Supplementary Information for "Explainable machine learning–guided integrated multiomics analysis reveals macrophage-driven immune suppression in breast cancer"

(Dated: April 18, 2026)

I. SUPPLEMENTARY METHODS

A. Deconvolution

Deconvolution of bulk gene expression or RNA-seq data is the process of estimating the proportions of different cell types that contribute to the overall measurement. Assuming the final measurement is a linear combination of the expressions from various cell types, it can be written as [1]:

$$\mathbf{P}_j = \mathbf{SM} \cdot \boldsymbol{\beta}_j + \boldsymbol{\epsilon} \tag{1}$$

In this equation, $\mathbf{P}j$ represents the measured expression/RNA-seq values from tumor sample j, $\mathbf{SM}$ is the signature matrix, $\boldsymbol{\beta}j$ is a vector of mixing fractions for sample j, and $\boldsymbol{\epsilon}$ represents the noise. For Pij, the expression of gene i in sample j, we have: $Pij = \sum_k SM_{ik}\beta_{kj}$, where SM_{ik} is the average expression of gene i in cell type k and β_{kj} is the fraction of cell type k in sample j.

We used 10 signature matrices from our previous work [2], created using high-resolution scRNA-seq data from [3]. To estimate cell fractions, we used a common method called CIBERSORTx [4]. Bulk RNA-seq data from TCGA and GEPs data from MBRC were downloaded from cBioPortal. TCGA data were log2-normalized and scaled, while MBRC data were scaled. We estimated cell fractions using either the "Impute Cell Fractions" module of the CIBERSORTx webtool with default settings or the Docker version of CIBERSORTx with similar settings. We then averaged the results from all 10 estimates.

B. NAC cohort

Inclusion criteria were breast tumors that underwent NAC, with available bGEP microarray data and information on treatment response. We identified a total of 15 datasets, which are presented in Table S1, along with treatment regimen, subtypes, and outcome evaluation. The majority of datasets relied on either the Residual Cancer Burden (RCB) classification or the standard binary definition of pCR (absence of residual invasive cancer in breast and axillary lymph nodes, with carcinoma in situ permitted). In some cases, studies applied a study-specific responder vs non-responder classification that combined clinical examination (tumor shrinkage or RECIST criteria) with pathologic findings at surgery. Notably, none of the included cohorts employed the Miller–Payne (MP) grading system.

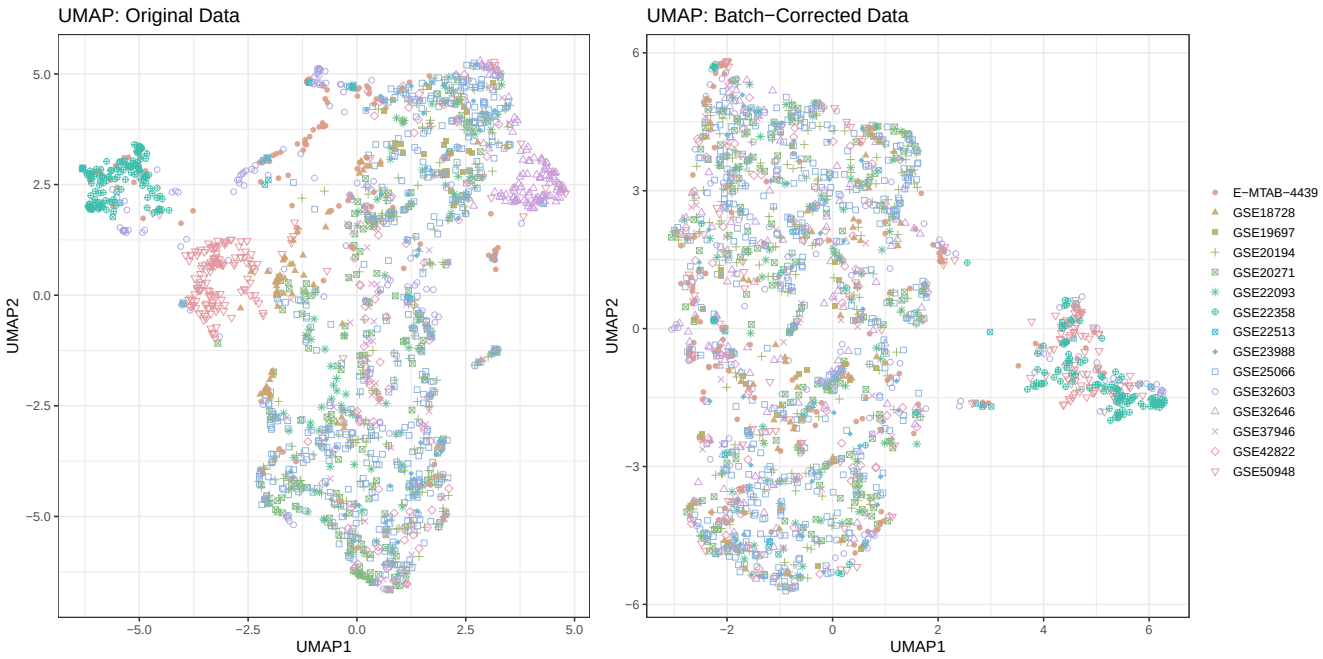
Datasets								
Expression Data Access Number /Reference Paper for Results	Information / Name and Clinical Trial Number	Number of Samples	Stage	Treatment	Her2	ER	Outcome Measurement	Cohort
E-MTAB-4439 silwal2017 [5]	Randomized, Phase II clinical trial for Avastin in combination with neoadjuvant treatment (NeoAva) NCT00773695	131	< IV	Fluorouracil, Epirubicin and Cyclophosphamide, Taxanes (Docetaxel or Paclitaxel) + Bevacizumab	Neg	Mix	RCB	Disc
GSE18728/ korde2010 [6]	Phase II trial with protocol available at lebowitz2004 [7]	61	II/III	Docetaxel and Capecitabine	Neg	Mix	study-specific responder / non-responder classification	Disc
GSE19697 lin2010 [8]	Prospective clinical trial at Washington University (WASH2) with protocol available at lin2010 [8]	24	-	Epirubicin and Docetaxel+ Zoledronic Acid	Neg	Neg	pCR vs RD	Disc
GSE20194 popovici2010 [9]	Prospective pharmacogenomic marker discovery study with corresponding protocol available at shi2010 [10]	278	I-III	Paclitaxel, 5-fluorouracil, Cyclophosphamide and Doxorubicin	Mix	Mix	pCR vs RD	Disc
GSE20271 tabchy2010 [11]	Prospective, randomized, international clinical trial with protocol available at tabchy2010 [11]	178	I-III	Paclitaxel, 5-fluorouracil, Cyclophosphamide and Doxorubicin	Mix	Mix	pCR vs RD	Disc
GSE22093 iwamoto2011 [12]	MDACC/IGR dataset with protocol available at iwamoto2011 [12]	97	I-III	Fluorouracil, Adriamycin, and Cytoxan	Nor	Mix	pCR vs RD	Disc
GSE22358 gluck2012[13]	A multicenter, open-label study on treating operable early-stage breast cancer with further information available at gluck2012[13]	122	I-III	Capecitabine, Docetaxel and Trastuzumab	Mix	Mix	RCB	Disc
GSE42822 shen2012 [14]	Phase II trial with protocol available at shen2012cell [14]	90	II-III	5-fluorouracil, Epirubicin, Cyclophosphamide; Docetaxel Capecitabine	Mix	Mix	pCR vs RD	Disc
GSE22513 (VAN) bauer2010 [15]	High-risk, operable breast cancers treated at Vanderbilt University or New York University with the protocol available at: bauer2010 [15]	28	II -III	Paclitaxel + Radiation	Mix	Neg	pCR vs RD	Disc
GSE25066 hatzis2011 [16]	Prospective multicenter study with protocols according to hatzis2011 [16]	482	II -III	Paclitaxel, Docetaxel with Capecitabine	Mix	Mix	RCB	Valid
GSE32603 magbanua2015 [17]	Clinical trial to study magnetic resonance imaging (I-SPY1) with clinical trial number of NCT00033397	138	< IV	Taxane	Mix	Mix	RCB	Valid
GSE32646 miyake2012 [18]	Primary breast cancer patients at Osaka University Hospital (Osaka) miyake2012 [18]	115	II -III	Paclitaxel, 5-fluorouracil, Epirubicin and Cyclophosphamide	Mix	Mix	pCR vs RD	Valid
GSE37946 liu2012 [19]	HER2+ patients received treatment at the MDACC (details here: liu2012 [19]).	50	II -III	Fluorouracil/ Epirubicin or Adriamycin/ Cyclophosphamide - Taxol plus Trastuzumab	Pos	Mix	pCR vs RD	Valid
GSE50948 prat2014 [20]	NOAH trial (randomized clinical trial) ISRCTN86043495 and control cohort of 42 HER2- tumors	156	< IV	Doxorubicin/Paclitaxel, Cyclophosphamide Methotrexate, Fluorouracil+ Trastuzumab, AT-CMF for control cohort	Mix	Mix	pCR vs RD	Valid
GSE23988 iwamoto2011 [12]	Phase II trial with details available at iwamoto2011 [12]	61	II -III	5-fluorouracil, Doxorubicin, Cyclophosphamide, Docetaxel and Capecitabine	Mix	Mix	pCR vs RD	Valid

Table S1: NAC cohort trials and their corresponding bGEPs datasets.

1. *Batch Correction*

To assess the potential impact of batch effects on NAC cohort, we initially attempted to correct for batch variation using the ComBat algorithm [21] (see I E 4 for discussion on TCGA and MBRC). For batch correction with ComBat, the fractions matrix was transposed so that rows corresponded to features and columns to samples. A null model design matrix was constructed, and ComBat was applied to adjust for batch effects. The corrected data were then transposed back, ensuring that each row represented a sample.

Uniform Manifold Approximation and Projection (UMAP) was performed on both the original and the batch-corrected data to visually inspect batch-related clustering, as shown in Fig. S1 (a). Additionally, kBET (k-nearest neighbor Batch Effect Test) was used to quantitatively assess batch mixing, with rejection rates compared between the original and corrected datasets (see Fig. S1 (b).



(a)

(b) kBET

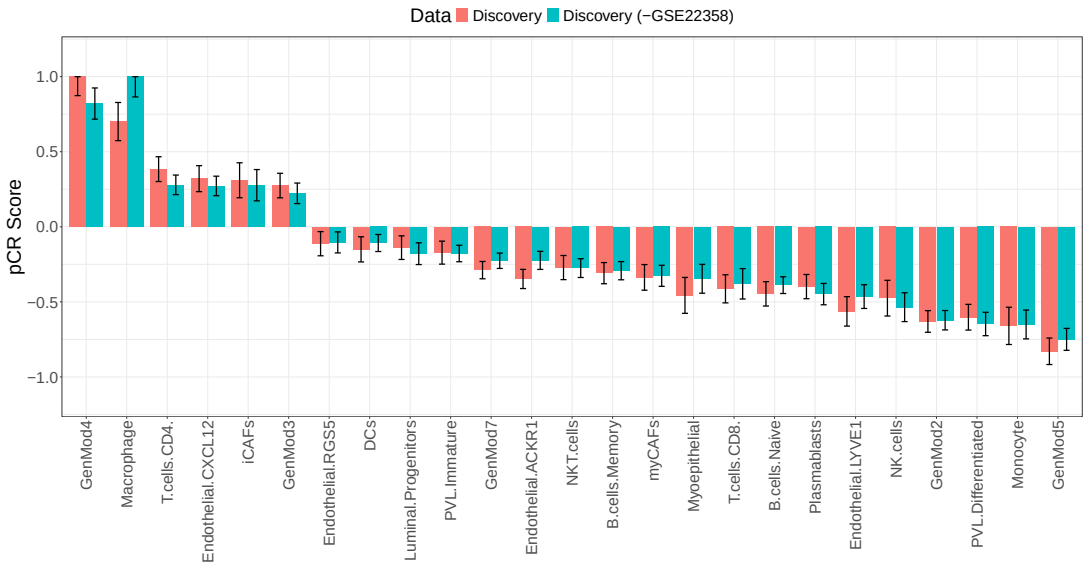
Figure S1: UMAP for the NAC cohort before and after batch correction using ComBat (a), and kBET values in both scenarios (b). As the comparison shows, while batch correction has an effect, it is limited. Source data are provided as a Source Data file.

