular phenotype density (count of cells of each phenotype per unit area), and cellular neighborhood information for analysis, as detailed below:

Clinical Features: The clinical features set comprises the Tumor Grade, ER Status, PR Status, and HER2 Status for each patient. Tumors in which cells appear highly dissimilar to normal cells tend to proliferate, and the tumor grade is assigned based on the extent of proliferation. In the traditional clinicopathological classification of breast cancer, patients are classified into Luminal A, Luminal B, Triple Negative, and HE2-Enriched classes. This classification is based on patients' clinical features such as ER (positive/negative), PR (positive/negative), and HER2 (positive/negative) status (Perou et al., 2000; Sørlie et al., 2001).

Cellular phenotypic features: For each patient, the cell phenotype density is quantified as the counts of each cellular phenotype per unit area in the IMC image of the tumor tissue. Out of the original 27 cellular phenotypes, there are six immune, seven stromal, and fourteen epithelial cellular phenotypes. The six immune phenotypes are B Cell, T and B Cell, T-Cell$_1$, Macrophage$_1$, T-Cell$_2$, Macrophage$_2$. The seven stroma phenotypes are Endothelial, Vimentin^{hi} Fibroblasts, Small circular Fibroblasts, Small elongated Fibroblasts, Fibronectin^{hi} Fibroblasts, Large Elongated Fibroblasts, SMA^{hi}-Vimentin^{hi} Fibroblasts. The fourteen epithelial cellular phenotypes are Hypoxic Epithelial Cells, Apoptotic Epithelial Cells, Proliferative Epithelial Cells, p53^{+} EGFR^{+} Epithelial Cells, Basal CK Epithelial Cells, CK7^{+}CK^{hi}Cadherin^{hi} Epithelial Cells, CK7^{+}CK^{+} Epithelial Cells, Epithelial^{low} Epithelial Cells, CK^{low}HR^{low} Epithelial Cells, CK^{+}HR^{hi} Epithelial Cells, CK^{+}HR^{+} Epithelial Cells, CK^{+}HR^{low} Epithelial Cells, CK^{low}HR^{hi}p53^{+} Epithelial Cells, Myoepithelial Cells. The annotations are shown in **Table S2**.

Cell-Cell Interaction Features: The phenotypic features in the original report are limited, as they do not assess the interactions between cellular phenotypes and between tumor-tumor microenvironment, which are important parts of the tissue heterogeneity. We utilized the available phenograph (Levine et al., 2015) neighborhood-information data from each patient and quantified the binary interactions between cellular phenotypes. The phenograph neighborhood of the

IMC image is described as a numerous cellular network spread out in the mass cytometry image, where individual cells are represented as nodes of the cellular network. Starting with 27 cellular phenotypes, we calculated 378 pair-wise phenotype-phenotype features.

We iterate through each cell of a cellular community in the image, enumerate the binary interactions the cell makes with its neighbors, and multiply the sum by the clustering coefficient C of the cellular community. Then, we repeat this process for all the cellular communities across the mass cytometry image, sum the 378 interactions, and divide them by the area occupied by all the cells.

We define the result of this procedure as the 'Cell-Cell Interaction Score' (CCIS).

$$CCIS_{\forall\, Pair(x,y), x\in P, y\in P} = \frac{\sum_{\forall\, N} (C * B_{xy})}{A} \quad \textbf{(Eq.1)}$$

where C is the clustering coefficient for each cluster, B_{xy} is the binary interaction counts between a particular pair of phenotypes x and y, A is the total area occupied by all the cells in a mass cytometry image, P is the set of all phenotypes and N is the number of cellular communities in an image.

At more detailed level, these 378 can be further categorized by immune/stromal/epithelial cell types in the interacting pairs and classified as immune-immune (21 features), immune-stromal (42 feature), immune-epithelial (84 features), stromal-stromal (28 features), stromal-epithelial (98 features), epithelial-epithelial (105 features) interactions. We plot the feature heatmaps using the `heatmap` package in R.

Survival modeling

We use a variety of neural-network based Cox-nnet models as the one-stage Cox-nnet-v2 models for the phenotypic feature set, tumor-microenvironment feature set, tumor-core feature set, and pairwise combinations. Further, we develop a two-stage Cox-nnet model by combining the hidden layer output of previously trained one-stage models. For each one-stage Cox-nnet-v2 model, the hidden layer nodes are equal to the square root (rounded up) of input nodes. The Cellular Phenotypic, TMI, and TCI feature-based models have 27, 268, and 105 input features, respectively, and hence we obtain 6, 17, and 11 nodes in the respective hidden layer. The two-stage Cox-nnet model combines the hidden layer output features. As a result 34 input features are used in the second stage. For comparison, we use the Cox-PH model on clinical features as the baseline model.

Survival feature ranking

We calculate the feature importance of the 34 input features using the 'Variable Importance' function of the Cox-nnet-v2 model. Out of the 34 input features in the two-stage Cox-nnet model, we calculate the importance scores, separate them into the original three sets of 6, 17, and 11 features respectively, and associate them with the one-stage Cox-nnet-v2 models. For each model, we took a dot product of the hidden layer feature-importance vector with the model weights to get the importance scores vector of the original features.

$$\hat{O} = W^T . \hat{H}$$ (**Eq. 2**),

where $\hat{O}$ is the importance vector of the original features (shape I).

W is the model weight matrix (shape $h\ x\ p$).

H is the importance vector for hidden layer (shape $h\ x\ 1$).

Unsupervised analysis for patient subtype detection

We normalize the feature importance values between 0 and 1 and select the top features from each of the three feature sets. For the phenotypic set, we set the threshold as 0.5 and selected all the features with a higher importance score. For the tumor-microenvironment and tumor-core feature sets, we select features with an importance score greater than 0.75. In total, out of the 400 features we choose 50 features and perform NMF based consensus clustering using the `NMF` R package v0.23.0. This technique has been used in molecular subtype detection in cancer (Frigyesi & Höglund, 2008; Ma et al., 2019). We carry out a hyperparameter search for the NMF rank and varied it from 3 to 15, finding that the maximum values of 'cophenetic scores' and 'silhouette coefficients' are reached at an NMF rank of seven, and hence we choose optimum NMF rank as seven for subsequent analysis.

Feature correlation analysis and comparison between NMF-defined and clinically defined subtypes

We calculate the correlations between the features associated with each NMF-defined subtype using Spearman's Correlation Coefficient in R and plot the Circos plot of the correlation using the `circlize` package in R (Gu, Gu, Eils, Schlesner, & Brors, 2014). How the NMF-defined classes intersect with the clinicopathological classification is determined by a Sankey Plot. We plot the Sankey plot using the `plotly` package in Python.

Class label transfer and comparison with TCGA-BRCA and METABRIC datasets

We verify our results for 'TNBC - Good Survival' and 'Luminal A - Poor Survival' patient subpopulation in two external datasets TCGA-BRCA and METABRIC (Cancer Genome Atlas, 2012; Curtis et al., 2012). Here, we utilize the mass cytometry counts for 30 protein-based biomarkers and calculate the average expression of each biomarker over all the cells present in the IMC image. We treat it as pseudo-bulk protein expression and a proxy for bulk mRNA expression to facilitate the comparison with bulk mRNA expression-based external datasets. Then, we repurpose the UNION-COM algorithm (Cao et al., 2020) to perform patient matching between our dataset, the TCGA-BRCA dataset, and the METABRIC dataset. The UNION-COM method was originally developed to perform topological alignment and label transfer on single-cell multi-omics datasets. It takes two expression matrices as input, calculates the joint embedding between the two datasets, and maps samples from one dataset to another. In the TCGA-BRCA dataset, we find 28 genes corresponding to 28 protein-based biomarkers out of 30 (except TWIST and mTOR). In the METABRIC dataset, we find all 30 genes corresponding to the 30 protein-based biomarkers. For each TNBC and Luminal A subpopulations, the pseudo bulk protein expression matrix from the single cell dataset is used as input, and the mRNA expression matrix is applied as the external dataset in the UNION-COM method. We run the UNION-COM method using default parameters, and it returns the pairwise distance between patients between our dataset and the external dataset. We set a 99% similarity cutoff and select patients matching our 'TNBC - Good Survival' and 'Luminal A - Poor Survival' patient subpopulations, respectively. Survival plots are made by the `lifelines` Python package, and differential gene expressions are done using the `limma` package in R (Ritchie et al., 2015).

