Deep-learning and transfer learning identify new breast cancer survival subtypes from single-cell imaging data

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

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

* These authors contributed equally to the work

[#]Corresponding author

Email addresses:

SY: shyadav@med.umich.edu

SZ: shuzh@med.umich.edu

BH: hbing@med.umich.edu

YD: yhdu@med.umich.edu

LXG: lgarmire@med.umich.edu

Running title: new breast cancer survival subtypes from single-cell imaging data

Key words: breast cancer; survival; subtype; deep-learning; transfer-learning

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 computationally 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 (marked by GATA3 over-expression) and Luminal A patients with poor prognosis (marked by KRT6 and ACTA2 over-expression and CDH1 under-expression). These atypical subpopulations are validated in TCGA-BRCA and METABRIC datasets. This work provides important guidelines on bridging single-cell level information towards population-level survival prediction.

STATEMENT OF TRANSLATIONAL RELEVANCE

Our findings from a breast cancer population cohort demonstrate the clinical utility of using the single-cell level imaging mass cytometry (IMC) data as a new type of patient prognosis prediction marker. Not only did the prognosis prediction achieve high accuracy with a Concordance index score greater than 0.8, it also enabled the discovery of seven survival subtypes that are more distinguishable than the molecular subtypes. These new subtypes present distinct profiles of epithelial, immune, fibroblast cells, and their interactions. Most importantly, this study identified and validated atypical subpopulations of TNBC patients with moderate prognosis (GATA3 over-expression) and Luminal A patients with poor prognosis (KRT6 and ACTA2 over-expression and CDH1 under-expression), using multiple large breast cancer cohorts.

INTRODUCTION

Breast cancer became the most commonly diagnosed cancer in 2020, with an estimated 2.3 million new cases globally [1]. 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 [2]. 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' is composed of several types of immune cells, fibroblasts, adipocytes, and mesenchymal cells [3,4]. 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 a comprehensive understanding of breast cancer heterogeneity by elucidating the contribution of tumor, tumor microenvironment, and their intricate interactions [5,6].

Investigations at single-cell resolution [7,8] 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[9–11]. Potential biomarkers have also been discovered for personalized cancer immunotherapy through Single-cell RNA sequencing (scRNA-seq) [12]. However, one major limitation of scRNA-seq is the loss of spatial information crucial in understanding tumor heterogeneity *in situ* [13]. 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 [14]. 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 [15,16]. 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 [17]. 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 [16]. We applied a neural network-based method Cox-nnet previously developed in our group [18] 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 the transfer learning approach and validated the presence of such unconventional subgroups and their biomarkers in the TCGA and METABRIC breast cancer datasets.

MATERIALS AND METHODS

Dataset and extracted features

The dataset analyzed in the study is obtained from a previously published study containing 259 patients [16]). 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, and direct neighbor cells are defined within 4 pixels (4μm) distance to the target cell. 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 [19,20].

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 stromal 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 [21] 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.

Cell-cell interactions are restricted to those within the same cellular community. 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 of the interactions by the clustering coefficient C of the cellular community as the weight. 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).

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

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. Among the feature pairs, 1 immune-stromal feature (macrophage interacting with Large Elongated Fibroblasts), 2 immune-epithelial features (macrophage interacting with Hypoxic Epithelial Cells and macrophage interacting with $CK7^{+}CK^{hi}Cadherin^{hi}$ Epithelial Cells), 1 stromal-stromal feature (Large Elongated Fibroblasts self interaction) and 1 stromal-epithelial (Large Elongated Fibroblasts interacting with Myoepithelial cells) are dropped since these pairwise phenotype-phenotype interactions calculations returned 0 (lack of neighbors within the immediate vicinity), hence the final number of features would be 373 features. 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.

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

where $\hat{O}$ is the importance vector of the original features (shape *1*).

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 40 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 [22,23]. 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 [24]. 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 [25,26]. 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 [27] 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 99.5% 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 [28].

Cross-check the TNBC survival subtypes with the molecular subtypes of TNBC

We cross-compared the seven subtypes in the previous study [29] with the two survival subtypes we identified. We used the six subtype markers (MKI67, MYC, EGFR, CDH1, SNAI2 and TWIST1) in that study that overlap with the 30 protein-based biomarkers in the single cell imaging dataset here. The six markers are MKI67 and MYC as the proliferation markers, EGFR as the growth factor/Myoepithelial markers, and SNAI2 and TWIST1 as the epithelial-mesenchymal transition–associated (EMT-associated) markers. The raw data from the previous study are not available, we relied on the pixels in this heatmap as the normalized gene expression values for the comparative study. We first transformed the heatmap of Figure S6 [29] of the six markers from RGB (red, green and blue) to HSV(Hue, Saturation and Value) scale, and filtered out all the red and green related colors by setting the HSV thresholds. Then we transformed the HSV scale to the gray scale, and mapped the intensity to log-transformed gene values according to the figure legend. Specifically, we map green related colors from -2 to 0 and red related colors from 0 to 2. We plotted the level of the six markers from the imaging data alongside the heatmap from the previous study. We performed hierarchical clustering with the average expression level of markers from the two survival subtypes defined by us and the TNBC molecular subtypes obtained from the previous study.

RESULTS

Feature engineering from the breast cancer single-cell images

The breast cancer cohort contains 259 patients with survival data, as reported earlier [16]. 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 between epithelial cells of various cellular phenotypes (**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 [30,31], compared to the conventional Cox-PH method [32]. 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 and TNM (tumor-node-metastasis) stagings (Set I) as the baseline model. We constructed a series of two-stage Cox-nnet models (**Fig. 2A**) on the combinations (Set II, III, IV, and V) of CP features, TMI features and TCI features (Set II, III, IV, and V). Two-stage Cox-nnet models are complex models that use simpler Cox-nnet models as individual building blocks [30] to integrate different feature sets or data modalities. 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 and 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. 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.

We also compared the new survival clusters with the 18 single-cell pathology (SCP) subgroups based on the 27 cellular phenotype features, as identified from the original study [16]. As shown in **Figure S1**, cluster 1 has a strong association (90%) with the SCP_11 (CK7 and pan-CK) subgroup. Cluster 5 is enriched with SCP_17 (Hypoxic) subgroup. Cluster 6 is equally distributed in SCP_7 (Epithelial.low) and SCP_10 Epithelial.low mixed) subgroups. Cluster 7 shows a connection with SCP_14 (Proliferative) subgroup. Cluster 2, Cluster3 and Cluster 4 do not show a clear association with some particular SCP subgroups.

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. We dichotomized the 7 survival subgroups into "good survival (subgroups 1-4)" and "bad survival (subgroups 5-7)" groups, and noticed that for luminal A and TNBC cancers these subgroups show significantly different survival outcomes (log-ranked p-value < 0.05) from their counterparts (**Fig. 5B, 5E**). Despite being a molecular subtype regarded as having the worst prognosis, TNBC actually has a large proportion (61.36%) of atypical good survival patients from clusters 1-4, where cluster 3 is the major cluster among the four (66.66%). On the other hand, while the luminal A subtype generally has the best prognosis among the four molecular subtypes, around 11% of the luminal A breast cancers still belong to the relatively "poor survival" group. We next 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, Supplementary Table 7).** On the other hand, the atypical "poor survival" subpopulation in the luminal A subtype has over-expression of KRT7 and ACTA2, and an under-expression of CDH1 (**Fig. 5E, Supplementary Table 8**).

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, Supplementary Tables 9-10**). Additionally, CA9 shows a common under-expression in the TCGA-BRCA dataset (**Fig. 5B-C)**. Similarly, the matching "poor survival" luminal 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, Supplementary Tables 10-12**). Loss of CDH1 is a characteristic of invasive lobular breast cancer [33]. Confirming this, we found that lobular type increases from 20.72% to 50.7%, comparing the good vs. bad survival subgroups in TCGA luminal A breast cancers. It also increases from 8.34% to 12.7%, for the good vs bad survival subgroups in the luminal A cancers in METABRIC dataset.