Although batch correction with ComBat resulted in a modest decrease in batch-associated metrics (e.g., kBET observed rejection rate decreased from 100% to 80%, the improvement was limited. Furthermore, our data represent fractions and negative values produced by some batch correction methods would be biologically implausible. Given the limited benefit and the need to preserve the inherent non-negativity and biological interpretability of the data, we opted not to apply batch correction to the final analyses.

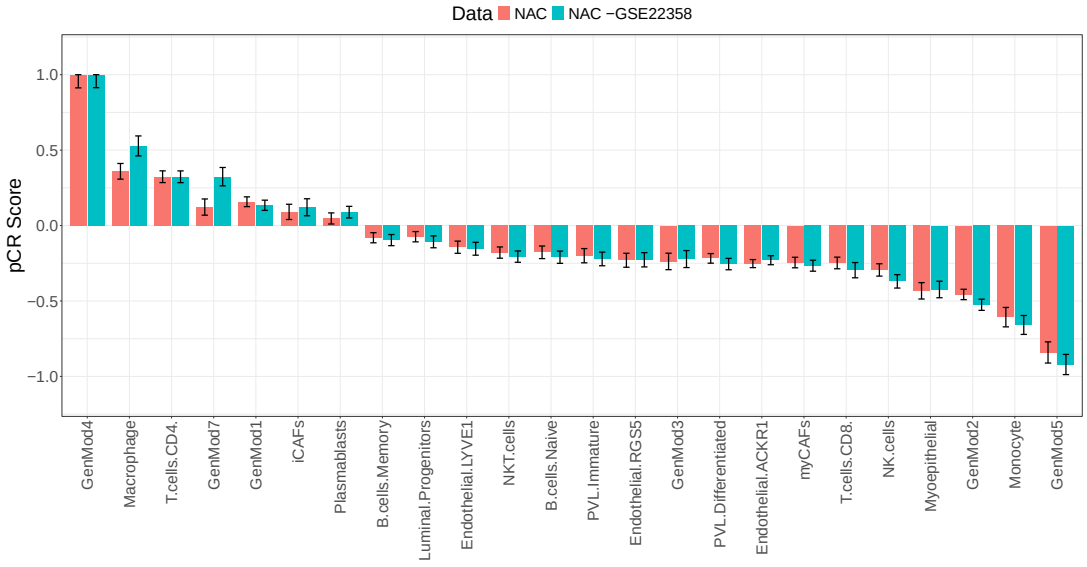
To avoid any bias to data sets, NAC cohort was randomly divided the data into discovery (9 datasets, 1009 samples) and validation (6 datasets, 998 samples) subsets. Then, pCR scores from the two subsets were cross validated against each others, keeping only cell types with consistent scores across two subsets. The same approach was applied to all tumor subtypes as well (see [2] for more details).

Excluding a dataset (GSE22358) To further assess the potential impact of batch effects, we performed a sensitivity analysis in which the dataset GSE22358, due to its distinct separation from the rest of data in UMAP plots, was excluded from the NAC discovery cohort. We retrained the model without this dataset and recalculated pCR scores for both the discovery set and the full NAC cohort.

As shown in Fig. S2, the results with and without GSE22358 were highly consistent. These findings suggest that while some individual cohorts may display stronger batch-related signatures, our overall framework, based on discovery/validation partitioning and cross-validation across independent datasets, remains robust. It also should be noted that batch effects may sometimes be driven by specific features (e.g., non-coding RNAs measured with differing methodologies), and that excluding such genes or, in certain cases, excluding problematic cohorts could be a reasonable strategy.



(a) Discovery cohort



(b) Full NAC cohort

Figure S2: Comparison of pCR scores with and without GSE22358. (a) Discovery cohort vs. discovery cohort excluding GSE22358. (b) Full NAC cohort vs. NAC cohort excluding GSE22358. Source data are provided as a Source Data file.

C. Breast tumor subtypes

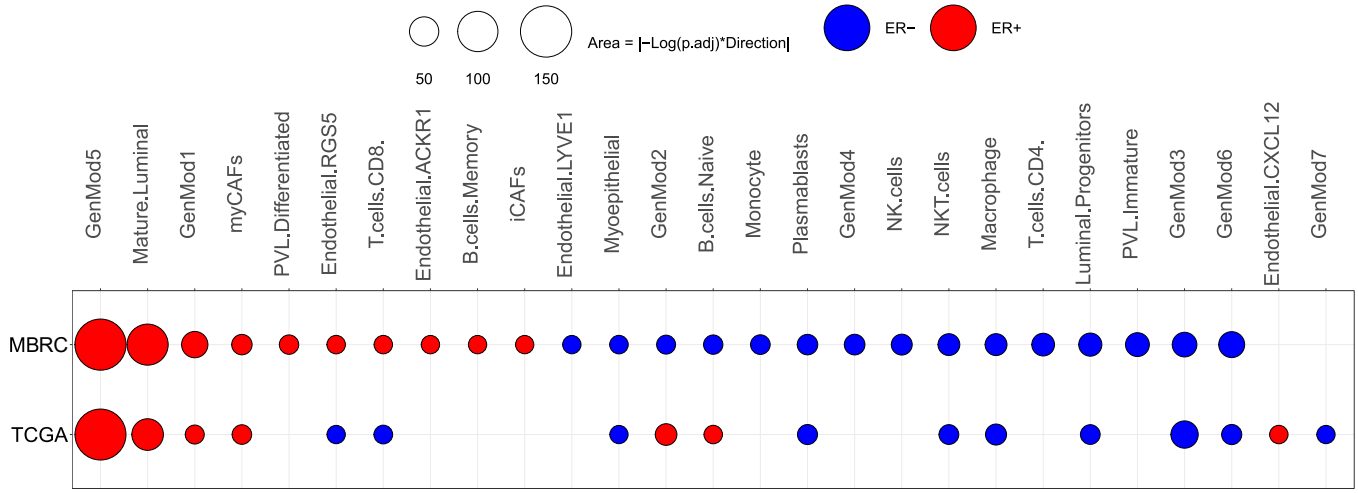
Breast cancer is a heterogeneous disease with several subtypes defined by distinct histological and molecular characteristics and prognosis. Main subtypes characterized by the status of Estrogen Receptor (ER), Progesterone Receptor (PR) and Human Epidermal Growth Factor Receptor 2 (HER2) [22], or by their expression profile using the PAM50 method [23]. Most ER+ tumors include the Luminal A (LumA) and Luminal B (LumB) subtypes; these generally respond well to endocrine therapies and have a relatively favorable prognosis. In contrast, ER- tumors, which includes most of Basal-like and HER2-enriched subtypes, tend to be more aggressive, with limited targeted treatment options available. Table S2 includes frequency of these subtypes across all data used in this study (PAM50 Normal-like subtype was not explored in this study).

Data	MBRC	TCGA	NAC	IMC	scRNA-seq
ER+	1347	751	1097	539	26
ER-	417	219	910	142	31
LumA	597	360	589	227	15
LumB	466	221	464	189	8
Basal	310	196	553	111	17
Her2	244	123	252	77	10
GP	1764	971	2007	681	57

Table S2: Number of samples in GP and subtypes of MBRC, TCGA, NAC, IMC and scRNA-seq data.

1. *TME composition across ER subtypes*

Differences between breast cancer subtypes is also reflected in the composition of TME. Here, we compared the fraction of cell types across ER+ and ER- samples in both TCGA and MBRC, revealing consistent differences in the fraction of cancer and immune cells, as shown in Fig. S3.



(a)

Figure S3: Comparison of cell type fractions in ER+ vs. ER- samples. Source data are provided as a Source Data file.

D. Machine Learning Models

1. *Model Selection*

In this study, three popular modeling options were selected for survival analysis: Gradient Boost Machine (GBM) for survival, Random Survival Forest (RSF), and Survival Support Vector Machine (SSVM). Hyperparameter optimization was carried out using GridSearchCV for each of the models.

2. *Evaluation Metric*

The concordance index (C-index), also referred to as Harrell's C [24], is a widely used performance measure in survival analysis and risk prediction models. It assesses how well a model's predicted risk scores agree with the actual ordering of events (such as time to failure or death). Intuitively, the C-index represents the probability that, for a randomly selected pair of subjects, the subject with the higher predicted risk experiences the event earlier than the subject with the lower predicted risk.

Consider a dataset with n subjects, where each subject i is characterized by $(t_i, \delta_i, \hat{f}(x_i))$ where: t_i is the observed time (which could be either the event time or the censoring time), δ_i is an indicator variable such that $\delta_i = 1$ if the event is observed and $\delta_i = 0$ if the observation is censored, $\hat{f}(x_i)$ is the predicted risk score for subject i. For such a configuration, the C-index is defined mathematically as [25]:

$$C = \frac{\sum_{i,j} \mathbb{1}\left\{\hat{f}(x_i) > \hat{f}(x_j)\right\} \cdot \mathbb{1}\{t_i < t_j\}}{\sum_{i,j} \mathbb{1}\{t_i < t_j\}} \tag{2}$$

where: $\mathbb{1}\{\cdot\}$ is the indicator function, which equals 1 if the condition is true and 0 otherwise. The summation is over all comparable pairs (i, j) such that subject i experiences the event before subject j (i.e., $t_i < t_j$). The numerator counts the number of pairs for which the predicted risk scores are correctly ordered (i.e., the subject with the earlier event has a higher risk score). Denominator represents the total number of comparable pairs. A C-index value of 1 indicates perfect concordance, meaning the model perfectly ranks all subjects according to their risk. A value of 0.5 suggests that the model's predictions are no better than random guessing.

3. Hyperparameter Tuning using GridSerachCV

MBRC

For RSF, the grid search was performed over `n_estimators` in the range 50 to 450 (increments of 50), `min_samples_split` as 5, 10, 15, and 20, and `min_samples_leaf` from 3 to 27 (increments of 3). The optimal performance was achieved consistently with 50 estimators and a `min_samples_leaf` of 18, while the `min_samples_split` varied among 5, 10, 15, and 20. All these configurations resulted in a C-Index of 0.69 ($\pm$ 0.01).

For SSVM, the parameter grid was defined as `alpha` as 2.2^i (with $i = -20, \ldots, 0$), `max iter`=j as (with $j = 0, \ldots, 6$), and `tol` as 2.2^k with k (with $k = -10, \ldots, 0$). The optimal parameters were an `alpha` of 7.8×10^{-5}, `max_iter` equal to 2, and a `tol` of 3.8×10^{-4}, yielding a C-Index of 0.683.

For GBM, the hyperparameters were tuned by varying `n_estimators` as $10+30i$ (with $i = 1, \ldots, 5$), `learning_rate` as 1.8^j (with $j = -8, \ldots, -1$), and `max_depth` as k (with $k = 1, \ldots, 10$). The best model obtained had 100 estimators, a learning rate of 0.095, and a maximum depth of 2, resulting in a C-Index of 0.70 ($\pm$0.02).

TCGA

For RSF, the grid search was performed over `n_estimators` in the range 50 to 450 (increments of 50), `min_samples_split` as 5, 10, 15, and 20, and `min_samples_leaf` from 3 to 27 (increments of 3). The optimal performance was achieved with 150 estimators and a `min_samples_leaf` of 18, while the `min_samples_split` could be any of 5, 10, 15, or 20. All of these configurations resulted in a C-Index of 0.676 ($\pm$0.041).