CODE AVAILABILITY: The source code to generate the figures and feature datasets for this work are available at https://github.com/lanagarmire/BC_imaging

AUTHORS' CONTRIBUTIONS

LG envisioned this project and supervised the study. SY performed the data analysis, generated the figures, and wrote the manuscript. SZ conducted drug reposition prediction, tested the code used in this study and made it available on the Github repository. BH helped with subgroup selection and interpretation.

ACKNOWLEDGEMENT

We thank Dr. Hartland Jackson for providing data for this project. This research was supported by grants R01 LM012373 and LM012907 awarded by NLM, and R01 HD084633 awarded by NICHD to L.X. Garmire.

FIGURE LEGENDS

Figure 1: Overview of different feature sets in the data:

(a) Methodology for extracting the cell-cell interaction features. We utilize the cellular neighborhood information obtained from Phenograph. The Phenograph result contains different locally connected cellular communities. For each cellular community (the neighborhood graph), we iterate through each cell(node) and count the number of interactions between the particular cell and its neighbors. We repeat this process for all the different cellular communities to assess the 378 different pairwise cell-cell interactions.

(b) Heatmaps illustrate the different feature types for all the patients. The patients are arranged in the order defined by the hierarchical clustering based on the epithelial-epithelial interactions.

Figure 2: Two-stage Cox-nnet model comparison with baseline and Cox-nnet-v2. (a) Model architecture of the Cox-nnet-v2 model for survival prediction based on a single data type. **(b)** Model architecture for the two-stage Cox-nnet model. Three individual Cox-nnet models were built for each data type. The hidden nodes from the first stage Cox-nnet models were combined to form the input to construct a new Cox-nnet model in the second stage. **(c)** Comparison of Concordance-index across different Feature Set-Model pairs.

Figure 3: NMF based subpopulation detection associated with survival

(a) NMF heatmap for 259 patients in our cohort illustrating the seven subpopulations arranged in order of decreasing survival. **(b)** Cophenetic Score and Silhouette Coefficient versus the NMF Rank. **(c)** Kaplan-Meier plots illustrate the Overall Survival for patients in the seven NMF-derived subpopulations clusters. **(d)** Kaplan-Meier plots illustrate the Overall Survival for patients based on clinicopathological classification**. (e)** Heatmap of top-ranked features from best performing two-stage Cox-nnet model, in associations with tumor grade, clinical features and molecular subtypes.

Figure 4. Characterization of the new single image based survival subtypes.

(A-R) Scoring and profiling for the seven survival subtypes based on various cellular phenotype and cell-cell interaction features. Cell counts and interactions were normalized between 0-1 to make comparison possible on the same scale. **(a)** $Macrophage_2$ Cells **(b)** T and B Cells **(c)** $Vimentin^{hi}$ Fibroblasts **(d)** Small Circular Fibroblasts

(e) Hypoxic Epithelial Cells **(f)** Proliferative Epithelial Cells **(g)** $Macrophage_2$ - $Vimentin^{hi}$ Fibroblasts **(h)** T and B Cells - $Vimentin^{hi}$ Fibroblasts **(i)** $Macrophage_2$ - Hypoxic Epithelial Cells **(j)** $Macrophage_2$ - Proliferative Epithelial Cells **(k)** T and B Cells - Hypoxic Epithelial Cells **(l)** T and B Cells - Proliferative Epithelial Cells **(m)** $Macrophage_2$ - Small Circular Fibroblasts **(n)** $Vimentin^{hi}$ Fibroblasts - Small Circular Fibroblasts **(o)** $Vimentin^{hi}$ Fibroblasts - Hypoxic Epithelial Cells **(p)** $Vimentin^{hi}$ Fibroblasts - Proliferative Epithelial Cells **(q)** Hypoxic Epithelial Cells - Proliferative Epithelial Cells **(r)** Proliferative Epithelial Cells - Proliferative Epithelial Cells. Statistical testing was done using Mann-Whitney U Test with Benjamini-Hochberg based FDR adjustment. The significant pairs are marked as follows: *: p-value < 0.01. **: 0.01 < p-values < 0.05. (**S)** Circos plots demonstrate the correlation between feature pairs associated with each subpopulation.

Figure 5: Discovery and validation of atypical subpopulations in TNBC and luminal A patients

(a) Sankey plot showing the distribution of patients in the seven subpopulations vs. clinical subtypes. **(b) KM** survival plot for the two TNBC subpopulations in the single cell dataset and the differentially expressed protein biomarkers for the two subpopulations. **(c)** Validation of the two TNBC subpopulations in TCGA breast cancer data corresponding to those in (B), by KM plot and differentially expressed genes. **(d)** Validation of the two TNBC subpopulations in METABRIC breast cancer data corresponding to those in (B), by KM plot and differentially expressed genes. **(e)** KM survival plot for the two luminal A subpopulations in the single cell dataset and the differentially expressed protein biomarkers for the two subpopulations. **(f)** Validation of the two luminal A subpopulations in TCGA breast cancer data corresponding to those in (E), by KM plot and differentially expressed genes. **(g)** Validation of the two luminal A subpopulations in METABRIC breast cancer data corresponding to those in (E), by KM plot and differentially expressed genes. (For each subplot, the blue curve represents the subpopulation in context)

Supplementary Figure 1: Single Cell IMC images based on biomarkers

Single Cell IMC images; illustrating differences between patients based on specific biomarkers. Representative images for each subpopulation highlight the corresponding biomarker for each cell phenotype. E.g., Carbonic Anhydrase IX for hypoxia and KI67 for proliferation