TNBC was previously reported to have six molecular subtypes[29], including 2 basal-like (BL1 and BL2), an immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem–like (MSL), and a luminal androgen receptor (LAR) subtype, as well as an extra unstable (UNS) subtype. We further cross-checked the survival based TNBC subtypes here with those reported earlier (**Figure S2**). We selected six markers (MKI67, MYC, EGFR, CDH1, SNAI2, TWIST1) that are in common with the Figure S6 of the original study and our single-cell imaging dataset, and relied on the pixels in this heatmap as the normalized gene expression values for the comparative study. We scaled the marker gene intensities from the single cell imaging data by the RGB values of the pixels. The good TNBC survival group of the image data shows a lower expression in the proliferation markers and myoepithelial markers and a mixed expression in EMT markers (**Figure S2b**). And the bad TNBC survival group of the image data shows a combination of low and high expressions in Proliferation markers and EMT markers (**Figure S2c)**. From the clustering result of averaged gene expression in two types of subtype systems, the good TNBC survival subgroup is the closest to MSL and LAR subgroups, where as the bad TNBC survival group is closest to BL2 and M subgroups (**Figure S2d**)

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 improved 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 amend the molecular subtypes originally based on tumor cell signatures, as each of the molecular subtypes is highly heterogeneous in terms of patients prognosis. It also extends beyond previous work that classified breast cancers by immuno-subtypes [34], by considering additionally the interactions among tumor, immune cells etc.

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 [35]. 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 [36]. 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 [37,38]. Despite most of the patients being 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) [39]. It also has the 2nd highest level of proliferative epithelial cells, both of which may contribute to poor survival. Subtype 7 has the highest level of proliferative epithelial cells marked by KI-67 expression, associated with poor survival [40,41]. It also has some of the strongest interactions with cells in EMT, suggesting the highest degree of tumor infiltration by immune cells [42]. 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 [43]. 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 [44], 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 [45], consistent with prior observation of a close relationship between proliferation-hypoxia in other solid cancers such as squamous cell carcinoma [46].

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 small population of "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 [47]. 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 [48,49]. Similarly, ACTA2 was reported to be a marker for poor prognosis in lung cancer [50]. 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 [51]. Thus 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.

We further investigated the marker genes that have high correlations (Pearson's correlation coefficient >0.5 or <-0.5) with the cell-cell interactions, using the two extreme subtypes 1 and 7 (**Figure S3**). Many above described relationships with regards to prognosis are confirmed again. For subgroup 1 (**Figure S3a**), there exist positive correlations between CDH1 and epithelial-epithelial cell interactions. Since the loss of CDH1 is a characteristic of invasive lobular breast cancer [33], a strong positive correlation between CDH1 and the epithelial-epithelial interactions does indicate good

survival. On the other hand, positive correlations of MKI67 with epithelial-epithelial interaction and T-cell-epithelial cell interaction in group 7 (**Figure S3b**) indicates poor survival, as expected [52]. Interestingly, GATA3 is negatively correlated with T-cell-epithelial cell interaction in group 7, a cell-cell interaction signature for poor survival. However, it is positively correlated with epithelial-epithelial interaction in group 1 with good survival.

While it is true that the cost and time of the multiplex imaging platform currently do not allow live-time assistance to clinicians to precision treatment yet, this technology is highly reproducible and futuristic. It has already been used for the diagnosis of other diseases like leukemia [53] This technology has provided information of disseminated tumor cell phenotypes that frequently deviate from the clinical disease subtype [54]. It also significantly advanced our understanding of the intra-tumor heterogeneity and how the interactions among different cell types in tumor and tumor microenvironment affect patient outcome [15,16,55]. It is expected that in the near future, with the reduction of cost and time and assistance of AI tools and user-friendly computational tools, such technologies will eventually be adopted by pathology and be more informative to clinicians.

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.

CONFLICT OF INTEREST STATEMENT

The author claims no conflict of interest.

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 initial data analysis, generated the main figures, and wrote the manuscript. SZ characterized the atypical molecular subtypes, revised the manuscript and figures, tested the code used in this study and made it available on the Github repository. BH helped with subgroup selection and interpretation. YD helped test the atypical molecular subtypes with regard to the cell proportions.

ACKNOWLEDGEMENT

We thank Dr. Hartland Jackson for providing data for this project. We thank Yuheng Du in the Garmire group for valuable discussions. 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 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. **(b)** 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 FIGURES

Figure S1: Sankey plot showing the distribution of patients in the seven subpopulations vs. the original 18 single-cell pathology (SCP) subgroups.

Figure S2: Cross-check the good and bad TNBC survival subtypes with previous defined TNBC molecular subtypes. (a) Heatmap of the expression level of the six common marker genes in the seven previously defined TNBC molecular subtypes. **(b)** Heatmap of the expression level of the six common marker genes in the good survival TNBC subtype. **(c)** Heatmap of the expression level of the six common marker genes in the bad survival TNBC subtype. (**d**) Hierarchical clustering results on the average expression levels of the TNBC survival subtypes and the seven TNBC molecular subtypes.

Figure S3: Bipartite graph demonstrating the significant correlations between cell-cell interaction pairs (ellipses) and gene markers (circles) in subgroups 1 (a) and 7 (b). Ellipse: cell-cell interaction pair; circle: gene marker. Red line: positive correlations; blue lines: negative correlations. The sizes of the ellipses and circles are proportional to the average expression level of cell-cell interactions and gene expression respectively.

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

Table S7: Results of the differential analysis to the TNBC subtypes of the single-cell imaging cohort

Gene Symbol	logFC	AveExpr	t	P.Value	adj.P.Val
CA9	-1.66532	1.181947	-3.43789	0.001317	0.037967
GATA3	1.735083	1.894308	3.208068	0.002531	0.037967
PTPRC	2.099658	2.148328	3.016766	0.004288	0.042875
CDH1	-1.29972	2.695109	-2.6223	0.012046	0.090348
MKI67	-0.39527	0.455639	-1.98155	0.053973	0.323837
PGR	0.31247	0.670969	1.870176	0.068305	0.341527
CD68	0.263879	0.694707	1.555042	0.127297	0.545558
H3F3B	-2.03734	8.670026	-1.48123	0.145868	0.547003

MYC	-0.25257	0.561824	-1.32181	0.193251	0.604844
MTOR	0.32937	0.957116	1.276558	0.208638	0.604844
KRT14	0.127985	0.24281	1.239867	0.221776	0.604844
EGFR	-0.29323	0.633435	-1.14713	0.257698	0.644245
ESR1	-0.04331	0.154521	-0.99907	0.323374	0.728471
ERBB2	-0.20024	1.177114	-0.96502	0.339953	0.728471
SNAI2	-0.07591	0.292052	-0.87052	0.388872	0.766365
CASP3	0.122374	0.87393	0.799978	0.42814	0.766365
KRT7	-0.05051	0.187082	-0.77142	0.444694	0.766365
CD3E	-0.02687	0.119141	-0.74587	0.459819	0.766365
ACTA2	-0.07116	0.331779	-0.60659	0.547323	0.864193
TP53	0.007247	0.093337	0.395215	0.694646	0.94636
VIM	0.35772	4.605131	0.311459	0.756963	0.94636
KRT5	0.016114	0.184095	0.288969	0.774	0.94636
VWF	-0.00535	0.109268	-0.28682	0.775636	0.94636
CD44	-1.57545	33.46401	-0.25769	0.79788	0.94636
MS4A1	-0.01068	0.142771	-0.23376	0.816284	0.94636
KRT19	-0.01825	0.381231	-0.17763	0.85985	0.94636
TWIST1	0.006286	0.238186	0.1507	0.88092	0.94636
KRT8	0.048855	0.645513	0.147705	0.883269	0.94636
FN1	0.009486	0.345973	0.072506	0.942537	0.975038
RPS6	-0.00023	0.9968	-0.00104	0.999173	0.999173

Table S8: Results of the differential analysis to the luminal A subtypes of single-cell imaging cohort