For SSVM, the parameter grid was defined as `alpha` in 3^i (with $i = -23, \ldots, -3$), `max_iter` in $\{0, 1, 2, 3\}$, and `tol` in 4^j (with $j = -5, \ldots, -1$). The optimal parameters were `alpha` $= 6.27 \times 10^{-7}$, `max_iter` $= 2$, and `tol` $= 9.76 \times 10^{-4}$, yielding a C-Index of 0.731.

For GBM, the hyperparameters were tuned by varying `n_estimators` from 50 to 300 (in increments of 50), `learning_rate` in 1.9^k (with $k = -6, \ldots, -1$), and `max_depth` in $[1, 10]$. The best model obtained had 100 estimators, a learning rate of 0.02, and a maximum depth of 3 leading to C-index of 0.69 ($\pm$ 0.06).

4. Comparison to Classical Cox Proportional Hazards Model

In order to assess the performance of the machine learning models, a classical Cox Proportional Hazards model was also considered. The Cox model was tuned using GridSearchCV with a dynamic parameter grid for the regularization parameter `alpha`, defined as 1.5^j for $j = -10, \ldots, 9$. For MBRC, the optimal value obtained was $\alpha = 38.4$, which resulted in a best cross-validated concordance index of 0.65. When the final model was evaluated on the test set, it achieved a concordance index of 0.64. In TCGA, the same optimal value was obtained for α, resulting in a best cross-validated C-index of 0.7 and C-index of 0.67 on test set.

5. Comparison to existing ML models

The obtained C-index of approximately 0.7 indicates a robust discriminative ability in predicting relapse-free survival. This level of performance is in line with—and in some cases exceeds—that of many existing models [26].

Classical methods such as the Cox Proportional Hazards model typically yield lower C-index values, while ensemble approaches like RSF and other modern techniques have reported C-index values in a similar range (e.g., 0.65–0.73) in comparable studies. Thus, a C-index of 0.7 is considered reasonably good, reflecting a strong capability to differentiate between patients with varying risks of relapse. It should be noted that prediction of overall survival can lead to better performance [27], mainly due to inclusion of relapse-free survival as a predictive feature in such models.

E. Model Interpretation and SHAP Analysis

After developing models predicting risk of relapse, we focus on understanding the specific roles of different cell types in driving these predictions. To evaluate the contribution of individual features, we applied SHAP (SHapley Additive exPlanations) [28], a post-hoc explanation method that computes the marginal contribution of each feature to the model's output on a per-sample basis. In our analysis, a positive SHAP value for a given feature indicates its propensity to elevate the likelihood of relapse, while the absolute magnitude of the SHAP value serves as an indicator of the feature's overall importance.

1. Theoretical Foundation

SHAP is one of the most widely used explainable machine learning (XML) approaches, grounded in the concept of Shapley values from cooperative game theory [29]. Named after Lloyd Shapley, who was awarded the Nobel Memorial Prize in Economic Sciences in 2012, the Shapley value fairly distributes the total "gain" (or contribution) among all participating features. For a given feature i, the SHAP value is defined as:

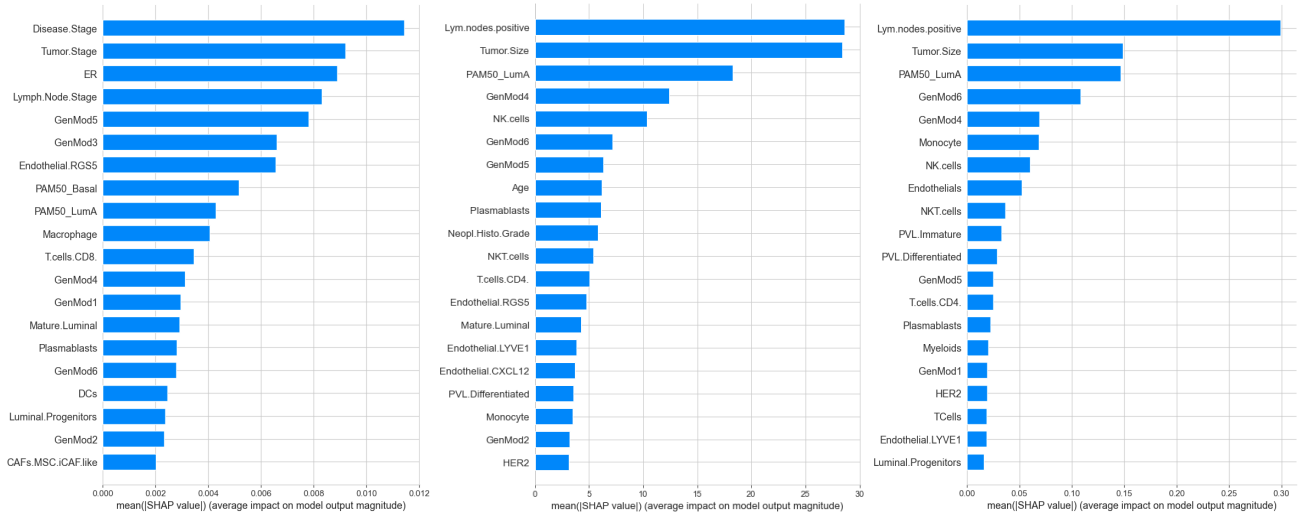
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right) \tag{3}$$

In this equation, $f_{S \cup \{i\}}$ denotes the model trained with feature i included, while f_S is the model trained with feature i withheld. The vector x_S represents the values of the input features in the subset S. By computing the contribution of feature i across all possible subsets $S \subseteq F \setminus \{i\}$, SHAP assigns a fair share of importance to each feature based on its influence on the prediction.

2. Model specific bias

SHAP analysis was conducted on the various models to assess the contribution of each feature to the predictions. As anticipated, the SHAP values differ among models, with the same feature sometimes receiving different contributions depending on the model. This variation is reflective of model-specific biases inherent to each algorithm. In other words, each model captures the relationships in the data differently—owing to differences in their assumptions, structure, and capacity to model non-linear interactions—which in turn affects how feature importance is calculated. These differences underscore that while SHAP values offer valuable insights into feature contributions, their interpretation must account for the context of the specific model, as the inherent bias in each machine learning method can influence the resulting values.

To account for the variation in SHAP values across different models, SHAP analyses were performed separately on each dataset (TCGA and MBRC) using models that provided meaningful feature contributions. Initially, it was observed that GBM tended to assign very small contributions to many features (see Fig. S4), which made its SHAP values impractical for our objectives. Consequently, GBM was excluded from the cross-validation process. Instead, the remaining two models were utilized for cross-validation to mitigate model-specific bias and to ensure a robust assessment of feature importance across datasets.



(a) SSVM (b) RSF (c) GBM

3TCGA

33

Figure S4: SHAP values from different models for the MBRC dataset. SSVM (a) and RSF (b) show a smoother decline in feature importance scores, whereas GBM (c) shows a sharp drop in SHAP values, with only the top features making meaningful contributions. Source data are provided as a Source Data file.

3. Survival Score Calculations

For each cell type, we begin by obtaining its SHAP values from the survival models, which reflect how the cell type's estimated fraction contributes to the risk of relapse. In cases where the SHAP values show a clear linear association with the cell fractions—indicative of a consistent effect on survival risk—we quantify this relationship by fitting a linear model and computing its slope. This slope, after normalization, serves as the measure of the cell type's contribution to survival risk, which we refer to as the Survival Score.

To mitigate the influence of model-specific biases, this process is performed independently using two different models (RSF and SSVM). Only cell types that demonstrate a similar association in both models (i.e., a comparable slope sign and magnitude) are considered valid. For these cell types, the normalized slopes from the two models are averaged to produce the final Survival Score, and the cell types are subsequently ranked based on these scores.

Implementation of our pipeline Using MBRC as an example and the same color scheme as Fig. 1, we show how the pipeline works in Fig. S5. Out of 28 cell types, 4 do not show a clear correlation (green). When comparing the two models, 2 more cell types have opposite associations and are dropped. The remaining 22 cell types, which show either a positive (blue) or negative (red) association, are used to calculate survival scores.

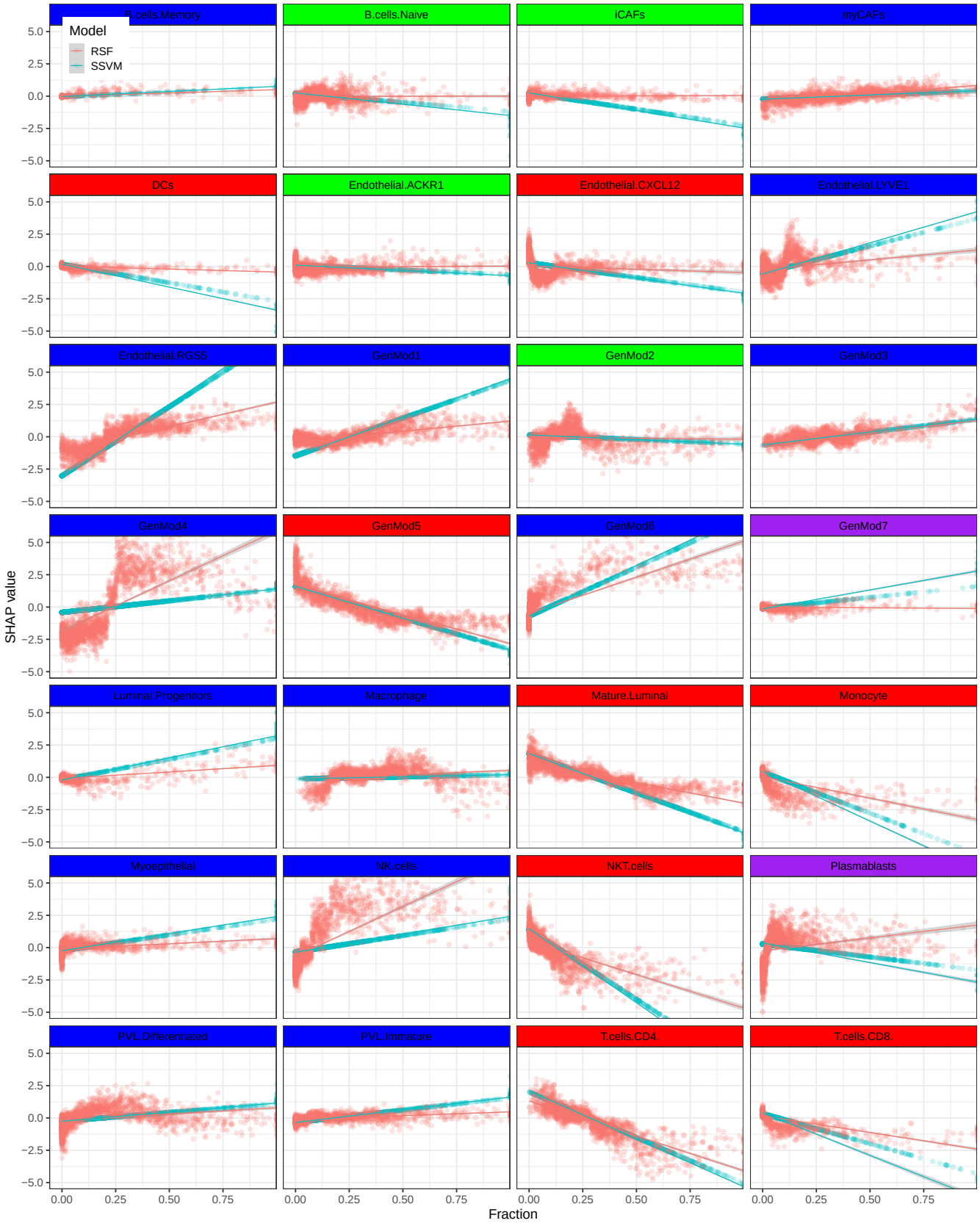


Figure S5: SHAP values for SSVM and RSF vs normalized cell type fractions, following the same color scheme introduced in Fig. 1. Some cell types are filtered out (green and purple) and for the remaining cells we calculate Survival scores. Source data are provided as a Source Data file.