Figure 1

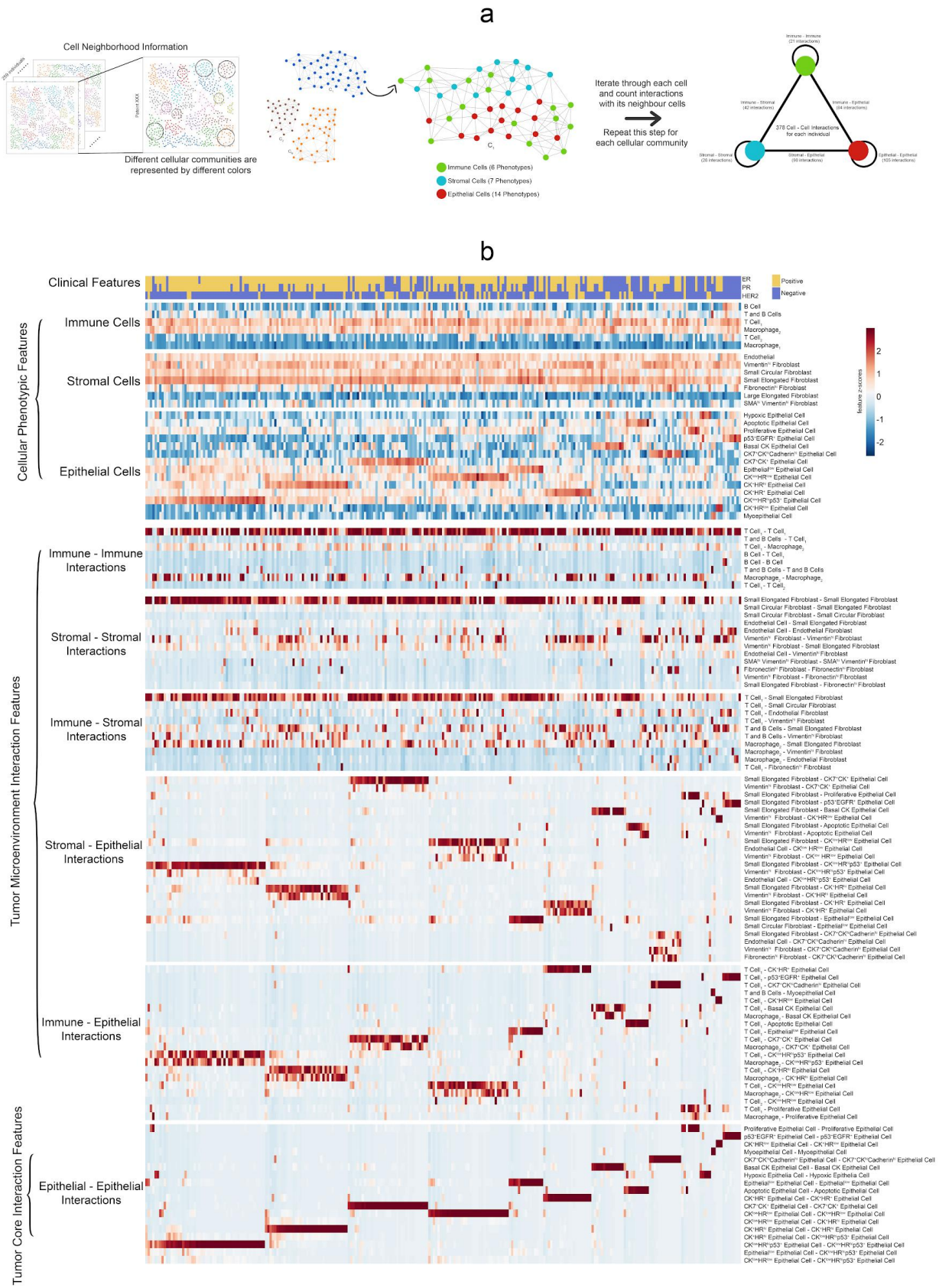
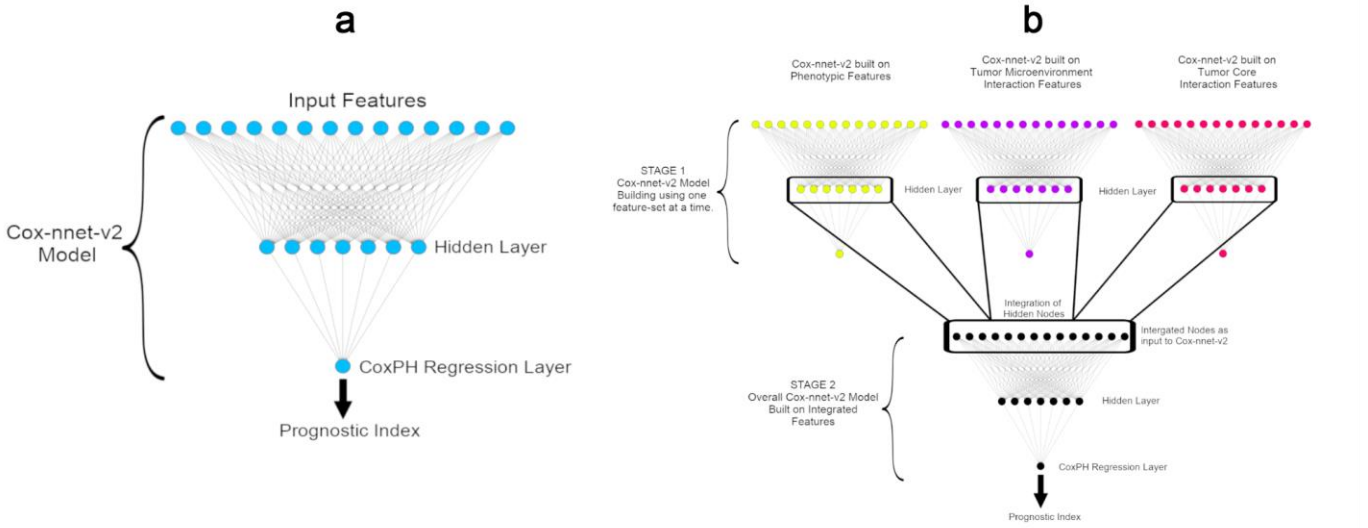


Figure 2



Legend	
Set	**Feature Set**
I	Clinical
II	Phenotypic
III	Tumor Microenvironment Interaction (TMI)
IV	Tumor Core Interaction (TCI)
V	Phenotypic + TMI
VI	Phenotypic + TCI
VII	TMI + TCI
VIII	Phenotypic + TMI + TCI
IX	Phenotypic* + TMI*
X	Phenotypic* + TCI*
XI	TMI* + TCI*
XII	Phenotypic* + TMI* + TCI*