Gene Symbol	logFC	AveExpr	t	P.Value	adj.P.Val
KRT7	0.621095	0.252991	3.392664	0.000867	0.022652
CASP3	-0.30149	0.716597	-3.18383	0.001738	0.022652
ERBB2	-0.71272	1.428197	-3.10159	0.002265	0.022652
MTOR	-0.47299	1.085437	-2.92243	0.003961	0.025079
H3F3B	-3.88308	8.254407	-2.90479	0.00418	0.025079
TP53	-0.02037	0.068909	-2.74515	0.00672	0.02899
GATA3	-0.43225	0.831243	-2.7429	0.006764	0.02899
ACTA2	1.450311	2.908928	2.64274	0.009017	0.030062

FN1	0.966874	1.939285	2.642688	0.009019	0.030062
CDH1	-1.49285	3.209465	-2.53702	0.012109	0.036327
ESR1	-0.20296	0.333834	-2.36442	0.019223	0.052427
TWIST1	-0.06338	0.228428	-2.14722	0.033239	0.082595
VWF	-0.02091	0.091888	-2.09964	0.037285	0.082595
KRT19	0.611251	1.045851	2.085735	0.038544	0.082595
SNAI2	-0.10615	0.223992	-1.89526	0.05981	0.11962
CD68	-0.20877	0.480148	-1.85104	0.065955	0.123665
KRT14	-0.05732	0.15019	-1.72856	0.085763	0.143635
RPS6	-0.27315	0.831391	-1.70041	0.090942	0.143635
PTPRC	-0.08485	0.20329	-1.70026	0.090969	0.143635
PGR	-0.4174	0.99885	-1.63683	0.103576	0.155364
CA9	0.215853	0.747	1.449928	0.14898	0.212829
CD3E	-0.01952	0.075158	-1.39362	0.165311	0.225424
KRT5	-0.01767	0.131671	-1.00806	0.314905	0.410746
KRT8	0.487363	1.86885	0.962083	0.337418	0.421773
VIM	-0.45621	2.978811	-0.89505	0.372069	0.446483
CD44	-5.46826	22.58901	-0.74825	0.455379	0.525438
MS4A1	-0.05775	0.11631	-0.67831	0.498525	0.553917
MYC	-0.01087	0.407169	-0.13294	0.894403	0.926926
EGFR	-0.00688	0.370803	-0.13088	0.896028	0.926926
MKI67	0.002209	0.138045	0.056161	0.955281	0.955281

Table S9: Results of the differential analysis to the TNBC subtypes of the TCGA data

Gene Name	logFC	AveExpr	t	P.Value	adj.P.Val
KRT5	0.359085	0.353412	3.602914	0.000415	0.008454
CA9	-0.733360	-0.780730	-3.496940	0.000604	0.008454
GATA3	0.944708	-0.016850	3.098935	0.002282	0.021295
CD3E	-0.451420	0.369958	-2.201850	0.029055	0.203382
MS4A1	-0.445460	0.301218	-2.030820	0.043868	0.245663
KRT8	-0.261330	0.585154	-1.725830	0.086238	0.375687
EGFR	0.283720	0.989292	1.684717	0.093922	0.375687
TP53	-0.374510	0.199245	-1.616030	0.107987	0.377956

PTPRC	-0.335660	0.323452	-1.474140	0.142338	0.423541
MYC	0.293505	0.615234	1.441721	0.151265	0.423541
ACTA2	0.316745	-0.131780	1.322205	0.187919	0.478338
H3F3B	0.183001	0.058156	1.009132	0.314380	0.717750
CD44	0.181941	0.170509	0.920988	0.358393	0.717750
CDH1	0.144242	-0.144540	0.845772	0.398897	0.717750
CD68	0.163085	0.233515	0.754943	0.451353	0.717750
FN1	0.151448	-0.245670	0.744126	0.457852	0.717750
VIM	-0.145540	0.583144	-0.711480	0.477788	0.717750
MKI67	0.120782	0.761689	0.687042	0.493015	0.717750
PGR	0.176797	-2.415230	0.672580	0.502150	0.717750
ERBB2	-0.085600	-0.730870	-0.656080	0.512678	0.717750
CASP3	-0.116840	-1.556450	-0.584980	0.559356	0.745808
VWF	0.114310	-0.418070	0.485462	0.627989	0.799259
RPS6	-0.090500	0.383519	-0.351920	0.725345	0.864719
ESR1	-0.041970	-1.605010	-0.302680	0.762516	0.864719
KRT7	-0.057680	0.814242	-0.276940	0.782170	0.864719
KRT19	0.064105	-0.544870	0.249926	0.802953	0.864719
KRT14	0.038968	0.403564	0.191417	0.848433	0.867726
SNAI2	-0.042370	0.135287	-0.166810	0.867726	0.867726

Table S10: Results of the differential analysis to the luminal A subtypes of the TCGA data

Gene Name	logFC	AveExpr	t	P.Value	adj.P.Val
KRT14	0.866476	-0.059990	7.098703	3.48E-12	9.74E-11
SNAI2	0.783525	-0.030140	6.636017	7.07E-11	9.90E-10
KRT7	0.787224	-0.103080	6.549817	1.22E-10	1.14E-09
ACTA2	0.678161	0.027382	5.873735	6.97E-09	4.88E-08
VWF	0.646944	0.152885	5.629324	2.75E-08	1.49E-07
VIM	0.653774	-0.093400	5.602645	3.18E-08	1.49E-07
MYC	0.467220	-0.058380	4.031164	0.000063	0.000250
CDH1	-0.506140	-0.049480	-3.872330	0.000119	0.000418

EGFR	0.403637	-0.249010	3.765384	0.000182	0.000567
PTPRC	0.405830	-0.031410	3.471977	0.000553	0.001548
MKI67	-0.411220	-0.271530	-3.399980	0.000717	0.001826
ESR1	-0.225690	0.446677	-3.262840	0.001164	0.002716
KRT5	0.395878	-0.181310	3.227836	0.001314	0.002829
KRT8	-0.380290	0.114790	-3.202750	0.001431	0.002863
GATA3	-0.240510	0.378847	-3.093340	0.002069	0.003862
CD3E	0.350318	-0.074010	2.877213	0.004151	0.007264
MS4A1	0.326671	-0.136920	2.846465	0.004568	0.007524
KRT19	-0.288910	0.098399	-2.590640	0.009806	0.015254
PGR	0.261333	0.377742	2.499440	0.012698	0.018713
RPS6	0.252696	-0.064950	2.259145	0.024223	0.033912
ERBB2	0.149491	-0.186800	2.104345	0.035753	0.047272
CASP3	-0.229720	-0.187110	-2.088730	0.037142	0.047272
CA9	-0.184530	-0.477180	-1.784310	0.074865	0.091140
H3F3B	-0.209870	-0.018020	-1.728320	0.084432	0.098504
FN1	0.168311	-0.020360	1.360655	0.174120	0.195015
CD44	0.114186	0.031027	0.937303	0.348970	0.375814
TP53	0.078552	0.078036	0.813627	0.416173	0.431587
CD68	0.043611	-0.054470	0.354237	0.723282	0.723282

Table S11: Results of the differential analysis to the TNBC subtypes of the Metabric data

Gene Name	logFC	AveExpr	t	P.Value	adj.P.Val
ACTA2	0.573868	0.077270	3.670104	2.87E-04	7.61E-03
KRT7	-0.549600	0.484024	-3.515140	5.07E-04	7.61E-03
GATA3	0.638064	-0.102740	3.319498	1.01E-03	1.01E-02
H3F3B	0.278453	-0.369990	1.992902	4.72E-02	2.94E-01
EGFR	-0.338870	1.203013	-1.976670	4.90E-02	2.94E-01
MS4A1	0.290526	0.117949	1.779548	7.62E-02	3.81E-01
CD68	0.266183	0.410858	1.592282	0.112372	0.481593
TP53	0.158366	-1.685070	1.428578	0.154163	0.578111
VWF	0.190017	-0.222290	1.277852	0.202287	0.612230