4. Dataset specific bias

TCGA and MBRC are two major cohorts in breast cancer studies, and they differ in several ways such as sampling, population composition, and follow-up protocols. In addition, TCGA data comes from RNA sequencing, while MBRC data is generated using microarrays, so some differences are expected. Moreover, the characteristics of the data vary significantly: TCGA has a censoring rate of 90% and MBRC has a censoring rate of 55%, based on data from cBioPortal.

These variations, combined with our aspiration to provide measures that are reliable and work across different data sets. Thus, as final filtering, we only keep features that show consistent scores. We believe that this filter is reliable and rigorous as it does compare the internal structures of the models and cross validate the associations captured by models for the two datasets.

F. Survival Scores in breast cancer subtypes

To further quantify the influence of cell type fractions on the predicted survival outcome, we performed a linear regression analysis between the computed SHAP values and the corresponding cell type fractions. By evaluating the slope of the regression line along with its confidence interval, we established a tangible measure of the association between a cell type's fraction and its impact on the prediction.

ER and PAM50 subtypes were analyzed using the same pipeline to calculate Survival Scores. Fig. S6 shows the results for all subtypes in TCGA and MBRC separately.

1. Survival Scores in ER subtypes

Cell types with positive Survival Scores in ER+: Among immune cells, NKT cells, CD8 T cells, and DCs have positive Survival Scores. In cancer cells, GenMod5 and GenMod2 show positive Survival Scores. From normal epithelial cells, ACKR1 endothelials and cancer-associated fibroblasts with inflammatory features (iCAFs) show positive Survival Scores.

Cell types with negative Survival Scores in ER+: RGS5 endothelials and differentiated PVLs from microvascular cells have negative Survival Scores. Among immune cells, NK cells and macrophages show negative Survival Scores. GenMod3 from cancer cells, myCAFs from stromal cells, and luminal progenitors from normal epithelial cells also show negative Survival Scores.

Cell types with positive Survival Scores in ER-: NKT cells, plasmablasts, CD4 T cells, and naive B cells, all immune cells, show positive Survival Scores.

Cell types with negative Survival Scores in ER-: GenMod6 and luminal progenitors show negative Survival Scores.

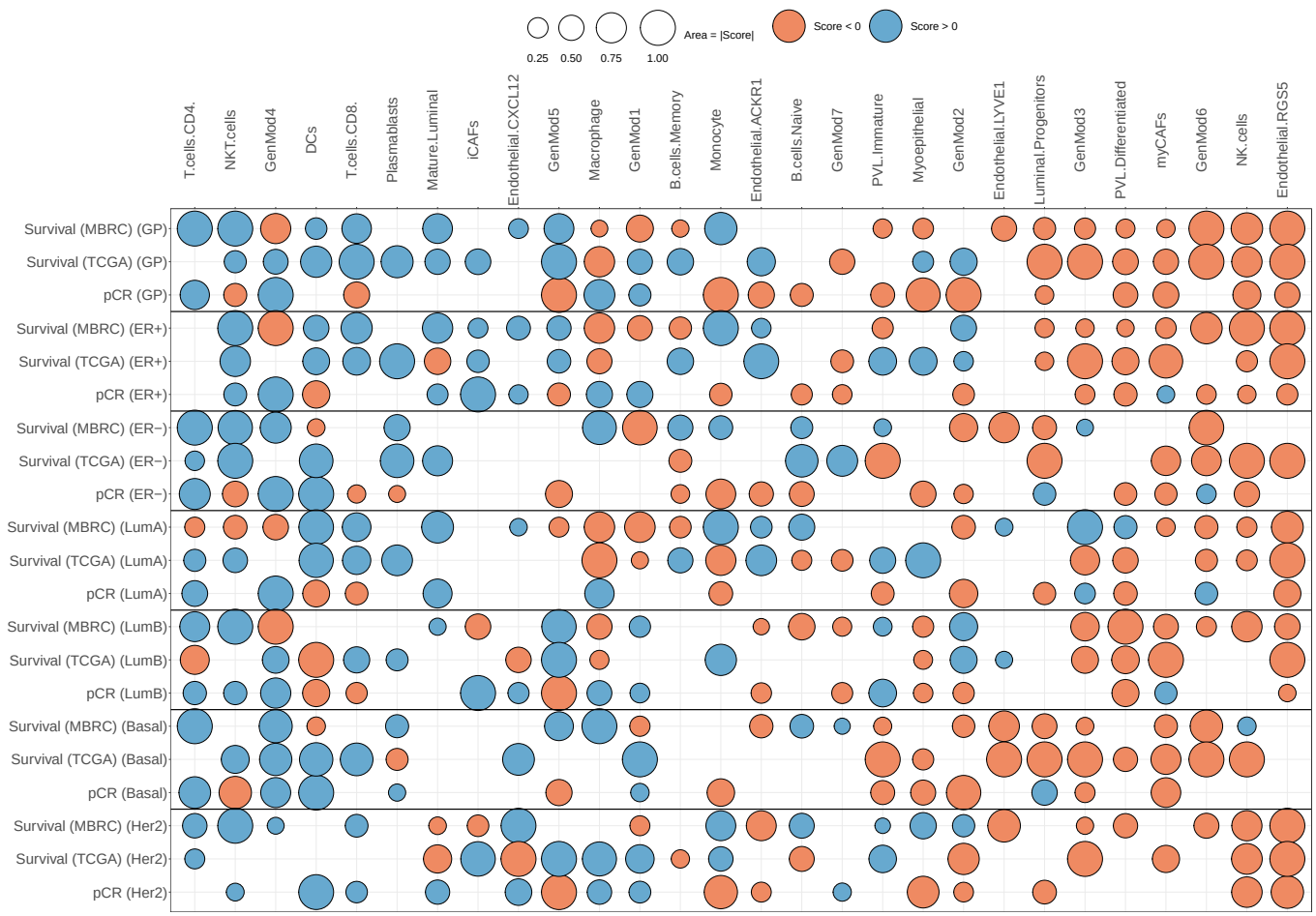


Figure S6: Survival Scores for subtypes in MBRC and TCGA datasets, alongside pCR scores. Source data are provided as a Source Data file

.

2. Survival Scores in PAM50 subtypes

CD8 T cells, DCs, and ACKR1 endothelials show positive Survival Scores in LumA tumors. GenMod1, macrophages, GenMod6, NK cells, and RGS5 endothelials show negative Survival Scores in LumA. GenMod5 and GenMod2 show positive Survival Scores in LumB tumors. Myoepithelials, macrophages, differentiated PVLs, GenMod3, myCAFs, and RGS5 endothelials show negative Survival Scores in LumB. GenMod4 shows positive Survival Scores in Basal. Luminal progenitors, LYVE1 endothelials, GenMod3, GenMod6, and myCAFs show negative Survival Scores in Basal. Monocytes, CD4 T cells, and immature PVLs show positive Survival Scores in HER2. Mature luminals, NK cells, and RGS5 endothelials show negative Survival Scores in HER2.

G. Alignment of pCR and Survival

Generally, in the breast cancer population, ER+ tumors have a lower pCR rate but better survival compared to ER- tumors. This well-established fact in breast tumors highlights the importance of evaluating the relationship between pCR and survival. To quantify this alignment, we developed an intuitive method to explore the alignment between pCR and survival from a TME perspective. We defined and calculated an "Alignment Score" as $\sum_{i=1}^{28} \text{pCR}_i \cdot \text{Survival}_i$, where pCR_i and Survival_i are the corresponding scores for cell type i.

We calculated the Alignment Score separately for TCGA and MBRC datasets, as shown in Fig. S7. In the GP, as mentioned earlier, pCR does not imply better survival. This is evident in the Alignment Score of GP for both datasets. Within ER subtypes, ER- shows positive alignment, suggesting a positive association between these two

clinical outcomes. For ER+, the results are inconsistent. Interestingly, PAM50 subtypes show a clear positive or negative alignment. Both LumA and LumB show negative Alignment Scores across both TCGA and MBRC. HER2 and Basal subtypes show positive Alignment Scores, following the ER- results.

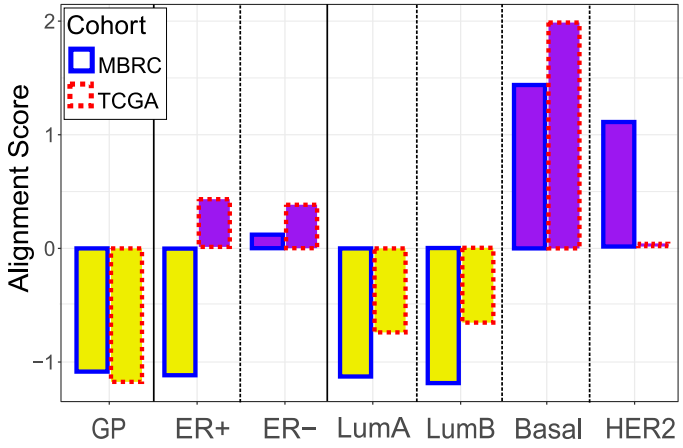
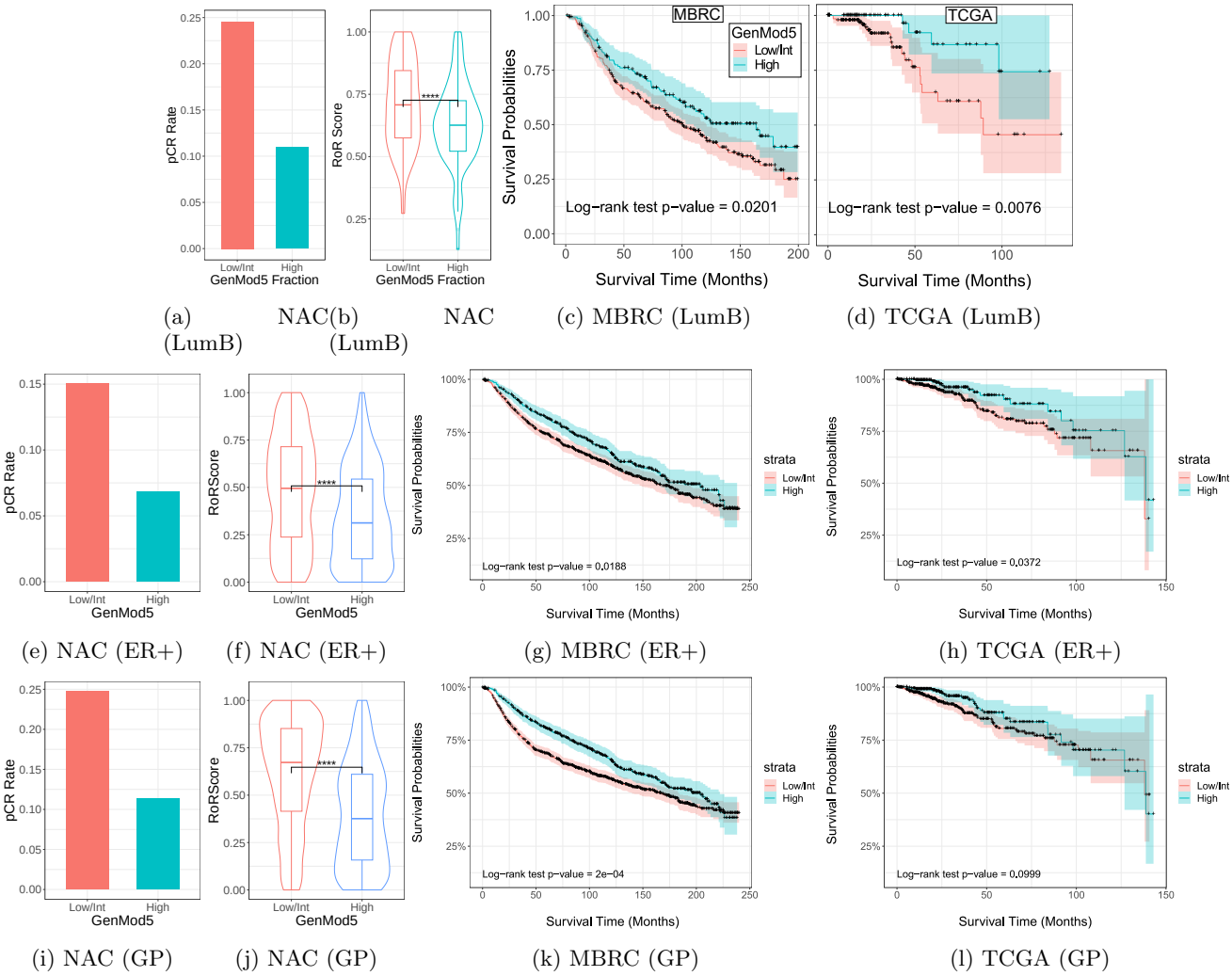


Figure S7: Alignment Score calculated for all groups. Source data are provided as a Source Data file.