*Denotes hidden layer features obtained from cox-nnet-v2

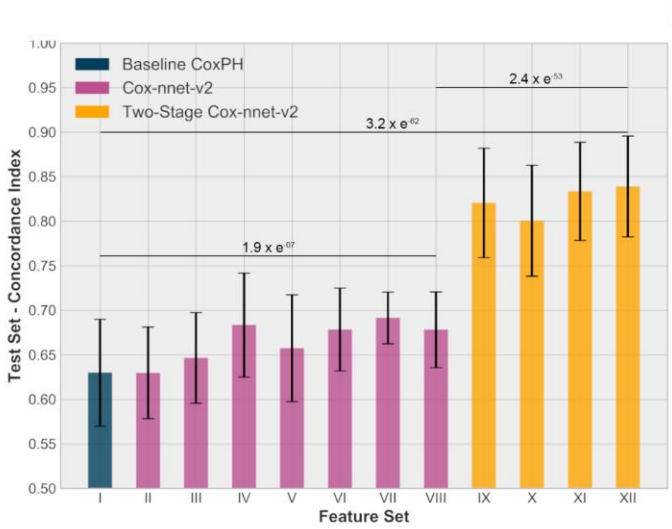


Figure 3

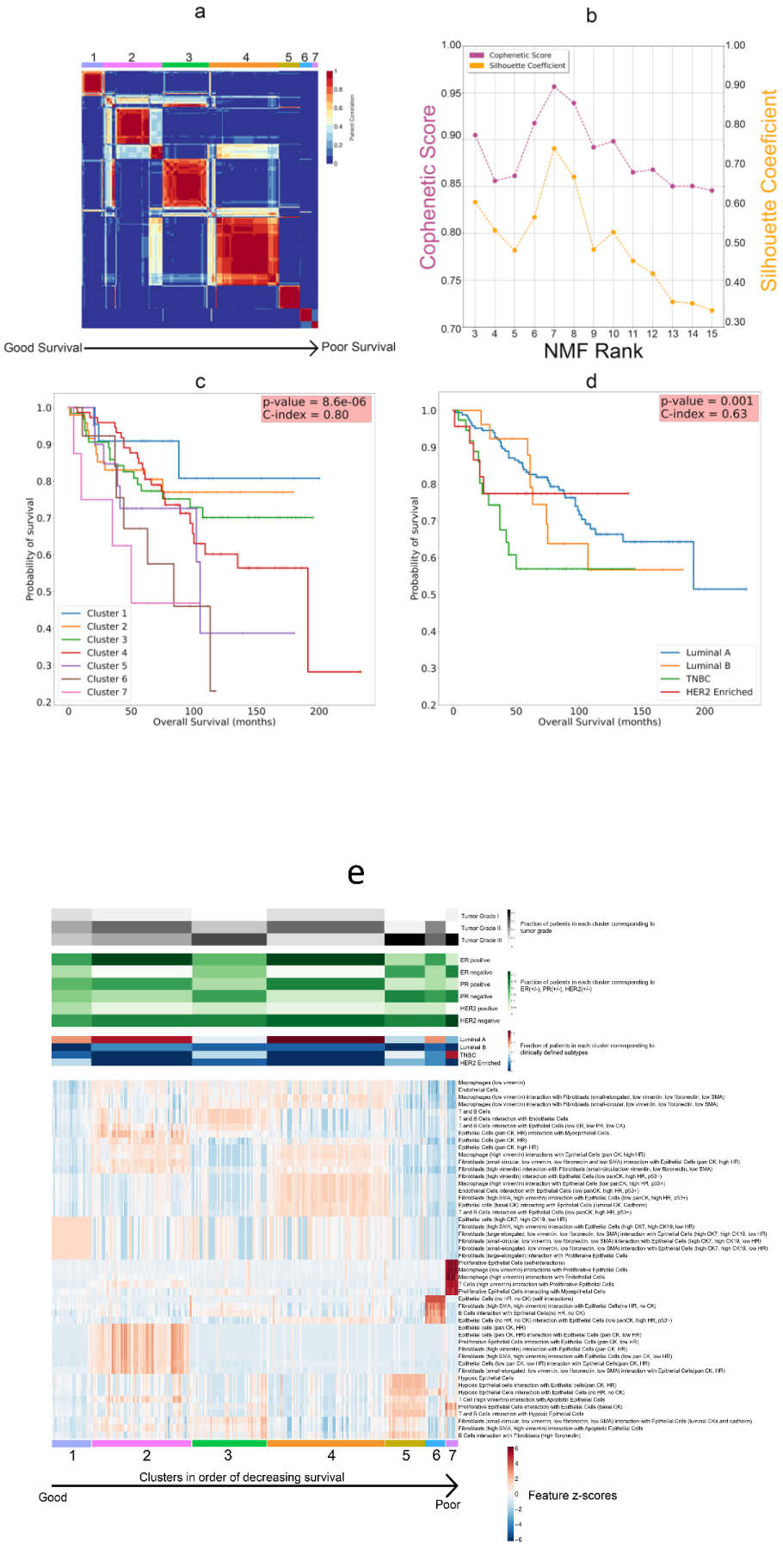


Figure 4

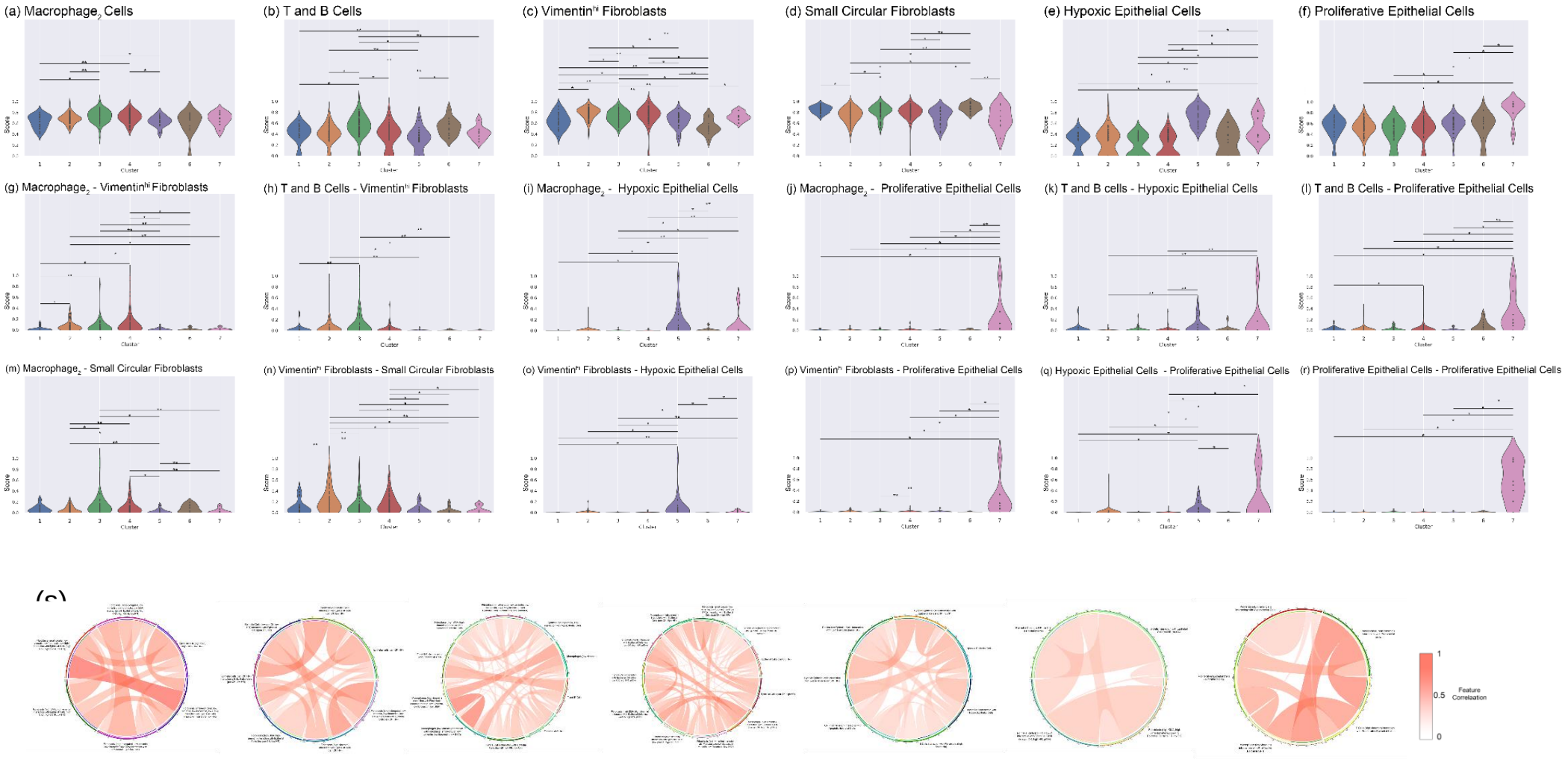
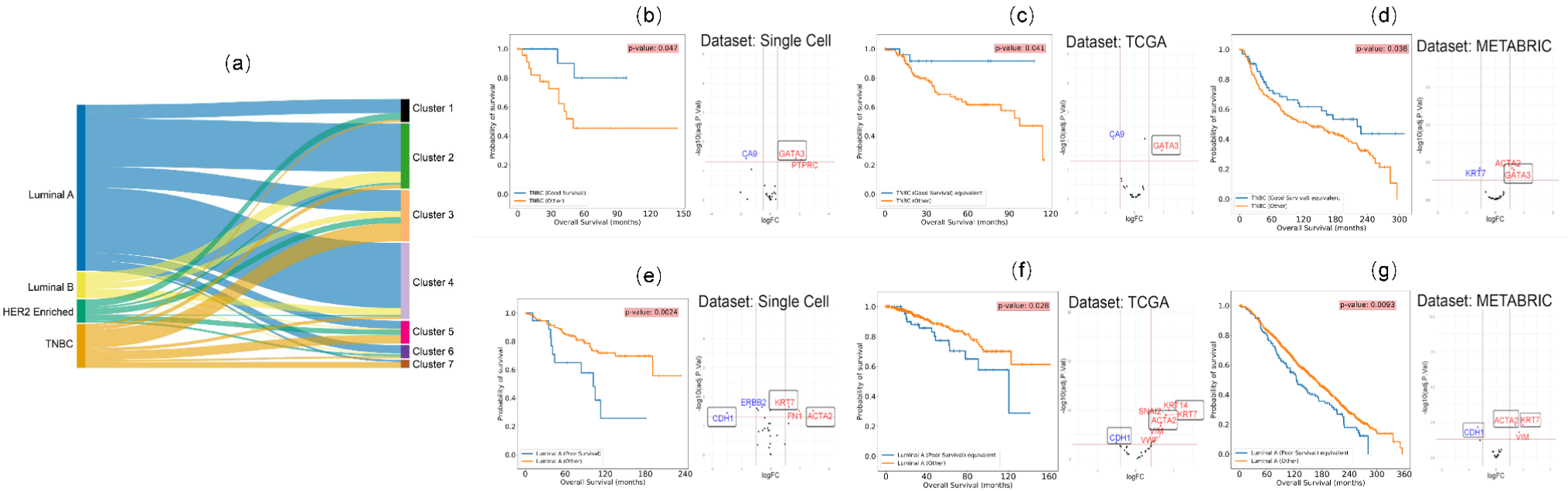
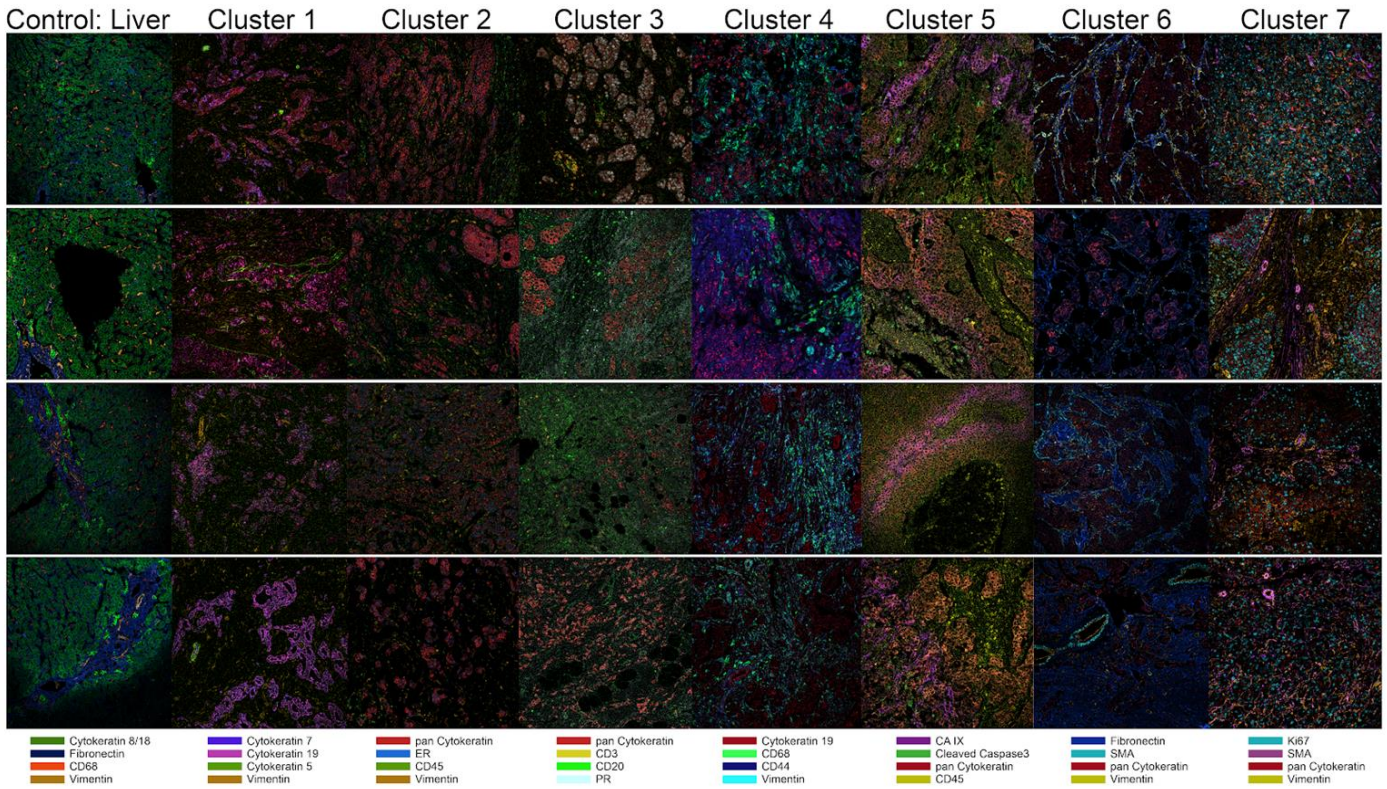