FN1	0.185938	0.086585	1.272786	0.204077	0.612230
KRT19	-0.227440	-0.582490	-1.178710	0.239444	0.653029
KRT8	-0.204340	-0.898690	-1.099790	0.272301	0.680753
CDH1	-0.125990	-0.211900	-0.945570	0.345129	0.751691
RPS6	-0.135820	0.079517	-0.907200	0.365026	0.751691
ERBB2	0.109310	-0.729280	0.886892	0.375845	0.751691
PTPRC	0.146653	0.539348	0.839543	0.401832	0.753434
CA9	-0.156040	0.910100	-0.764220	0.445338	0.776903
KRT5	-0.125250	0.950768	-0.697810	0.485832	0.776903
MTOR	-0.098440	0.138333	-0.641760	0.521519	0.776903
MYC	-0.088930	0.513788	-0.596690	0.551162	0.776903
CD44	0.074794	0.084156	0.569977	0.569119	0.776903
MKI67	0.091584	0.836039	0.539762	0.589761	0.776903
ESR1	0.026827	-1.662260	0.531267	0.595626	0.776903
VIM	0.040330	0.477915	0.314419	0.753421	0.909424
SNAI2	0.047458	0.241982	0.271032	0.786552	0.909424
CASP3	-0.038570	0.454359	-0.268930	0.788168	0.909424
CD3E	-0.031100	0.466719	-0.180230	0.857091	0.910779
KRT14	-0.025880	0.362967	-0.151640	0.879570	0.910779
PGR	-0.00352	-0.91613	-0.12531	0.900364	0.910779
TWIST1	-0.01595	0.154403	-0.11215	0.910779	0.910779

Table S12: Results of the differential analysis to the luminal A subtypes of the Metabric data

Gene Name	logFC	AveExpr	t	P.Value	adj.P.Val
KRT7	0.977340	-0.097440	3.803803	1.49E-04	4.46E-03
ACTA2	0.703218	-0.001900	3.603348	3.25E-04	4.88E-03
CDH1	-0.674130	-0.543030	-3.399420	6.95E-04	6.95E-03
VIM	0.845528	-0.193060	3.084499	2.08E-03	1.56E-02
MS4A1	-0.592870	-0.177460	-2.577770	1.00E-02	6.03E-02
VWF	0.191635	0.053709	1.934711	5.32E-02	2.66E-01
GATA3	0.096992	0.455363	1.832000	0.067168	0.287864
KRT14	0.154100	-0.197550	1.588594	0.112383	0.374411
SNAI2	0.148167	-0.055810	1.562958	0.118293	0.374411

TP53	-0.129190	0.007260	-1.488630	0.136816	0.374411
CA9	-0.096260	-0.256750	-1.486850	0.137284	0.374411
CD3E	-0.119040	-0.157230	-1.363400	0.172979	0.401474
KRT19	0.118648	0.097706	1.326799	0.184796	0.401474
KRT5	0.106327	-0.181740	1.283057	0.199689	0.401474
KRT8	0.104671	0.131979	1.280070	0.200737	0.401474
FN1	0.090496	-0.083140	0.950817	0.341865	0.622803
CD68	0.083207	-0.109750	0.891749	0.372684	0.622803
ERBB2	0.051047	-0.191510	0.889889	0.373682	0.622803
RPS6	0.081818	0.021178	0.837080	0.402693	0.635832
PGR	0.070527	0.308679	0.720813	0.471147	0.688093
EGFR	-0.046280	-0.306160	-0.691850	0.489147	0.688093
CD44	0.066715	0.038655	0.667443	0.504602	0.688093
H3F3B	-0.049260	0.133718	-0.518680	0.604066	0.787912
MTOR	0.046233	-0.057040	0.479503	0.631657	0.789572
MYC	0.036033	-0.053670	0.380451	0.703670	0.844404
CASP3	0.027063	-0.224020	0.299306	0.764752	0.850481
TWIST1	0.030437	-0.058700	0.298414	0.765433	0.850481
MKI67	-0.014180	-0.011680	-0.140220	0.888508	0.951973
PTPRC	-0.00288	-0.14908	-0.03257	0.974025	0.987457
ESR1	0.000747	0.502018	0.015724	0.987457	0.987457

REFERENCES

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Mackillop, W. J. The importance of prognosis in cancer medicine. *TNM Online* (2006) doi:10.1002/0471463736.tnmp01.pub2.
3. McAllister, S. S. & Weinberg, R. A. Tumor-host interactions: a far-reaching relationship. *J. Clin. Oncol.* **28**, 4022–4028 (2010).
4. Turashvili, G. & Brogi, E. Tumor Heterogeneity in Breast Cancer. *Front. Med.* **4**, 227 (2017).

5. Lüönd, F., Tiede, S. & Christofori, G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br. J. Cancer* **125**, 164–175 (2021).

6. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).

7. Casasent, A. K. *et al.* Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172**, 205–217.e12 (2018).

8. Yan, X. *et al.* Comprehensive description of the current breast cancer microenvironment advancements via single-cell analysis. *J. Exp. Clin. Cancer Res.* **40**, 142 (2021).

9. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).

10. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293–1308.e36 (2018).

11. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, 3588 (2018).

12. Ding, S., Chen, X. & Shen, K. Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Commun.* **40**, 329–344 (2020).

13. Saviano, A., Henderson, N. C. & Baumert, T. F. Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology. *J. Hepatol.* **73**, 1219–1230 (2020).

14. Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N. & Patrick, E. Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Front. Immunol.* **10**, (2019).

15. Ali, H. R. *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat Cancer* **1**, 163–175 (2020).

16. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).

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

18. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* **14**, e1006076 (2018).

19. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

20. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10869–10874 (2001).

21. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

22. Ma, X. *et al.* Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection. *Sci. Rep.* **9**, 1–8 (2019).

23. Frigyesi, A. & Höglund, M. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer Inform.* **6**, 275–292 (2008).

24. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

25. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

26. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

27. Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).

28. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

29. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).

30. Zhan, Z. *et al.* 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 Genom Bioinform* **3**, lqab015 (2021).

31. Wang, D., Jing, Z., He, K. & Garmire, L. X. Cox-nnet v2.0: improved neural-network-based survival prediction extended to large-scale EMR data. *Bioinformatics* **37**, 2772–2774 (2021).

32. Cox, D. R. *Regression Models and Life-tables*. (1972).

33. Dossus, L. & Benusiglio, P. R. Lobular breast cancer: incidence and genetic and non-genetic risk factors. *Breast Cancer Res.* **17**, 37 (2015).

34. Tekpli, X. *et al.* An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat. Commun.* **10**, 5499 (2019).

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

36. Menz, A. *et al.* Diagnostic and prognostic impact of cytokeratin 18 expression in human tumors: a tissue microarray study on 11,952 tumors. *Molecular Medicine* **27**, 1–16 (2021).

37. Jögi, A., Ehinger, A., Hartman, L. & Alkner, S. Expression of HIF-1α is related to a poor prognosis and tamoxifen resistance in contralateral breast cancer. *PLoS One* **14**, e0226150 (2019).

38. Mimeault, M. & Batra, S. K. Hypoxia-inducing factors as master regulators of stemness properties and altered metabolism of cancer- and metastasis-initiating cells. *J. Cell. Mol. Med.* **17**, 30–54 (2013).

39. Pellegrino, B. *et al.* Luminal Breast Cancer: Risk of Recurrence and Tumor-Associated Immune Suppression. *Mol. Diagn. Ther.* **25**, 409–424 (2021).

40. Kanyılmaz, G. *et al.* Prognostic Importance of Ki-67 in Breast Cancer and Its Relationship with Other Prognostic Factors. *Eur J Breast Health* **15**, 256–261 (2019).