H. GenMod5 fraction stratifies breast tumors into groups with distinct clinical outcomes

A possible explanation for the negative Alignment Score in LumB tumors is that there might be different subtypes within this group with unique pCR rates and survival outcomes. Looking back at Fig. **??** in the main text, we explored if the fraction of GenMod5 can stratify LumB samples. Therefore, we divided LumB samples into three groups with equal numbers of samples based on the density of GenMod5 (Low, Int, and High). We first selected the NAC cohort, which includes 464 samples. Divided into three groups, we separated the group with a higher fraction of GenMod5 and compared their pCR rate with the other two groups, which were combined into a single group "Low/Int". As expected, the pCR rate is less than 45% of that of the "Low/Int" group, as shown in Fig. S8(a). Interestingly, the risk of recurrence (RoR) score [30] for the same group is significantly lower (refer to Fig. S8(b)), suggesting the existence of a subset of LumB tumors with a low pCR rate but better survival. To further explore this observation, we divided LumB samples in the MBRC and TCGA datasets into two groups based on the fraction of GenMod5 and compared their survival using a Kaplan-Meier graph. In both datasets, their survival is significantly different based on the p-value of the log-rank test, as shown in Fig. S8(c) and (d). Similarly, the GenMod5 fraction can be used to divide ER+ or GP samples into two groups with distinct behaviors, as shown in the rest of Fig. S8.



(a) NAC (LumB) (b) NAC (LumB) (c) MBRC (LumB) (d) TCGA (LumB)

(e) NAC (ER+) (f) NAC (ER+) (g) MBRC (ER+) (h) TCGA (ER+)

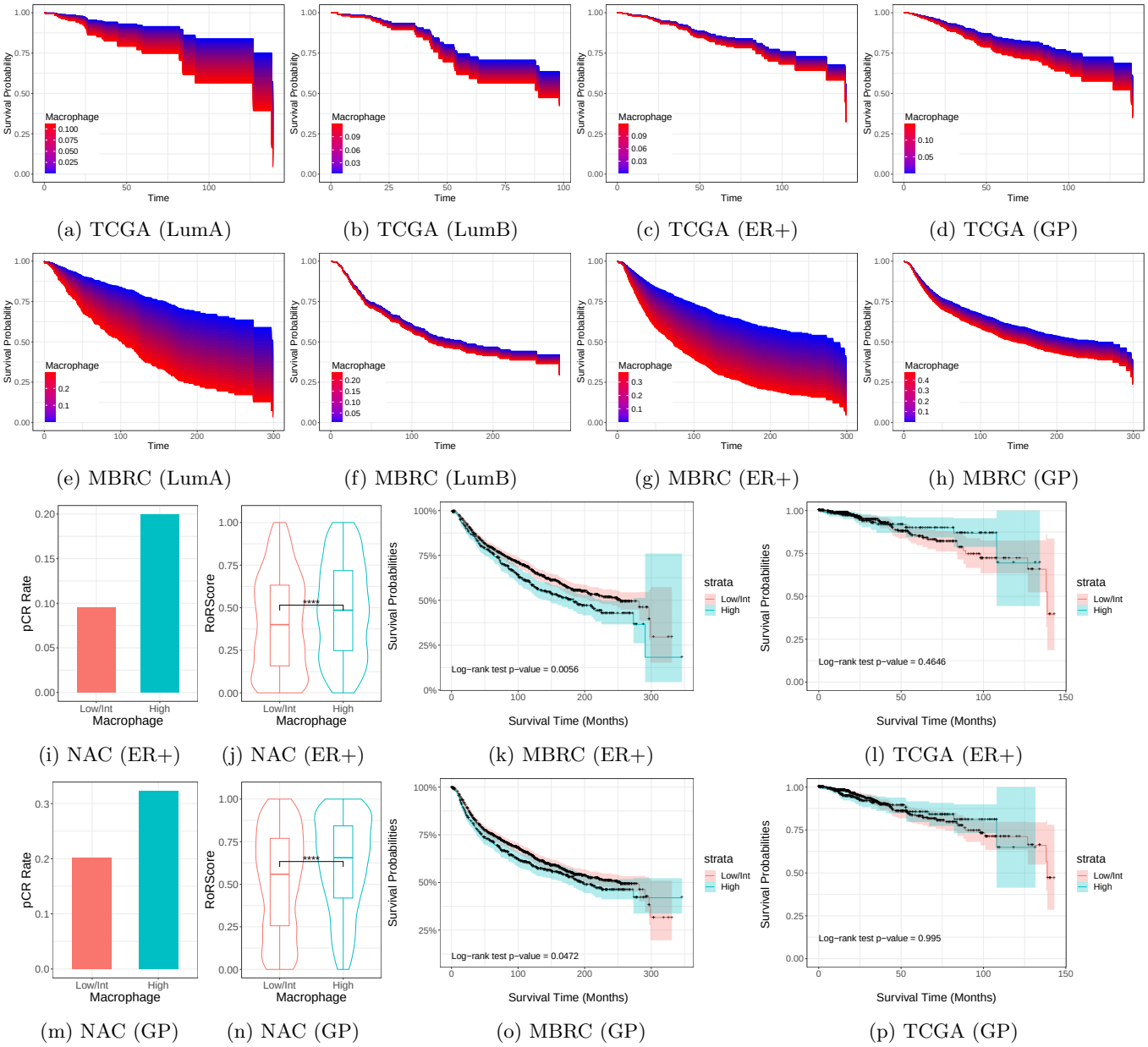
(i) NAC (GP) (j) NAC (GP) (k) MBRC (GP) (l) TCGA (GP)

Figure S8: GenMod5 fraction stratifies breast tumors into groups with distinct clinical outcomes in LumB (a, b, c, d), ER+ (e, f, g, h), and GP (i, j, k. l) tumors. Box plots show the median and 25th–75th percentiles and hiskers extend to the most extreme values within 1.5×interquartile range in b, f and j. **** represents p values of 2.29e-06, 1.1e-14 and 1.73e-51 for two-sided Student's t-test in b, f and j respectively. Source data are provided as a Source Data file.

I. Macrophages fraction stratifies breast cancer samples into groups with distinct clinical outcomes

Following the observation in the main text, we also divided ER+ and GP samples into two groups based on the fraction of macrophages and explored their characteristics.

Deconvolution: We first plot association of macrophage fractions and survival probability in deconvolution data. In both TCGA and MBRC, across LumA, LumB, ER+ and GP, higher levels of macrophages are associated with lower survival probabilities as Figs S9 a-h show. Additionally, for the NAC cohort, we compared pCR rates and RoR scores for these groups of samples across ER+ and GP tumors. Following the observation for LumA samples, samples with high macrophage fraction show higher pCR rates while also having significantly higher RoR scores, as shown in Fig. S9 (e), (f), (i), and (j). In MBRC, samples with a high fraction of macrophages show significantly lower survival probabilities, as shown in Fig. S9 (g) and (k). In TCGA, however, such significance was not observed. The lack of such observation does not undermine the role of macrophages in the survival probability of the TCGA dataset, as we consistently recovered negative Survival Scores in LumA, LumB, ER+, and GP in this dataset.



(a) TCGA (LumA) (b) TCGA (LumB) (c) TCGA (ER+) (d) TCGA (GP)

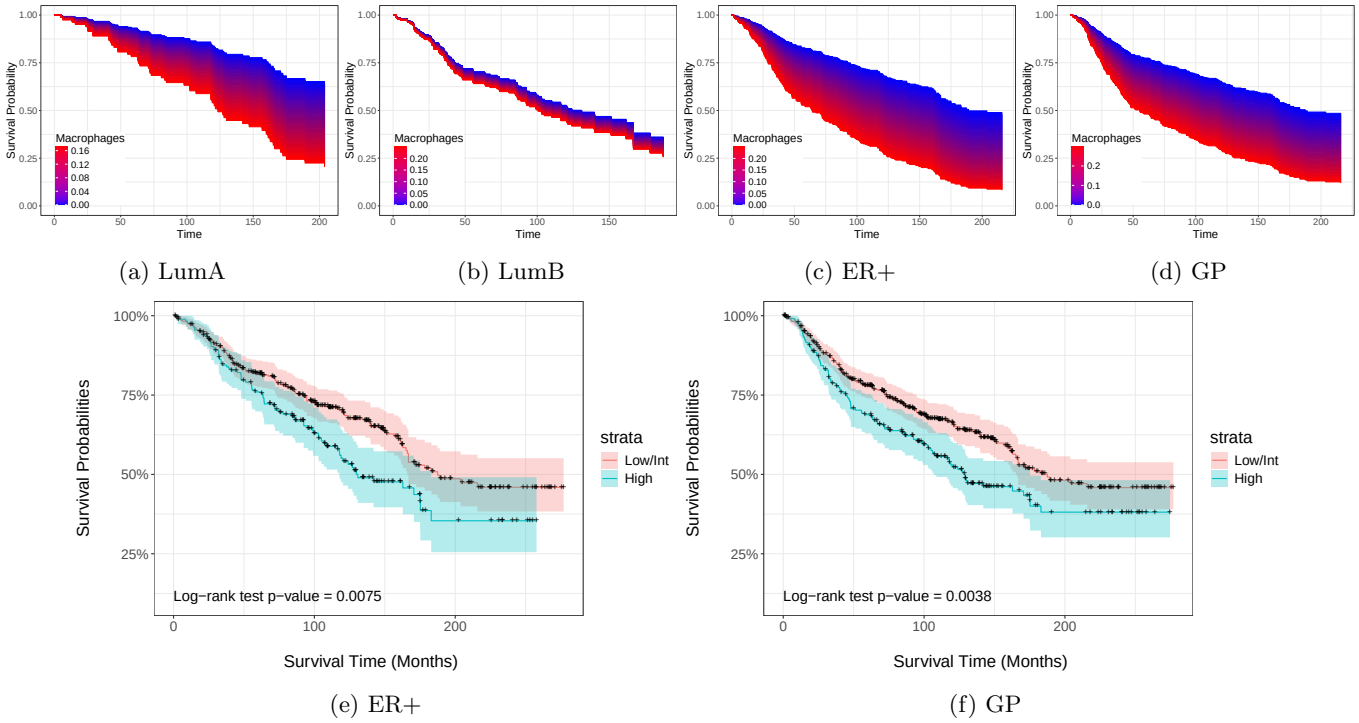
(e) MBRC (LumA) (f) MBRC (LumB) (g) MBRC (ER+) (h) MBRC (GP)

(i) NAC (ER+) (j) NAC (ER+) (k) MBRC (ER+) (l) TCGA (ER+)

(m) NAC (GP) (n) NAC (GP) (o) MBRC (GP) (p) TCGA (GP)

Figure S9: **Macrophage fraction inferred from bulk deconvolution stratifies breast cancer samples into groups with distinct clinical outcomes.** Association between macrophage fraction and survival probability in TCGA (a, b, c, d) and MBRC (e, f, g, h) datasets across LumA, LumB, ER+, and GP subtypes. Comparison of pCR rates and RoR scores and survival probabilities between macrophage-high and macrophage-low groups in NAC (i, j, m and n), MBRC (k and o), and TCGA (l and p) cohorts. Higher macrophage fractions correspond to worse overall survival and, in the NAC cohort, higher pCR rates and RoR scores. Source data are provided as a Source Data file. Box plots show the median and 25th–75th percentiles and hiskers extend to the most extreme values within 1.5×interquartile range in j and n. **** represents p values of 1.7 e-51 and 1.36e-11 for two-sided Student's t-test in j and n respectively.