(a) Macrophage$_2$ Cells
(b) T and B Cells
(c) Vimentin^{hi} Fibroblasts
(d) Small Circular Fibroblasts
(e) Hypoxic Epithelial Cells
(f) Proliferative Epithelial Cells
(g) Macrophage$_2$ - Vimentin^{hi} Fibroblasts
(h) T and B Cells - Vimentin^{hi} Fibroblasts
(i) Macrophage$_2$ - Hypoxic Epithelial Cells
(j) Macrophage$_2$ - Proliferative Epithelial Cells
(k) T and B cells - Hypoxic Epithelial Cells
(l) T and B Cells - Proliferative Epithelial Cells
(m) Macrophage$_2$ - Small Circular Fibroblasts
(n) Vimentin^{hi} Fibroblasts - Small Circular Fibroblasts
(o) Vimentin^{hi} Fibroblasts - Hypoxic Epithelial Cells
(p) Vimentin^{hi} Fibroblasts - Proliferative Epithelial Cells
(q) Hypoxic Epithelial Cells - Proliferative Epithelial Cells
(r) Proliferative Epithelial Cells - Proliferative Epithelial Cells
Score
Cluster
(s)
1
0.5
0
Feature Correlation

Figure 5


(a)
Luminal A
Luminal B
HER2 Enriched
TNBC
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
(b)
p-value: 0.047
Overall Survival (months)
Probability of survival
Dataset: Single Cell
-log10(adj.P.Val)
logFC
CA9
GATA3
PTPRC
(c)
p-value: 0.041
Dataset: TCGA
(d)
p-value: 0.038
Dataset: METABRIC
KRT7
ACTA2
(e)
p-value: 0.0024
CDH1
ERBB2
KRT7
FN1
ACTA2
(f)
p-value: 0.028
SNAI2
KRT14
(g)
p-value: 0.0093
VIM

SUPPLEMENTARY MATERIALS

Supplementary Fig. 1



Supplementary Tables

Table S1: Distribution of our cohort's 259 patients corresponding to Clinical Features and clinicopathological subtypes

Clinical Features										Subtypes			
	Tumor Grade			ER Status		PR Status		HER2 Status		Luminal A	Luminal B	TNBC	HER2 Enriched
Total	I	II	III	+	-	+	-	+	-				
259	34	109	116	191	68	143	116	49	210	166	26	44	23

Table S2: Annotations of the 27 Cell Phenotypes defined by (Jackson et al., 2020).

#	Cell Phenotype	Description
1	B Cell	Immune cells showing expression of CD20 (B-Cell biomarker), CD45 (pan-immune biomarker), and Vimentin Expression.
2	T and B Cell	Immune cell cluster (T Cells appearing together with B cells). Hierarchical Clustering is not able to unmix these cells based on the biomarkers. These cells express CD3 (T-Cell biomarker), CD45 (pan-immune biomarker), and CD20 (B-Cell biomarker). This cluster has a low Vimentin expression.
3	$T\ Cell_1$	Immune cells showing expression of CD3(T Cell Biomarker) and CD45 (pan-immune biomarker). This cluster has a low Vimentin expression.
4	$Macrophage_1$	Immune cells expressing CD68 (macrophage biomarker). This cluster has a high level of Vimentin expression.
5	$T\ Cell_2$	Immune cells showing expression of CD3(T Cell Biomarker) and CD45 (pan-immune biomarker). These T Cells have high levels of vimentin expression as compared to the T Cells in Cluster 3.
6	$Macrophage_2$	Immune cells expressing CD68 (macrophage biomarker). These macrophages have low levels of vimentin expression compared to $Macrophage_1$
7	Endothelial cell	These cells form the barrier between blood vessels and tissues. Expressing vWF(von Willebrand factor - endothelial marker), CD31 (endothelial marker). These cells also show high Vimentin Expression.
8	$Vimentin^{hi}$ Fibroblast	Fibroblasts cells with high vimentin expression, low smooth muscle actin (SMA) expression, low fibronectin expression. These cells also express high cMYC (a proto-oncogene).
9	Small circular Fibroblast	These fibroblasts cells are small and circular and show low expression of Vimentin, Fibronectin, and SMA.
10	Small elongated Fibroblast	These are fibroblasts cells small in size and elongated (high eccentricity) and show low expression of Vimentin, Fibronectin, and SMA.
11	$Fibronectin^{hi}$ Fibroblast	Fibroblasts cells with high fibronectin expression, low smooth muscle actin (SMA) expression, low vimentin expression.
12	Large Elongated Fibroblast	These fibroblasts cells are larger and elongated and show low Vimentin, Fibronectin, and SMA expression.
13	SMA^{hi} $Vimentin^{hi}$ Fibroblast	Fibroblasts cells with high SMA expression, high vimentin expression, and low fibronectin expression. They also show an increased expression of CD68.
14	Hypoxic epithelial	These epithelial cells have a high expression of CAIX (Carbonic Anhydrase), which is a hypoxia marker. Rest other epithelial markers such as cytokeratins, e-cadherins, ER, PR have a deficient expression.