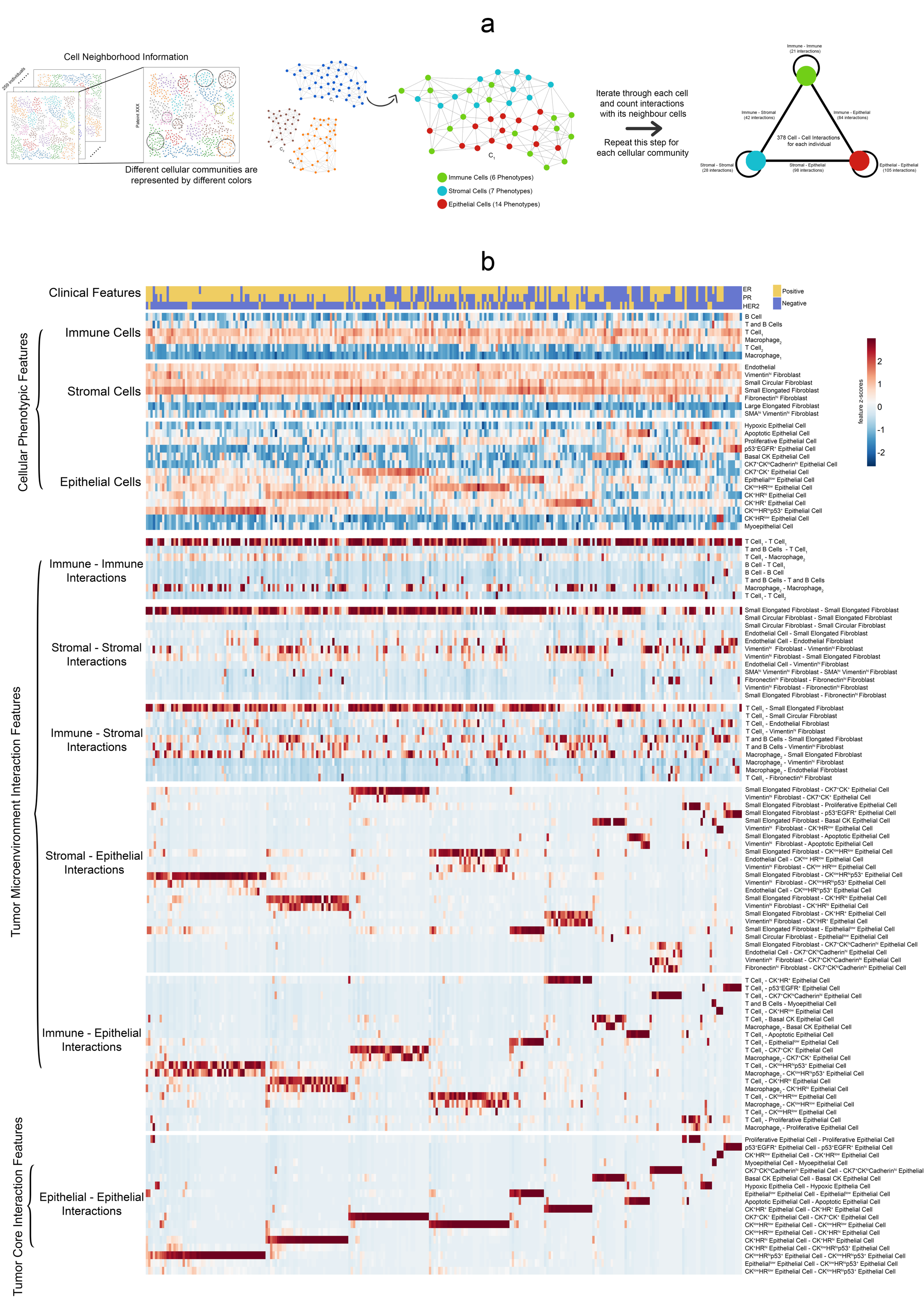
41. Yerushalmi, R., Woods, R., Ravdin, P. M., Hayes, M. M. & Gelmon, K. A. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol.* **11**, 174–183 (2010).

42. Liu, Z., Li, M., Jiang, Z. & Wang, X. A Comprehensive Immunologic Portrait of Triple-Negative

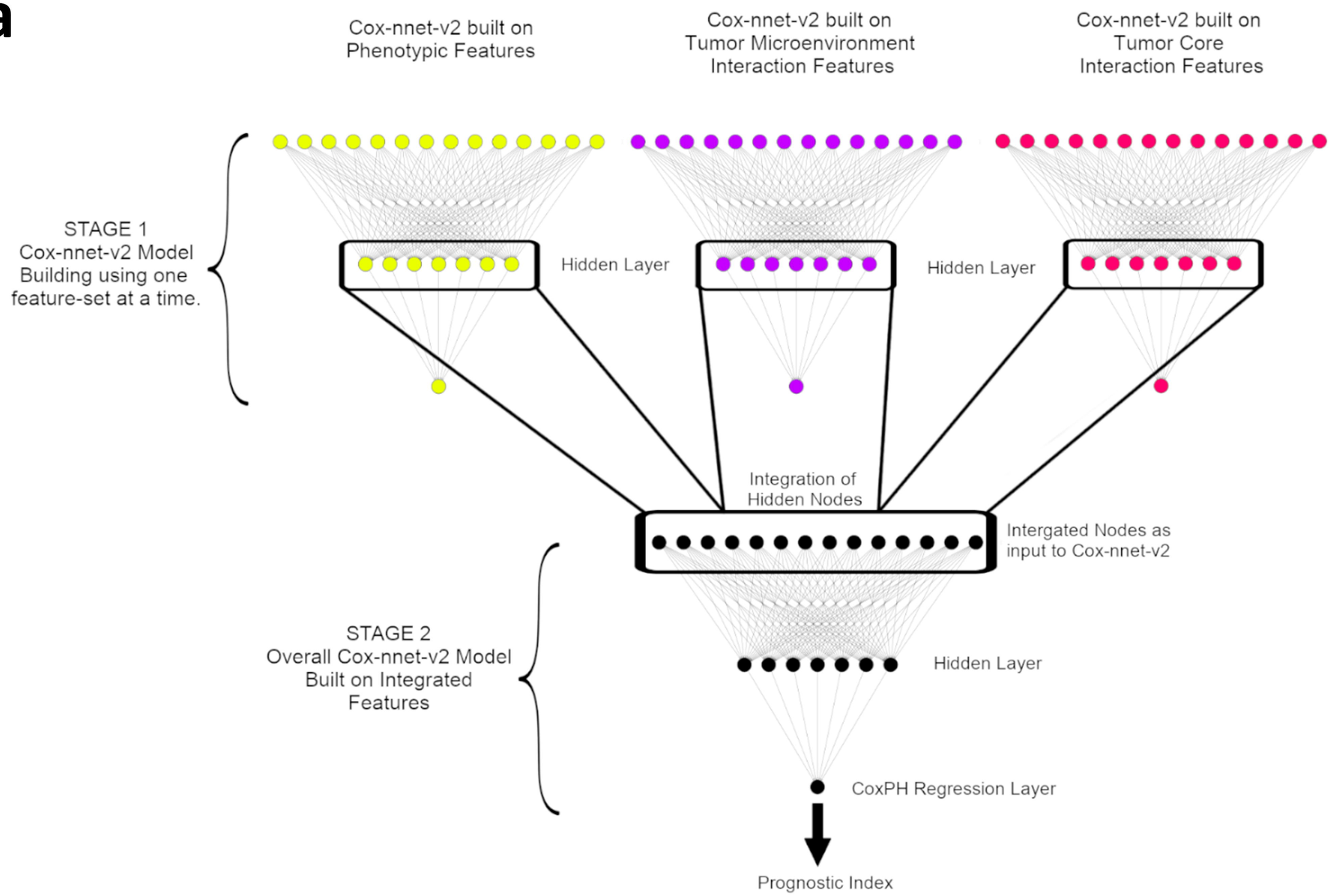
Breast Cancer. *Transl. Oncol.* **11**, 311–329 (2018).