IMC Data: Similar to deconvolution data, we first plot association of macrophage fractions and survival probability. Across LumA, LumB, ER+ and GP, higher levels of macrophages are associated with lower survival probabilities (S10 a-d). Then, we used the fraction of macrophages obtained from IMC data [31] to divide ER+ and GP samples into two groups following the same method. As Figs. S10 (e) and (f) show, in both cases, samples with a high fraction of macrophages show significantly lower survival probabilities.



(a) LumA (b) LumB (c) ER+ (d) GP

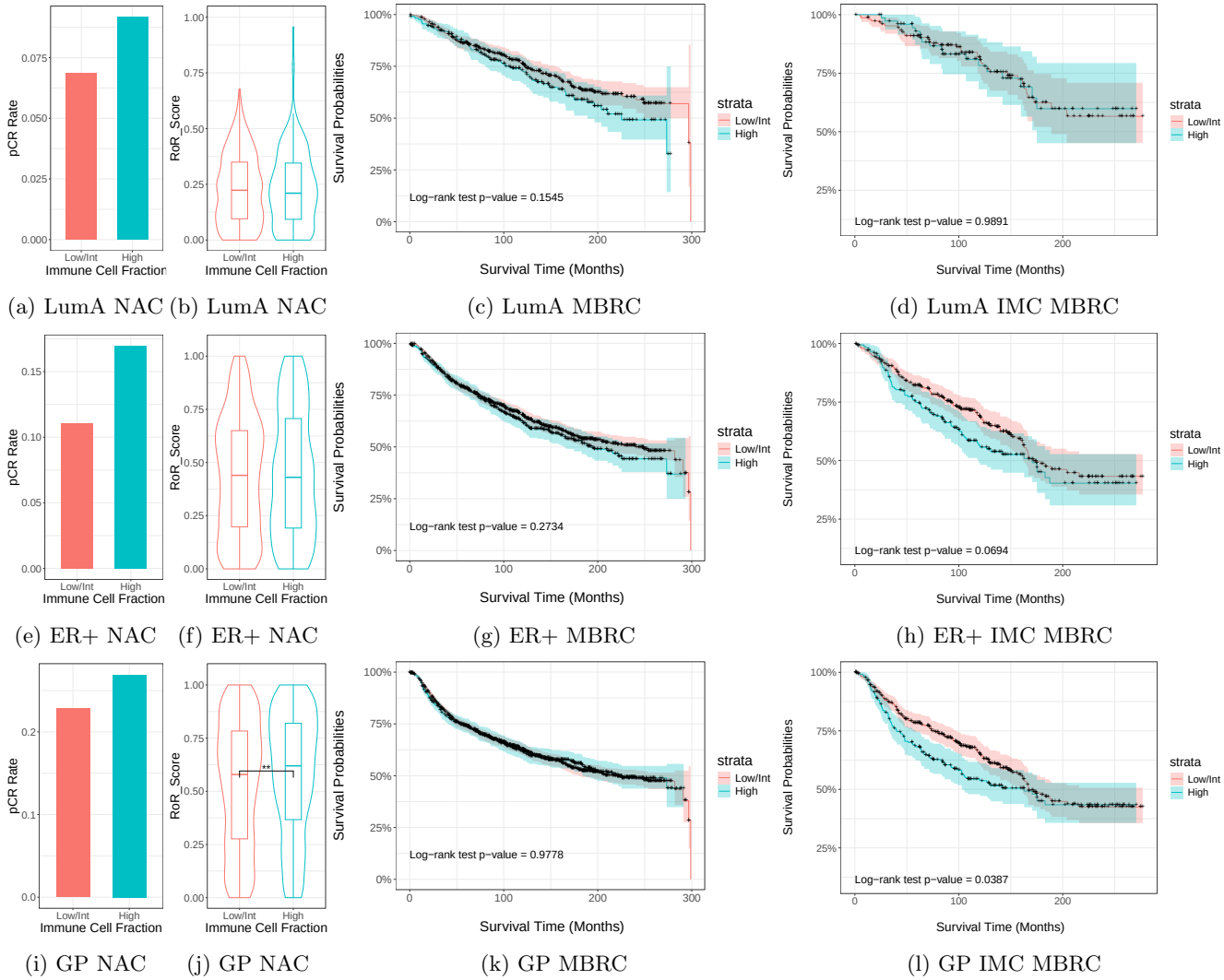
(e) ER+ (f) GP

Figure S10: **Macrophage fraction estimated from IMC data stratifies breast cancer samples into groups with distinct clinical outcomes.** (a–d) Association between IMC-derived macrophage fraction and survival probability across LumA, LumB, ER+, and GP subtypes. (e–f) Kaplan–Meier survival curves for ER+ and GP tumors, showing that macrophage-high samples exhibit significantly lower survival probabilities. Source data are provided as a Source Data file.

J. Immune infiltration and clinical outcomes

While we showed that Macrophage infiltration levels leads to distinct clinical outcomes, here we explore how immune infiltration, defined by fraction of all immune cells affects clinical outcomes. In deconvolution, immune cells include: DCs, B.cells.Memory, B.cells.Naive, T.cells.CD4., T.cells.CD8., Monocyte, NKT.cells, NK.cells and Plasmablasts. We define the summation of all these cell fractions as immune cells fraction and use that to stratify LumA, ER+ and GP samples. In NAC cohort, we compared pCR rates and RoR scores and in MBRC data we compare their survival probabilities. As Fig. S11 shows, LumA and ER+ samples dont show significant difference. In GP, difference shows up in RoR score.

In IMC data, we have Macrophages, B cells, $CD38^+$ lymphocytes, Macrophages & granulocytes, T_{Reg}&T_{Ex}, $CD4^+$ T cells & APCs, $CD38^+$ lymphocytes, $CD4^+$, T cells, $CD8^+$ T cells and Granulocytes. Similarly, we put the summation of these cell fractions as Immune Cells Fraction and use it to stratify the samples in LumA, ER+ and GP. As the Fog. S11 shows, immune cells fraction does not stratify the patients in LumA and ER+ samples.



(a) LumA NAC (b) LumA NAC (c) LumA MBRC (d) LumA IMC MBRC

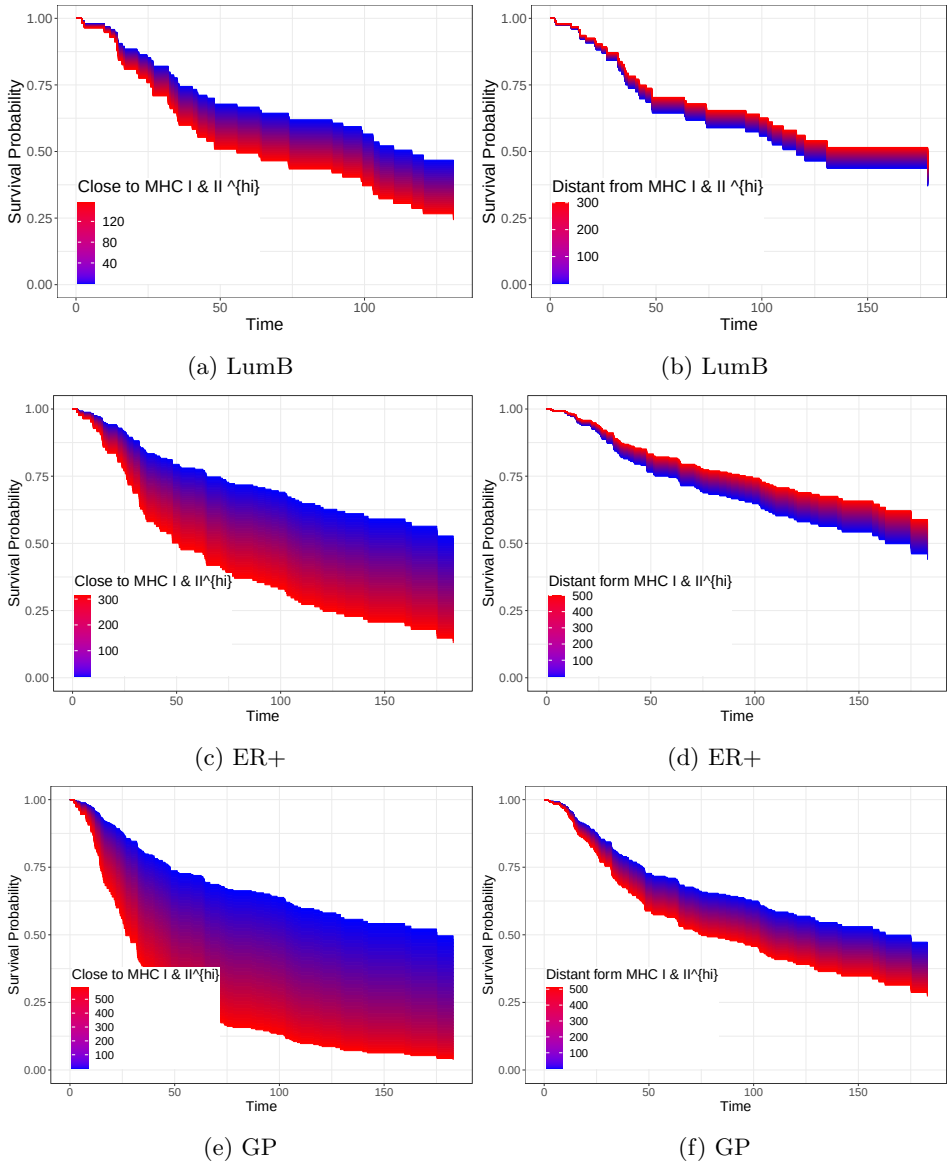
(e) ER+ NAC (f) ER+ NAC (g) ER+ MBRC (h) ER+ IMC MBRC

(i) GP NAC (j) GP NAC (k) GP MBRC (l) GP IMC MBRC

Figure S11: Immune cells fraction used to stratify LumA, ER+ and GP samples in NAC (a, b, e, f, i and j), MBRC (c, g and k) and IMC MBRC (d, h and l) data. In both LumA and ER+, immune cell fraction levels don't lead to different clinical outcomes. Source data are provided as a Source Data file. Box plots show the median and 25th–75th percentiles and hiskers extend to the most extreme values within 1.5×interquartile range in b, f and j. ** represents p value 0.00195 for two-sided Student's t-test in j.

K. Distance from MHC I & II^{hi} ECs reverses role of macrophages in RFS

In LumB and ER+ samples, we divided macrophages into two groups based on their minimum distance to MHC I & II^{hi} ECs. The threshold distance, r_T, is the median of the minimum distances, r_m, between macrophages and the target cell type across LumB and ER+ samples, respectively. We then fit a Cox proportional hazards model [32] to each subset of macrophages. As the survival area plot [33] in Fig. S12 shows, for both LumB and ER+ samples, a higher frequency of macrophages close to (distant from) MHC I & II^{hi} leads to lower (higher) survival probability.