15	Apoptotic epithelial	These epithelial cells have a high expression of PARP (Poly (ADP-ribose) polymerase) and Caspase 3, markers for apoptosis. These cells also show an increased expression of p53 and EGFR (epidermal growth factor receptor). Rest other epithelial markers such as cytokeratin, e-cadherins, ER, PR have a deficient expression.
16	Proliferative epithelial	These epithelial cells have high expressions of Ki67 (nuclear protein associated with cellular proliferation), pHH3 (Phosphorylated Histone H3), and Phospho-S6 ribosomal protein. Rest other epithelial markers such as cytokeratins, e-cadherins, ER, PR have a deficient expression.
17	$p53^{+}EGFR^{+}$ epithelial	These epithelial cells have high expressions of p53 and EGFR (epidermal growth factor receptor). Other significantly expressed biomarkers are CAIX (Carbonic Anhydrase), cMYC (a proto-oncogene). Rest other epithelial markers such as cytokeratins, e-cadherins, ER, PR have a deficient expression.
18	Basal CK epithelial	These epithelial cells have high expressions of CK5 and CK14 (basal cytokeratin biomarkers). Rest other epithelial markers such as e-cadherins, ER, PR have an extremely low expression.
19	$CK7^{+}CK^{hi}Cadherin^{hi}$ epithelial	These epithelial cells have high expressions of CK7, other luminal cytokeratins (CK8/18, CK19 biomarkers), and E/P Cadherins. Other epithelial markers, ER, PR have an extremely low expression.
20	$CK7^{+}CK^{+}$ epithelial	These epithelial cells have high expressions of CK19, other luminal cytokeratins (CK7 biomarkers). Rest other epithelial markers such as e-cadherins, ER, PR have an extremely low expression.
21	$Epithelial^{low}$	These epithelial cells have almost negligible expressions of any epithelial markers such as cytokeratins, e-cadherins, ER, PR. Also, they do not express other markers such as p53, EGFR, CAIX, Ki67, PHH3.
22	$CK^{low}HR^{low}$ epithelial	These epithelial cells have low expressions of epithelial markers such as cytokeratins, ER, PR. Also, they do not express other markers such as p53, EGFR, CAIX, Ki67, PHH3.
23	$CK^{+}HR^{hi}$ epithelial	These epithelial cells have very high expressions of epithelial markers such as luminal cytokeratins (CK 8/18, CK 19), pan-cytokeratin, ER, PR. They also express high HER2. However, luminal CK 7 is absent in these cells.
24	$CK^{+}HR^{+}$ epithelial	These epithelial cells have very high expressions of epithelial markers such as luminal cytokeratins (CK 8/18, CK 19), pan-cytokeratin. They do not have an increased expression of ER, PR. They also express high HER2. However, luminal CK 7 is absent in these cells.
25	$CK^{+}HR^{low}$ epithelial	These epithelial cells have very high expressions of epithelial markers such as luminal cytokeratins (CK 8/18, CK 19), pan-cytokeratin. However, they have extremely low ER, PR, HER2 expressions. Luminal CK 7 is absent in these cells.

26	$CK^{low}HR^{hi}p53^{+}$ epithelial	These epithelial cells have very low expressions of epithelial markers such as luminal cytokeratins (CK 8/18, CK 19), pan-cytokeratin. However, they have high ER, PR, HER2 expression. They also express a significant amount of p53, EGFR, PARP, and Caspase 3
27	Myoepithelial	These epithelial cells form semi-continuous protective sheets separating the human breast epithelium and the surrounding stroma and show higher expression of SMA, Luminal Cytokeratins, Basal Cytokeratins, pan Cytokeratins biomarkers.

Table S3: Significant features of the Cellular Phenotypic (CP) feature set.

Rank	Cellular Phenotype	Importance Score	Type
1	$Macrophage_2$	1.000	Immune
2	T and B Cell	0.692	Immune
3	$CK^{+}HR^{+}$	0.631	Epithelial
4	$CK7^{+}CK^{+}$	0.576	Epithelial
5	Endothelial	0.546	Stromal
6	Hypoxic	0.543	Epithelial
7	$CK^{+}HR^{low}$	0.508	Epithelial
8	$CK^{+}HR^{hi}$	0.507	Epithelial

Table S4: Significant features of the Tumor Microenvironment Interaction (TMI) feature set.

Rank	Cell Phenotype 1	Cell Phenotype 2	Importance Score	Interaction Type
1	T and B Cell	$CK^{low}HR^{low}$	1.000	Immune-Epithelial
2	Small Elongated Fibroblast	$CK7^{+}CK^{+}$	0.966	Stromal-Epithelial
3	$Macrophage_1$	$CK^{low}HR^{hi}p53^{+}$	0.965	Immune-Epithelial
4	$Vimentin^{hi}$ Fibroblast	Small Circular Fibroblast	0.964	Stromal-Stromal
5	Large Elongated Fibroblast	Proliferative	0.922	Stromal-Epithelial
6	$SMA^{hi}Vimentin^{hi}$ Fibroblast	$CK7^{+}CK^{+}$	0.892	Stromal-Epithelial
7	B Cell	$Epithelial^{low}$	0.885	Immune-Epithelial

8	T and B Cell	Hypoxic	0.867	Immune-Epithelial
9	Macrophage$_2$	Small Circular Fibroblast	0.853	Immune-Stromal
10	Macrophage$_2$	Small Elongated Fibroblast	0.848	Immune-Stromal
11	SMA ^{hi} Vimentin ^{hi} Fibroblast	CK ^{low} HR^{hi} p53 $^{+}$	0.846	Stromal-Epithelial
12	Small Circular Fibroblast	CK7^{+}CK^{+}	0.841	Stromal-Epithelial
13	Vimentin ^{hi} Fibroblast	CK ^{low} HR ^{hi} p53 $^{+}$	0.823	Stromal-Epithelial
14	T and B Cell	CK ^{low} HR ^{hi} p53 $^{+}$	0.819	Immune-Epithelial
15	Macrophage$_1$	Endothelial	0.814	Immune-Stromal
16	T and B Cell	Endothelial	0.809	Immune-Stromal
17	Endothelial	CK ^{low} HR ^{hi} p53 $^{+}$	0.808	Stromal-Epithelial
18	SMA ^{hi} Vimentin ^{hi}	Epithelial^{low}	0.795	Stromal-Epithelial
19	T Cell$_2$	Proliferative	0.792	Immune-Epithelial
20	Macrophage$_1$	CK^{+} HR ^{hi}	0.783	Immune-Epithelial
21	Small Circular Fibroblast	CK^{+} HR ^{hi}	0.773	Stromal-Epithelial
22	Small Elongated Fibroblast	CK^{+} HR^{+}	0.769	Stromal-Epithelial
23	B Cell	Fibronectin ^{hi} Fibroblast	0.766	Immune-Stromal
24	Vimentin ^{hi} Fibroblast	CK^{+} HR^{+}	0.765	Stromal-Epithelial
25	SMA ^{hi} Vimentin ^{hi}	CK^{low} HR^{low}	0.765	Stromal-Epithelial
26	SMA ^{hi} Vimentin ^{hi}	Apoptotic	0.764	Stromal-Epithelial
27	Small Circular Fibroblast	CK7 $^{+}$ CK ^{hi} cadherin ^{hi}	0.760	Stromal-Epithelial
28	T Cell$_2$	Apoptotic	0.760	Immune-Epithelial
29	Large Elongated Fibroblast	CK7^{+}CK^{+}	0.751	Immune-Stromal
30	Macrophage$_2$	Proliferative	0.751	Immune-Epithelial

Table S5: Significant features of the Tumor Core Interaction (TCI) feature set.