43. Saccani, A. *et al.* p50 nuclear factor-kappaB overexpression in tumor-associated macrophages inhibits M1 inflammatory responses and antitumor resistance. *Cancer Res.* **66**, 11432–11440 (2006).
44. Zhang, H., Qin, G., Yu, H., Han, X. & Zhu, S. Comprehensive genomic and immunophenotypic analysis of CD4 T cell infiltrating human triple-negative breast cancer. *Cancer Immunol. Immunother.* **70**, 1649–1665 (2021).
45. Keith, B. & Simon, M. C. Hypoxia-inducible factors, stem cells, and cancer. *Cell* **129**, 465–472 (2007).
46. Kennedy, A. S. *et al.* Proliferation and hypoxia in human squamous cell carcinoma of the cervix: first report of combined immunohistochemical assays. *Int. J. Radiat. Oncol. Biol. Phys.* **37**, 897–905 (1997).
47. Mehra, R. *et al.* Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res.* **65**, 11259–11264 (2005).
48. An, Q. *et al.* KRT7 promotes epithelial-mesenchymal transition in ovarian cancer via the TGF-β/Smad2/3 signaling pathway. *Oncol. Rep.* **45**, 481–492 (2021).
49. Communal, L. *et al.* A Keratin 7 and E-Cadherin Signature Is Highly Predictive of Tubo-Ovarian High-Grade Serous Carcinoma Prognosis. *Int. J. Mol. Sci.* **22**, (2021).
50. Lee, H. W. *et al.* Alpha-smooth muscle actin (ACTA2) is required for metastatic potential of human lung adenocarcinoma. *Clin. Cancer Res.* **19**, 5879–5889 (2013).
51. Oka, H. *et al.* Expression of E-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis. *Cancer Res.* **53**, 1696–1701 (1993).
52. Soliman, N. A. & Yussif, S. M. Ki-67 as a prognostic marker according to breast cancer molecular subtype. *Cancer Biol Med* **13**, 496–504 (2016).
53. Grimwade, L. F., Fuller, K. A. & Erber, W. N. Applications of imaging flow cytometry in the diagnostic assessment of acute leukaemia. *Methods* **112**, 39–45 (2017).
54. Fischer, J. R. *et al.* Multiplex imaging of breast cancer lymph node metastases identifies prognostic single-cell populations independent of clinical classifiers. *Cell Rep Med* **4**, 100977 (2023).
55. Danenberg, E. *et al.* Breast tumor microenvironment structures are associated with genomic features

and clinical outcome. *Nat. Genet.* **54**, 660–669 (2022).


a
Cell Neighborhood Information
259 individuals
Patient XXX
Different cellular communities are represented by different colors
Immune Cells (6 Phenotypes)
Stromal Cells (7 Phenotypes)
Epithelial Cells (14 Phenotypes)
Iterate through each cell and count interactions with its neighbour cells
Repeat this step for each cellular community
Immune - Immune (21 interactions)
Immune - Stromal (42 interactions)
Immune - Epithelial (84 interactions)
378 Cell - Cell Interactions for each individual
Stromal - Stromal (28 interactions)
Stromal - Epithelial (98 interactions)
Epithelial - Epithelial (105 interactions)
b
Clinical Features
ER
PR
HER2
Positive
Negative
Cellular Phenotypic Features
Immune Cells
Stromal Cells
Epithelial Cells
feature z-scores
Tumor Microenvironment Interaction Features
Immune - Immune Interactions
Stromal - Stromal Interactions
Immune - Stromal Interactions
Stromal - Epithelial Interactions
Immune - Epithelial Interactions
Tumor Core Interaction Features
Epithelial - Epithelial Interactions