(a) LumB (b) LumB

(c) ER+ (d) ER+

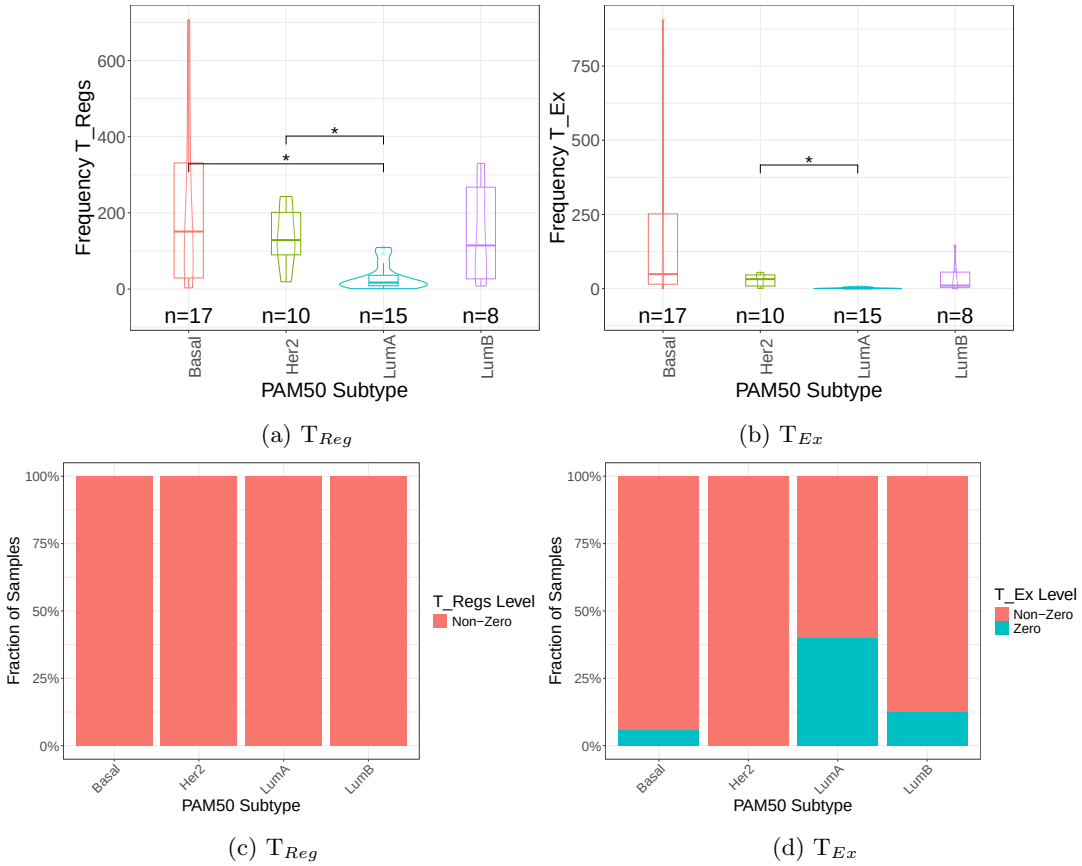
(e) GP (f) GP

Figure S12: Survival area plot for the effect of the frequency of macrophages close to MHC I & II^{hi} and distant from MHC I & II^{hi} in LumB (a and b), ER+ (c and d), and GP (e and f) samples. Source data are provided as a Source Data file.

L. T_{Reg} and T_{Ex} in tumor subtypes.

1. Frequency

scRNA-seq Analysis In the scRNA-seq dataset, we analyzed the frequency of T_{Reg} and T_{Ex} cells separately across the PAM50 subtypes (shown in Fig. S13). Our results indicate that while the overall frequency of these cells is significantly lower in LumA compared to other subtypes, all LumA samples contained Tregs, and the majority also included Tex cells. This suggests that while Tex and Tregs may be less abundant in LumA tumors, they are still present in most cases.

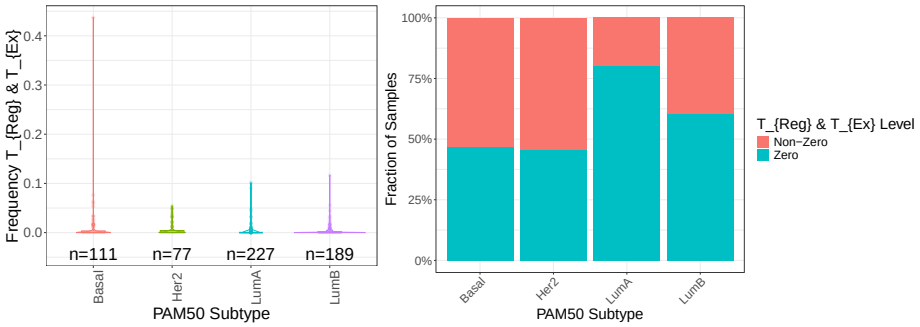


(a) T_{Reg} (b) T_{Ex}

(c) T_{Reg} (d) T_{Ex}

Figure S13: Frequency of T_{Reg} (a) and T_{Ex} (b) across PAM50 subtypes. Fraction of samples including T_{Reg} (c) and T_{Ex} (d) across PAM50 subtypes. Box plots show the median and 25th–75th percentiles and hiskers extend to the most extreme values within 1.5×interquartile range in a and b. * represents benferroni adjusted p values 0.05 for Basal vs LumA and 0.03 for Her2 vs LumA in a and 0.03 in b for two-sided Student's t-test. Source data are provided as a Source Data file.

IMC Analysis In the IMC dataset, we examined the presence of T_{Reg}&T_{Ex} across PAM50 subtypes. We observed that approximately 25% of LumA samples contained T_{Reg}&T_{Ex} (see S14). This proportion is lower than in Basal and Her2 subtypes, where these cells were detected in about 50% of the samples. Given the spatial nature of IMC data, the lower detection rate in LumA may be influenced by technical limitations, such as lower sensitivity for rare cell types in highly heterogeneous tumor environments.



(a) (b)

Figure S14: a) Frequency of T_{Reg}&T_{Ex} across PAM50 subtypes. b) Fraction of samples including T_{Reg}&T_{Ex} across PAM50 subtypes. Box plots show the median and 25th–75th percentiles and hiskers extend to the most extreme values within 1.5×interquartile range in a. Source data are provided as a Source Data file.

2. DEA and PEA

We run DEA and PEA across Luminal (LumA + LumB) vs Basal/Her2 to find differentially expressed genes and to reveal pathway-level differences in T_{Reg} and T_{Ex} as shown in Fig. S15.

For T_{Ex} cells (Figs. S15a and b), luminal tumours showed transcriptional signatures consistent with deeper exhaustion. Up-regulated pathways included TNFα/NF-κB, TGF-β signalling, mTORC1, and IL-2/STAT5, alongside increased expression of co-stimulatory molecules (CD81), proteasomal components (UBE2M/UBE2S), and extracellular matrix–associated genes (COL1A2, TNFSF9). Conversely, interferon responses, antigen-presentation genes, and certain TCR variable segments were down-regulated. Together, this pattern suggests that luminal T_{Ex} adopt a metabolically active but functionally impaired state, characterised by reduced effector capacity and enhanced stress adaptation.

For T_{Reg} cells (Figs. S15c and d), luminal tumours displayed marked up-regulation of pathways associated with stress adaptation and suppressive function, including TNFα via NF-κB, IL-2/STAT5 signalling, hypoxia, and the unfolded protein response. Notably, heat-shock proteins (e.g., HSPA1A/B, HSP90AA1, DNAJB1) and metabolic enzymes (e.g., GLS, UBE2F/UBE2S) were over-expressed, suggesting adaptation to a nutrient- and oxygen-deprived microenvironment that may enhance suppressive potency. In contrast, pathways linked to interferon-α/γ responses and antigen presentation were down-regulated, consistent with a shift away from classical immune activation toward a suppressive phenotype. These findings suggest that luminal T_{Reg} exert qualitatively stronger suppressive activity despite being numerically less abundant.

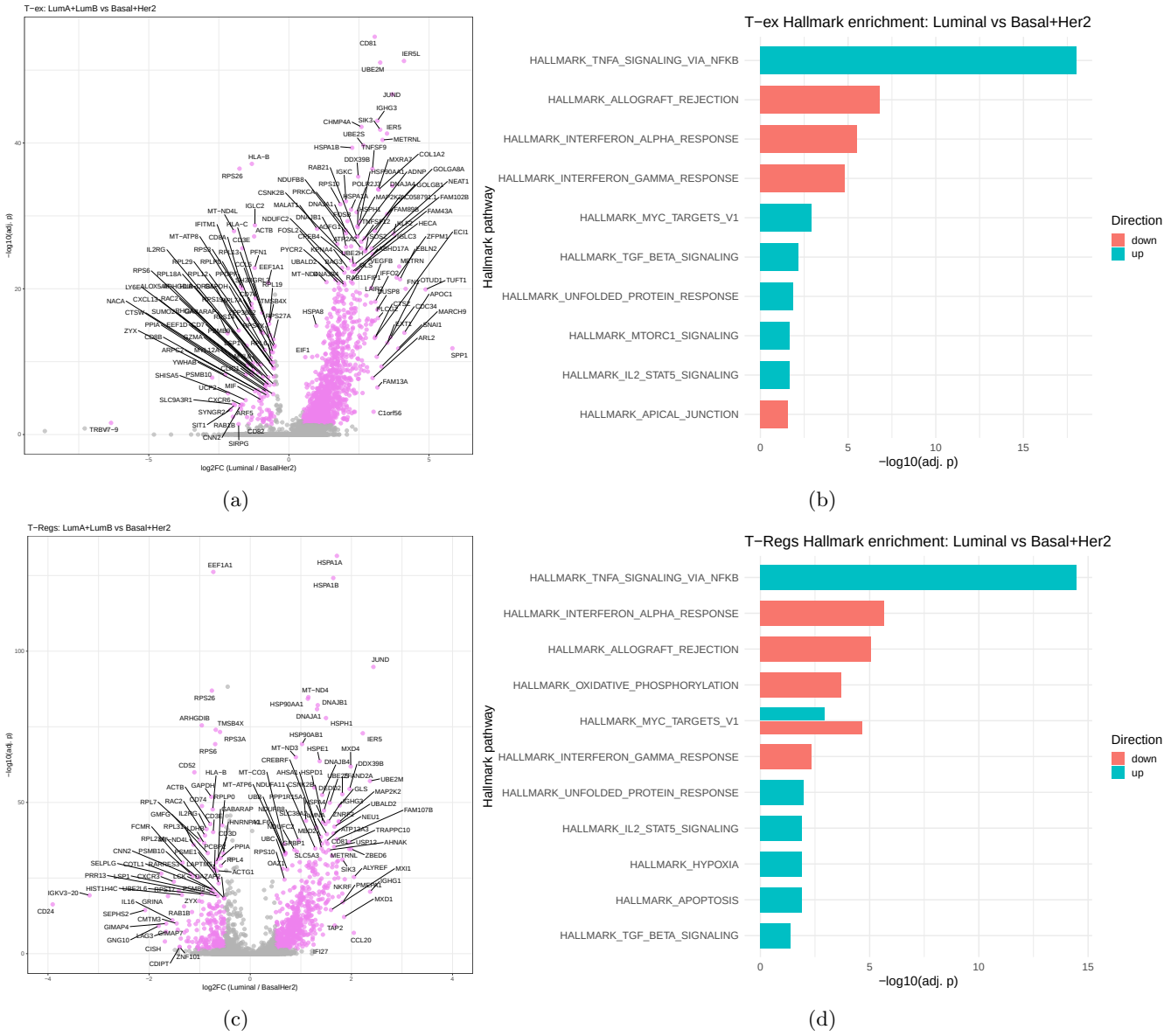
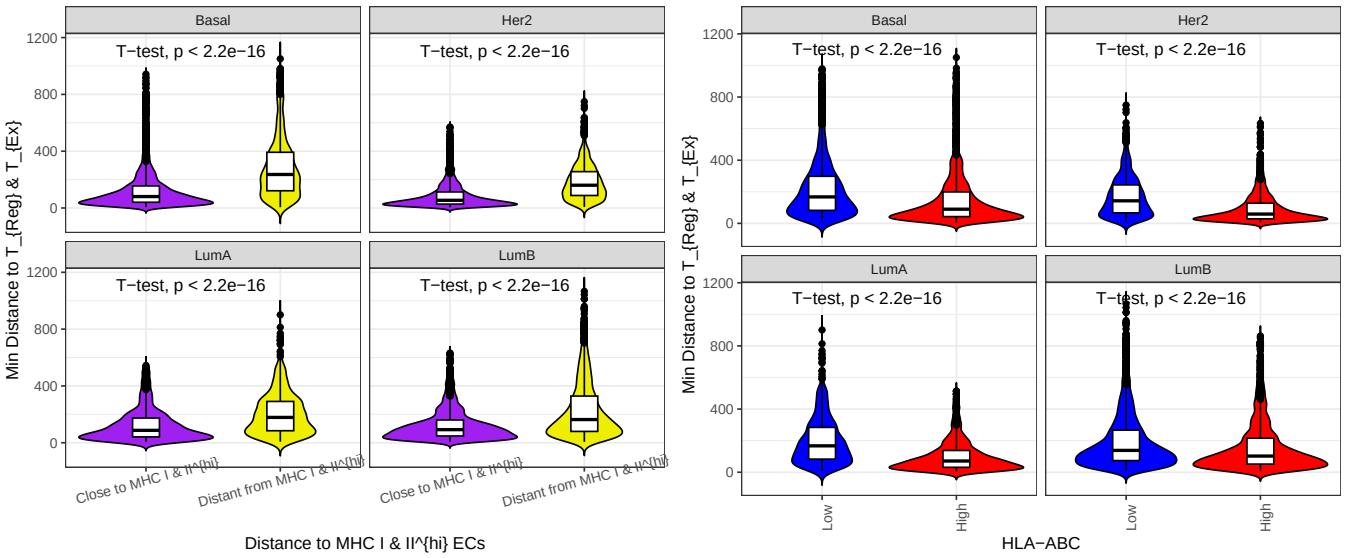