Rank	Cell Phenotype 1	Cell Phenotype 2	Importance Score	Interaction Type
1	Proliferative	Proliferative	1.000	Epithelial-Epithelial
2	$Epithelial^{low}$	$CK^{low}\ HR^{hi}\ p53^{+}$	0.870	
3	$CK^{+}\ HR^{+}$	$CK^{+}\ HR^{low}$	0.811	
4	Hypoxic	$CK^{+}\ HR^{+}$	0.806	
5	Basal CK	$CK7^{+}\ CK^{hi}\ cadherin^{hi}$	0.789	
6	$CK^{low}\ HR^{low}$	$CK^{+}\ HR^{+}$	0.788	
7	Proliferative	Basal CK	0.783	
8	Hypoxic	$Epithelium^{low}$	0.775	
9	Proliferative	$CK^{+}\ HR^{low}$	0.769	
10	Proliferative	Myoepithelial	0.766	
11	$Epithelial^{low}$	$Epithelial^{low}$	0.766	
12	$CK^{+}\ HR^{+}$	Myoepithelial	0.763	

Table S6: Description of patients in NMF-derived subpopulation clusters corresponding to clinical features and clinicopathological subtypes

Distribution of Patient in each Cluster Clinical Features														
Clinical Features											Subtypes			
Cluster ID No. Of Patients		Tumor Grade			ER Status		PR Status		HER2 Status		Luminal A	Luminal B	TNBC	HER2 Enriched
		I	II	III	+	-	+	-	+	-				
1	23	6	9	8	14	9	12	11	7	16	14	0	2	7
2	65	7	35	23	59	6	45	20	13	52	48	11	4	2
3	51	4	15	32	27	24	19	32	12	39	21	6	18	6
4	76	16	41	19	71	5	55	21	8	68	65	7	3	1
5	23	0	3	20	9	14	5	18	6	17	8	1	9	5
6	13	0	6	7	9	4	5	8	3	10	8	1	2	2
7	8	1	0	7	2	6	2	6	0	8	2	0	6	0

REFERENCES

Ali, H. R., Jackson, H. W., Zanotelli, V. R. T., Danenberg, E., Fischer, J. R., Bardwell, H., . . . Bodenmiller, B. (2020). Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nature Cancer, 1*(2), 163-175. doi:10.1038/s43018-020-0026-6

An, Q., Liu, T., Wang, M. Y., Yang, Y. J., Zhang, Z. D., Liu, Z. J., & Yang, B. (2021). KRT7 promotes epithelial-mesenchymal transition in ovarian cancer via the TGFbeta/Smad2/3 signaling pathway. *Oncol Rep, 45*(2), 481-492. doi:10.3892/or.2020.7886

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., . . . Pe'er, D. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell, 174*(5), 1293-1308.e1236. doi:10.1016/j.cell.2018.05.060

Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N., & Patrick, E. (2019). Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Frontiers in Immunology, 10*. doi:10.3389/fimmu.2019.02657

Byrne, D. J., Deb, S., Takano, E. A., & Fox, S. B. (2017). GATA3 expression in triple-negative breast cancers. *Histopathology, 71*(1), 63-71. doi:10.1111/his.13187

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70. doi:10.1038/nature11412

Cao, K., Bai, X., Hong, Y., & Wan, L. (2020). Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics, 36*(Suppl_1), i48-i56. doi:10.1093/bioinformatics/btaa443

Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., . . . Navin, N. E. (2018). Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell, 172*(1-2), 205-217 e212. doi:10.1016/j.cell.2017.12.007

Chen, Y.-C., Sahoo, S., Brien, R., Jung, S., Humphries, B., Lee, W., . . . Yoon, E. (2019). Single-cell RNA-sequencing of migratory breast cancer cells: discovering genes associated with cancer metastasis. *Analyst, 144*(24), 7296-7309. doi:10.1039/C9AN01358J

Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology, 14*(4), e1006076. doi:10.1371/journal.pcbi.1006076

Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., . . . Park, W.-Y. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications, 8*(1), 15081. doi:10.1038/ncomms15081

Communal, L., Roy, N., Cahuzac, M., Rahimi, K., Kobel, M., Provencher, D. M., & Mes-Masson, A. M. (2021). A Keratin 7 and E-Cadherin Signature Is Highly Predictive of Tubo-Ovarian High-Grade Serous Carcinoma Prognosis. *Int J Mol Sci, 22*(10). doi:10.3390/ijms22105325

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-202. doi:10.1111/j.2517-6161.1972.tb00899.x

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., . . . Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346-352. doi:10.1038/nature10983

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, *15*(2), 81-94. doi:10.1038/nrclinonc.2017.166

De Bin, R. (2016). Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computational Statistics*, *31*(2), 513-531. doi:10.1007/s00180-015-0642-2

Ding, S., Chen, X., & Shen, K. (2020). Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Communications*, *40*(8), 329-344. doi:10.1002/cac2.12078

Frigyesi, A., & Höglund, M. (2008). Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes. *Cancer Informatics*, *6*, 275-292.

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811-2812. doi:10.1093/bioinformatics/btu393

Hu, G., Xu, F., Zhong, K., Wang, S., Huang, L., & Chen, W. (2018). Activated Tumor-infiltrating Fibroblasts Predict Worse Prognosis in Breast Cancer Patients. *Journal of Cancer*, *9*(20), 3736-3742. doi:10.7150/jca.28054

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841-860. doi:10.1214/08-AOAS169

Jackson, H. W., Fischer, J. R., Zanotelli, V. R. T., Ali, H. R., Mechera, R., Soysal, S. D., . . . Bodenmiller, B. (2020). The single-cell pathology landscape of breast cancer. *Nature*, *578*(7796), 615-620. doi:10.1038/s41586-019-1876-x

Jögi, A., Ehinger, A., Hartman, L., & Alkner, S. (2019). Expression of HIF-1α is related to a poor prognosis and tamoxifen resistance in contralateral breast cancer. *PLOS ONE*, *14*(12), e0226150. doi:10.1371/journal.pone.0226150

Kanyılmaz, G., Yavuz, B. B., Aktan, M., Karaağaç, M., Uyar, M., & Fındık, S. (2019). Prognostic Importance of Ki-67 in Breast Cancer and Its Relationship with Other Prognostic Factors. *European Journal of Breast Health*, *15*(4), 256-261. doi:10.5152/ejbh.2019.4778

Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., . . . Ellisen, L. W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications, 9*(1), 3588. doi:10.1038/s41467-018-06052-0

Keam, B., Im, S. A., Kim, H. J., Oh, D. Y., Kim, J. H., Lee, S. H., . . . Bang, Y. J. (2007). Prognostic impact of clinicopathologic parameters in stage II/III breast cancer treated with neoadjuvant docetaxel and doxorubicin chemotherapy: paradoxical features of the triple negative breast cancer. *BMC Cancer, 7*, 203. doi:10.1186/1471-2407-7-203

Keam, B., Im, S. A., Lee, K. H., Han, S. W., Oh, D. Y., Kim, J. H., . . . Bang, Y. J. (2011). Ki-67 can be used for further classification of triple negative breast cancer into two subtypes with different response and prognosis. *Breast Cancer Res, 13*(2), R22. doi:10.1186/bcr2834