a

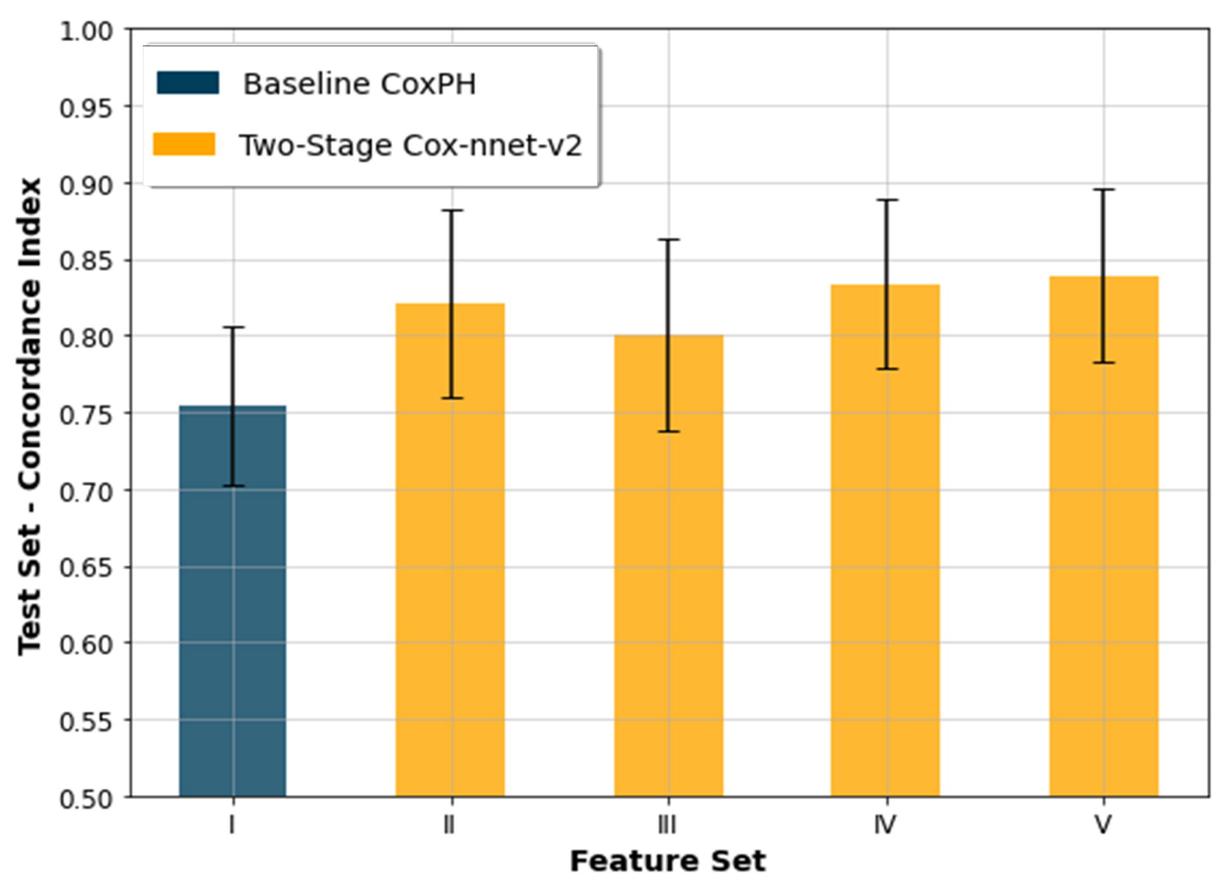


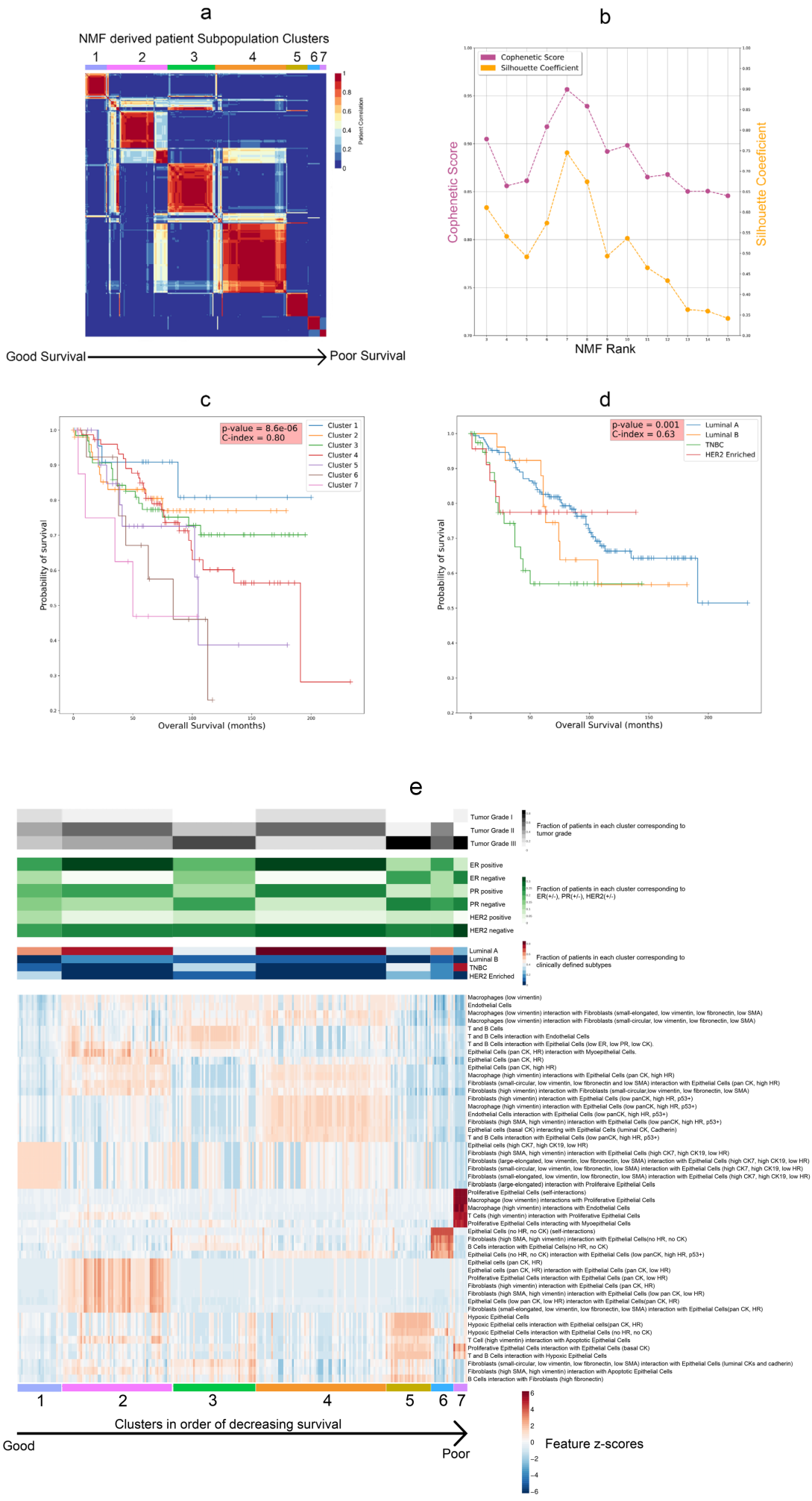
b

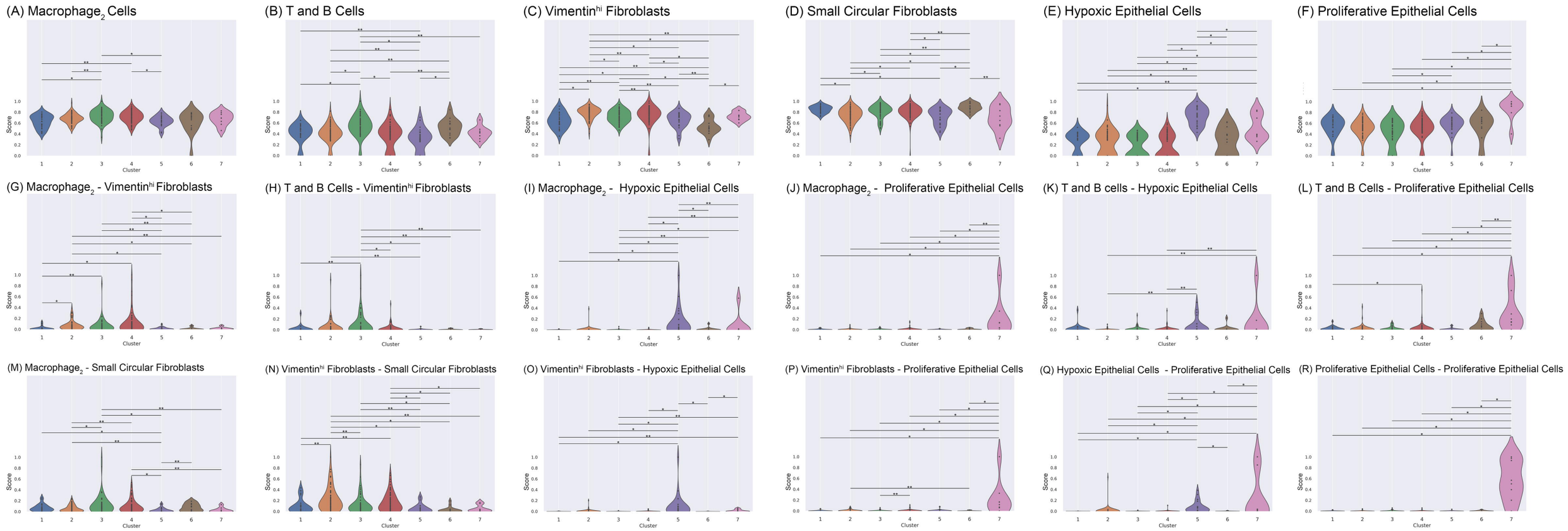
Legend

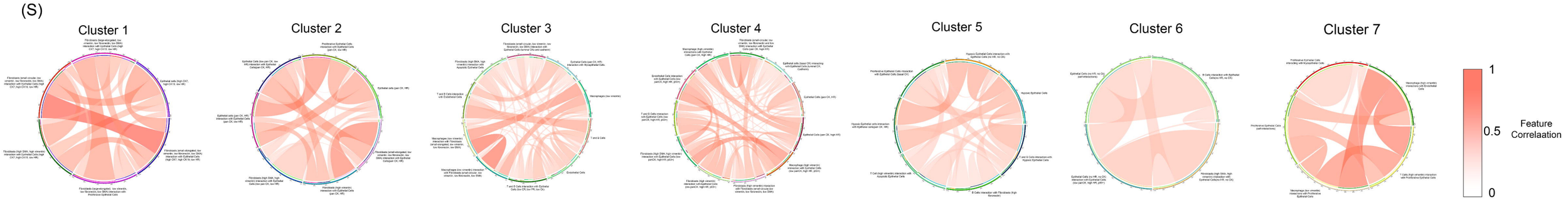
Set	Feature Set
I	Clinical
II	Phenotypic* + TMI*
III	Phenotypic* + TCI*
IV	TMI* + TCI*
V	Phenotypic* + TMI* + TCI*