Korsching, E., Packeisen, J., Liedtke, C., Hungermann, D., Wülfing, P., van Diest, P. J., . . . Buerger, H. (2005). The origin of vimentin expression in invasive breast cancer: epithelial-mesenchymal transition, myoepithelial histogenesis or histogenesis from progenitor cells with bilinear differentiation potential? *The Journal of Pathology, 206*(4), 451-457. doi:10.1002/path.1797

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir el, A. D., Tadmor, M. D., . . . Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell, 162*(1), 184-197. doi:10.1016/j.cell.2015.05.047

Lin, H., Hong, Y., Huang, B., Liu, X., Zheng, J., & Qiu, S. (2019). Vimentin Overexpressions Induced by Cell Hypoxia Promote Vasculogenic Mimicry by Renal Cell Carcinoma Cells. *Biomed Res Int, 2019*, 7259691. doi:10.1155/2019/7259691

Liu, Z., Li, M., Jiang, Z., & Wang, X. (2018). A Comprehensive Immunologic Portrait of Triple-Negative Breast Cancer. *Translational Oncology, 11*(2), 311-329. doi:10.1016/j.tranon.2018.01.011

Lu, S., Yakirevich, E., Wang, L. J., Resnick, M. B., & Wang, Y. (2019). Cytokeratin 7-negative and GATA binding protein 3-negative breast cancers: Clinicopathological features and prognostic significance. *BMC Cancer, 19*(1), 1085. doi:10.1186/s12885-019-6295-8

Lüönd, F., Tiede, S., & Christofori, G. (2021). Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British Journal of Cancer*, 1-12. doi:10.1038/s41416-021-01328-7

Ma, X., Gu, J., Wang, K., Zhang, X., Bai, J., Zhang, J., . . . Qu, K. (2019). Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection. *Scientific Reports, 9*(1), 7073. doi:10.1038/s41598-019-43548-1

Mackillop, W. J. (2003). The importance of prognosis in cancer medicine. *TNM Online*.

McAllister, S. S., & Weinberg, R. A. (2010). Tumor-Host Interactions: A Far-Reaching Relationship. *Journal of Clinical Oncology, 28*(26), 4022-4028. doi:10.1200/JCO.2010.28.4257

Menz, A., Weitbrecht, T., Gorbokon, N., Büscheck, F., Luebke, A. M., Kluth, M., . . . Simon, R. (2021). Diagnostic and prognostic impact of cytokeratin 18 expression in human tumors: a tissue microarray study on 11,952 tumors. *Molecular Medicine, 27*(1), 16. doi:10.1186/s10020-021-00274-7

Mimeault, M., & Batra, S. K. (2013). Hypoxia-inducing factors as master regulators of stemness properties and altered metabolism of cancer- and metastasis-initiating cells. *Journal of Cellular and Molecular Medicine, 17*(1), 30-54. doi:10.1111/jcmm.12004

Murdoch, C., Muthana, M., & Lewis, C. E. (2005). Hypoxia regulates macrophage functions in inflammation. *J Immunol, 175*(10), 6257-6263. doi:10.4049/jimmunol.175.10.6257

Oka, H., Shiozaki, H., Kobayashi, K., Inoue, M., Tahara, H., Kobayashi, T., . . . et al. (1993). Expression of E-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis. *Cancer Res, 53*(7), 1696-1701.

Oue, N., Noguchi, T., Anami, K., Kitano, S., Sakamoto, N., Sentani, K., . . . Yasui, W. (2012). Cytokeratin 7 is a predictive marker for survival in patients with esophageal squamous cell carcinoma. *Annals of Surgical Oncology, 19*(6), 1902-1910. doi:10.1245/s10434-011-2175-4

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., . . . Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752. doi:10.1038/35021093

Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., & Garmire, L. X. (2021). DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med, 13*(1), 112. doi:10.1186/s13073-021-00930-x

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res, 43*(7), e47. doi:10.1093/nar/gkv007

Saccani, A., Schioppa, T., Porta, C., Biswas, S. K., Nebuloni, M., Vago, L., . . . Sica, A. (2006). p50 nuclear factor-kappaB overexpression in tumor-associated macrophages inhibits M1 inflammatory responses and antitumor resistance. *Cancer Res, 66*(23), 11432-11440. doi:10.1158/0008-5472.CAN-06-1867

Saviano, A., Henderson, N. C., & Baumert, T. F. (2020). Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology. *Journal of Hepatology, 73*(5), 1219-1230. doi:10.1016/j.jhep.2020.06.004

Schulze, A. B., Schmidt, L. H., Heitkötter, B., Huss, S., Mohr, M., Marra, A., . . . Evers, G. (2020). Prognostic impact of CD34 and SMA in cancer-associated fibroblasts in stage I–III NSCLC. *Thoracic Cancer, 11*(1), 120-129. doi:10.1111/1759-7714.13248

Shao, W., Cheng, J., Sun, L., Han, Z., Feng, Q., Zhang, D., & Huang, K. (2018, 2018). *Ordinal Multi-modal Feature Selection for Survival Analysis of Early-Stage Renal Cancer*.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, 98*(19), 10869-10874. doi:10.1073/pnas.191367098

Su, L., Pan, P., Yan, P., Long, Y., Zhou, X., Wang, X., . . . Liu, D. (2019). Role of vimentin in modulating immune cell apoptosis and inflammatory responses in sepsis. *Scientific Reports, 9*(1), 5747. doi:10.1038/s41598-019-42287-7

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209-249. doi:10.3322/caac.21660

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine, 16*(4), 385-395. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Tibshirani, R. & Hastie, T. (2007). Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1), 2–8. doi.org/10.1093/biostatistics/kxl005

Turashvili, G., & Brogi, E. (2017). Tumor Heterogeneity in Breast Cancer. *Frontiers in Medicine, 4*. doi:10.3389/fmed.2017.00227

Wang, D., Jing, Z., He, K., & Garmire, L. X. (2021). Cox-nnet v2.0: improved neural-network-based survival prediction extended to large-scale EMR data. *Bioinformatics*(btab046). doi:10.1093/bioinformatics/btab046

Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., . . . Swarbrick, A. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet, 53*(9), 1334-1347. doi:10.1038/s41588-021-00911-1

Yan, X., Xie, Y., Yang, F., Hua, Y., Zeng, T., Sun, C., . . . Yin, Y. (2021). Comprehensive description of the current breast cancer microenvironment advancements via single-cell analysis. *J Exp Clin Cancer Res, 40*(1), 142. doi:10.1186/s13046-021-01949-z

Yerushalmi, R., Woods, R., Ravdin, P. M., Hayes, M. M., & Gelmon, K. A. (2010). Ki67 in breast cancer: prognostic and predictive potential. *The Lancet Oncology, 11*(2), 174-183. doi:10.1016/S1470-2045(09)70262-1

Zhan, Z., Jing, Z., He, B., Hosseini, N., Westerhoff, M., Choi, E.-Y., & Garmire, L. X. (2021). Two-stage Cox-nnet: biologically interpretable neural-network model for prognosis prediction and its application in liver cancer survival using histopathology and transcriptomic data. *NAR Genomics and Bioinformatics, 3*(1), lqab015. doi:10.1093/nargab/lqab015

Zhang, H., Qin, G., Yu, H., Han, X., & Zhu, S. (2021). Comprehensive genomic and immunophenotypic analysis of CD4 T cell infiltrating human triple-negative breast cancer. *Cancer Immunol Immunother, 70*(6), 1649-1665. doi:10.1007/s00262-020-02807-1

Zhang, L., Huang, G., Li, X., Zhang, Y., Jiang, Y., Shen, J., . . . Qian, C. (2013). Hypoxia induces epithelial-mesenchymal transition via activation of SNAI1 by hypoxia-inducible factor -1alpha in hepatocellular carcinoma. *BMC Cancer, 13*, 108. doi:10.1186/1471-2407-13-108