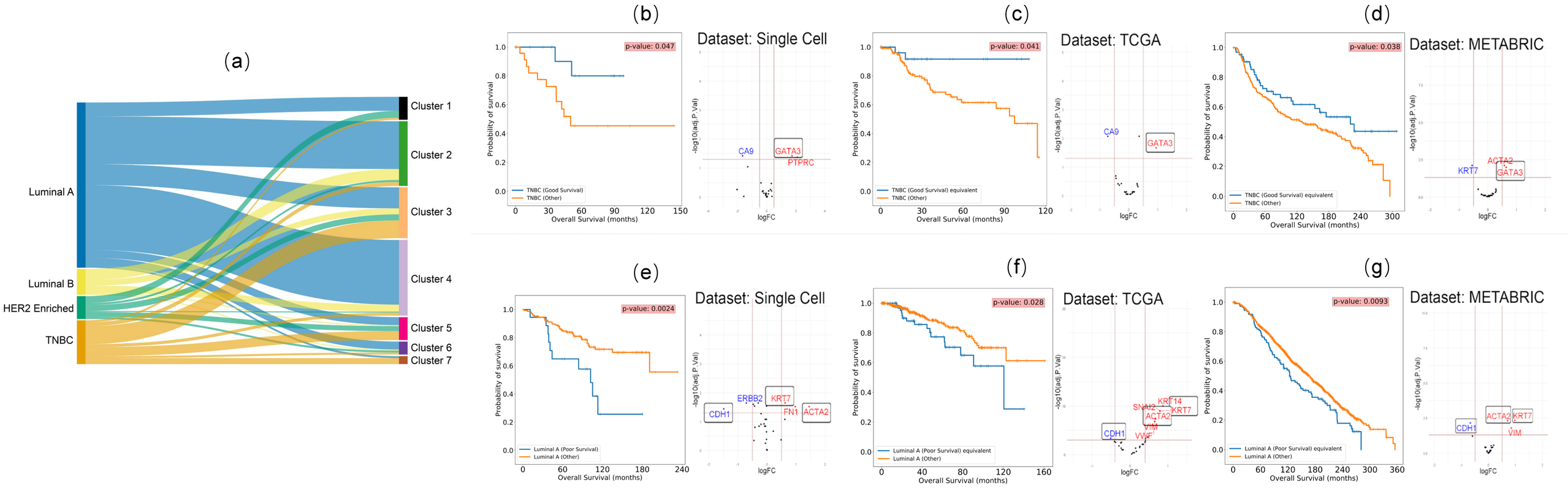
*Denotes hidden layer features obtained from cox-nnet-v2




a
NMF derived patient Subpopulation Clusters
1 2 3 4 5 6 7
Patient Correlation
Good Survival
Poor Survival
b
Cophenetic Score
Silhouette Coefficient
NMF Rank
c
p-value = 8.6e-06
C-index = 0.80
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Probability of survival
Overall Survival (months)
d
p-value = 0.001
C-index = 0.63
Luminal A
Luminal B
TNBC
HER2 Enriched
Probability of survival
Overall Survival (months)
e
Tumor Grade I
Tumor Grade II
Tumor Grade III
Fraction of patients in each cluster corresponding to tumor grade
ER positive
ER negative
PR positive
PR negative
HER2 positive
HER2 negative
Fraction of patients in each cluster corresponding to ER(+/-), PR(+/-), HER2(+/-)
Luminal A
Luminal B
TNBC
HER2 Enriched
Fraction of patients in each cluster corresponding to clinically defined subtypes
Macrophages (low vimentin)
Endothelial Cells
Macrophages (low vimentin) interaction with Fibroblasts (small-elongated, low vimentin, low fibronectin, low SMA)
Macrophages (low vimentin) interaction with Fibroblasts (small-circular, low vimentin, low fibronectin, low SMA)
T and B Cells
T and B Cells interaction with Endothelial Cells
T and B Cells interaction with Epithelial Cells (low ER, low PR, low CK).
Epithelial Cells (pan CK, HR) interaction with Myoepithelial Cells.
Epithelial Cells (pan CK, HR)
Epithelial Cells (pan CK, high HR)
Macrophage (high vimentin) interactions with Epithelial Cells (pan CK, high HR)
Fibroblasts (small-circular, low vimentin, low fibronectin and low SMA) interaction with Epithelial Cells (pan CK, high HR)
Fibroblasts (high vimentin) interaction with Fibroblasts (small-circular,low vimentin, low fibronectin, low SMA)
Fibroblasts (high vimentin) interaction with Epithelial Cells (low panCK, high HR, p53+)
Macrophage (high vimentin) interaction with Epithelial Cells (low panCK, high HR, p53+)
Endothelial Cells interaction with Epithelial Cells (low panCK, high HR, p53+)
Fibroblasts (high SMA, high vimentin) interaction with Epithelial Cells (low panCK, high HR, p53+)
Epithelial cells (basal CK) interacting with Epithelial Cells (luminal CK, Cadherin)
T and B Cells interaction with Epithelial Cells (low panCK, high HR, p53+)
Epithelial cells (high CK7, high CK19, low HR)
Fibroblasts (high SMA, high vimentin) interaction with Epithelial Cells (high CK7, high CK19, low HR)
Fibroblasts (large-elongated, low vimentin, low fibronectin, low SMA) interaction with Epithelial Cells (high CK7, high CK19, low HR)
Fibroblasts (small-circular, low vimentin, low fibronectin, low SMA) interaction with Epithelial Cells (high CK7, high CK19, low HR)
Fibroblasts (small-elongated, low vimentin, low fibronectin, low SMA) interaction with Epithelial Cells (high CK7, high CK19, low HR)
Fibroblasts (large-elongated) interaction with Proliferaive Epithelial Cells
Proliferative Epithelial Cells (self-interactions)
Macrophage (low vimentin) interactions with Proliferative Epithelial Cells
Macrophage (high vimentin) interactions with Endothelial Cells
T Cells (high vimentin) interaction with Proliferative Epithelial Cells
Proliferative Epithelial Cells interacting with Myoepithelial Cells
Epithelial Cells (no HR, no CK) (self-interactions)
Fibroblasts (high SMA, high vimentin) interaction with Epithelial Cells(no HR, no CK)
B Cells interaction with Epithelial Cells(no HR, no CK)
Epithelial Cells (no HR, no CK) interaction with Epithelial Cells (low panCK, high HR, p53+)
Epithelial cells (pan CK, HR)
Epithelial cells (pan CK, HR) interaction with Epithelial Cells (pan CK, low HR)
Proliferative Epithelial Cells interaction with Epithelial Cells (pan CK, low HR)
Fibroblasts (high vimentin) interaction with Epithelial Cells (pan CK, HR)
Fibroblasts (high SMA, high vimentin) interaction with Epithelial Cells (low pan CK, low HR)
Epithelial Cells (low pan CK, low HR) interacton with Epithelial Cells(pan CK, HR)
Fibroblasts (small-elongated, low vimentin, low fibronectin, low SMA) interaction with Epithelial Cells(pan CK, HR)
Hypoxic Epithelial Cells
Hypoxic Epithelial cells interaction with Epithelial cells(pan CK, HR)
Hypoxic Epithelial Cells interaction with Epithelial Cells (no HR, no CK)
T Cell (high vimentin) interaction with Apoptotic Epithelial Cells
Proliferative Epithelial Cells interaction with Epithelial Cells (basal CK)
T and B Cells interaction with Hypoxic Epithelial Cells
Fibroblasts (small-circular, low vimentin, low fibronectin, low SMA) interaction with Epithelial Cells (luminal CKs and cadherin)
Fibroblasts (high SMA, high vimentin) interaction with Apoptotic Epithelial Cells
B Cells interaction with Fibroblasts (high fibronectin)
1 2 3 4 5 6 7
Clusters in order of decreasing survival
Good
Poor
Feature z-scores


(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)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
1
0.5
0
Feature Correlaation


(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
Probability of survival
Overall Survival (months)
TNBC (Good Survival)
TNBC (Other)
Dataset: Single Cell
-log10(adj.P.Val)
logFC
CA9
GATA3
PTPRC
(c)
p-value: 0.041
TNBC (Good Survival) equivalent
TNBC (Other)
Dataset: TCGA
CA9
GATA3
(d)
p-value: 0.038
TNBC (Good Survival) equivalent
TNBC (Other)
Dataset: METABRIC
KRT7
ACTA2
GATA3
(e)
p-value: 0.0024
Luminal A (Poor Survival)
Luminal A (Other)
Dataset: Single Cell
ERBB2
KRT7
CDH1
FN1
ACTA2
(f)
p-value: 0.028
Luminal A (Poor Survival) equivalent
Luminal A (Other)
Dataset: TCGA
KRT14
SNAI2
KRT7
ACTA2
VIM
CDH1
VWF
(g)
p-value: 0.0093
Luminal A (Poor Survival) equivalent
Luminal A (Other)
Dataset: METABRIC
ACTA2
KRT7
CDH1
VIM