Figure S15: Differential expression (volcano plots) and pathway enrichment (barplots) of T_{Ex} (panels a–b) and T_{Reg} (panels c–d) in luminal vs. basal/HER2 tumours. Source data are provided as a Source Data file.

M. Macrophages and T_{Reg}&T_{Ex} across subtypes

1. *Distance between Macrophages and T_{Reg}&T_{Ex}*

Next, we compare the distance of subsets of macrophages to T_{Reg}&T_{Ex} withing PAM50 subtypes. First, divide macrophages into two groups based on their minimum distance from MHC I & II^{hi} and then compared the their distance to T_{Reg}&T_{Ex} (see Fig. S16 (a). Then, we divide macrophages into two groups with high or low levels of HLA-ABC and compared their minimum distance from T_{Reg}&T_{Ex}, as shown in Fig. S16 (b). Both these analysis suggests that colocalization of close to MHC I & IIhi ECs and HLA-ABC^{hi} macrophages with T_{Reg}&T_{Ex} is not limited to an specific subset of breast tumors.

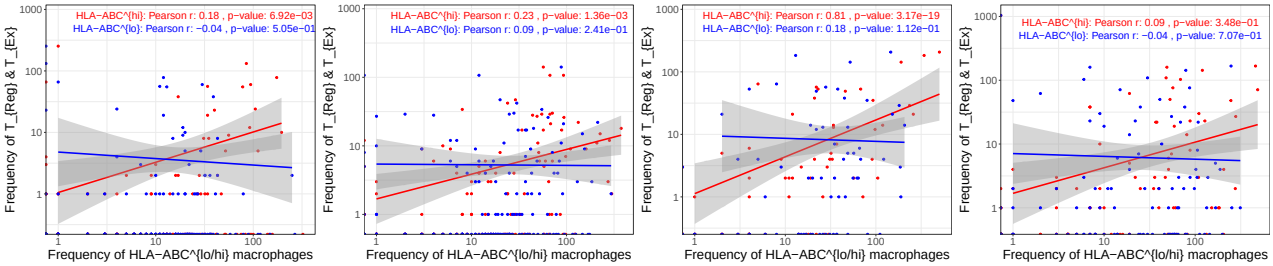


(a) (b)

Figure S16: Comparison of minimum distance of T_{Reg}&T_{Ex} with two subsets of macrophages that are divided based on their distance to MHC I & II^{hi} (a) and HLA-ABC level (b) in PAM50 subtypes. Source data are provided as a Source Data file.

2. *Correlations between Macrophages and T_{Reg}&T_{Ex}*

As a supplementary analysis, and similar to Fig. 5 (d), we compare correlation of HLA-ABC^{hi} macrophages to frequency of T_{Reg}&T_{Ex}. The results, shown in Fig. S17, further confirms that HLA-ABC^{hi} macrophages are exclusively correlated with T_{Reg}&T_{Ex}.



(a) LumA (b) LumB (c) Her2 (d) Basal

Figure S17: Correlation of frequency of HLA-ABC^{hi} and HLA-ABC^{lo} macrophages and T_{Reg}&T_{Ex} in LumA (a), LumB (b), Her2 (c) and Basal (d). Box plots show the median and 25th–75th percentiles and hiskers extend to the most extreme values within 1.5×interquartile range in all plots. Source data are provided as a Source Data file.

3. *Further stratification of T_{Reg}&T_{Ex}*

In Danenberg *et al.* [31], T_{Reg} and T_{Ex} were not analyzed separately. Here, we have now re-examined the data by explicitly separating the two populations. We stratified the combined T-cell compartment by FOXP3 and PD-1 expressions: among T_{Reg}&T_{Ex} cells, top 33% with FOXP3 labeled as T_{Reg}. Among the remaining 67%, cells with PD-1 > median were treated as T_{Ex}. This annotation led to $FOXP3^{hi}$ as T_{Reg}, and $FOXP3^{lo}$ PD-1^{hi} as T_{Ex} ($FOXP3^{lo}$ PD-1^{lo} remain unlabeled). This is an effort to annotate cells with high confidence without complicating the analysis and should be seen as support for our main results. An example of such re-annotation is shown in S18(a). Similar to Fig. 5(b), both T_{Ex} and T_{Reg} are located closer to macrophages near MHC I & II^{hi} epithelial cells, consistent with Fig.

Table S3: Significant correlations between HLA.ABC expression and T-cell populations across conditions.

Condition	r	p	n	Pair	stars	r_{lab}
GP	0.1695	8.67×10^{-6}	681	T_Reg vs. hi	***	0.17
GP	0.2303	1.20×10^{-9}	681	T_Ex vs. hi	***	0.23
ER+	0.2420	1.57×10^{-8}	532	T_Reg vs. hi	***	0.24
ER+	0.2456	9.44×10^{-9}	532	T_Ex vs. hi	***	0.25
ER-	0.2083	0.0108	149	T_Ex vs. hi	*	0.21
LumA	0.2571	8.90×10^{-5}	227	T_Reg vs. hi	***	0.26
LumA	0.2177	9.62×10^{-4}	227	T_Ex vs. hi	***	0.22
LumB	0.1894	0.0090	189	T_Reg vs. hi	**	0.19
LumB	0.2187	0.0025	189	T_Ex vs. hi	**	0.22
Her2	0.7094	5.14×10^{-13}	77	T_Reg vs. hi	***	0.71
Her2	0.7964	4.79×10^{-18}	77	T_Ex vs. hi	***	0.80

5(c). We then re-computed, the correlations between the frequencies of HLA-ABC$^{\text{hi}}$/HLA-ABC$^{\text{lo}}$ macrophages and each T-cell subset separately in all subsets of tumors. As Fig. S18 (d) (and table S3) shows, HLA-ABC$^{\text{hi}}$ macrophages are exclusively correlated with both T_{Ex} and T_{Reg}.

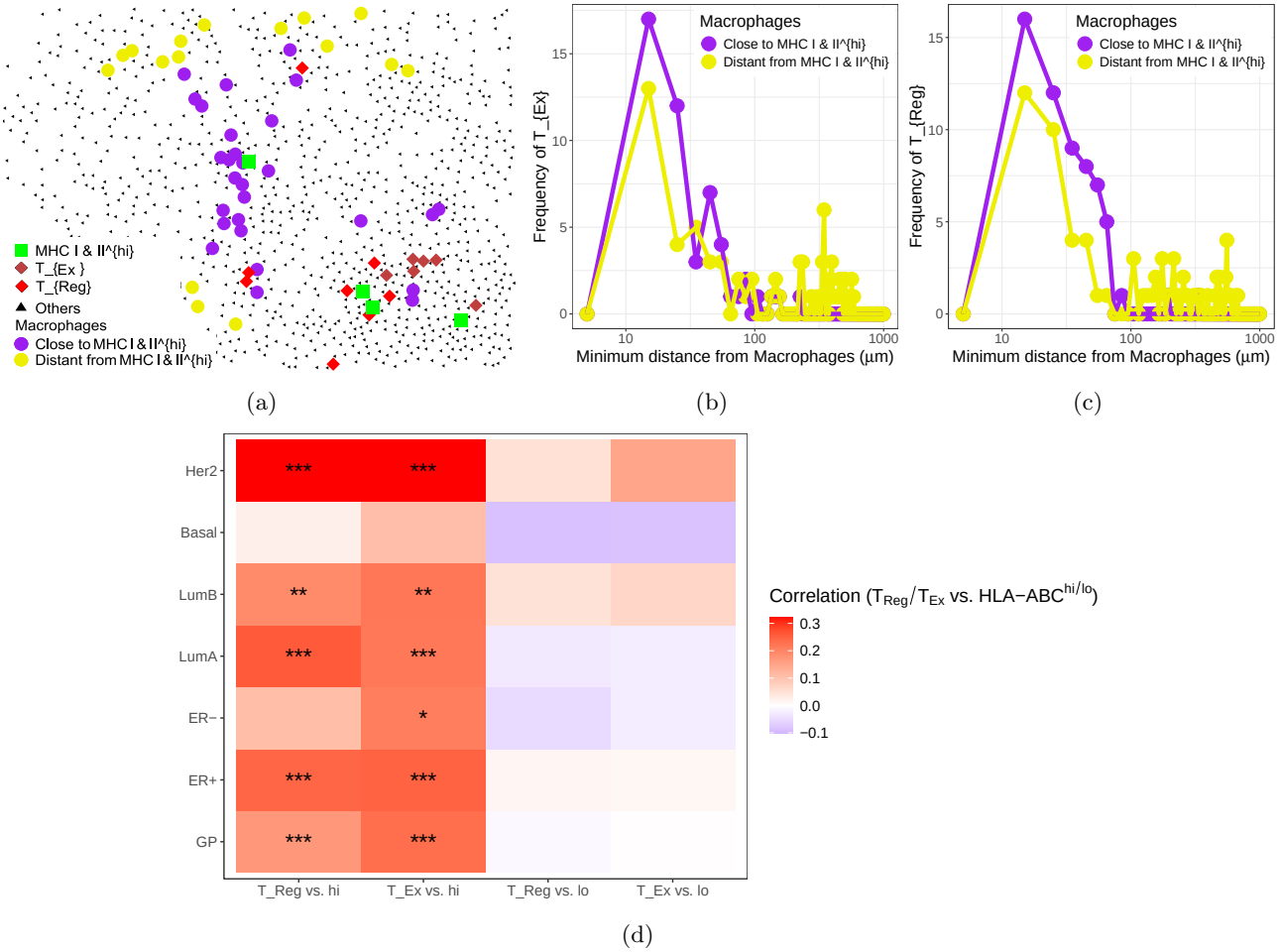


Figure S18: Spatial distribution of T_{Ex} and T_{Reg} cells (a) and their ferquencies in neighborhoods of two groups of macrophages (b-c). d) Correlations between the frequencies of T_{Ex} or T_{Reg} and HLA-ABC$^{\text{hi/lo}}$ macrophages across GP and tumor subtypes. See Table S3 for p-values calculated using two-sided Pearson correlation tests. Source data are provided as a Source Data file.

It should be noted that because cell-state annotation is inherently complex and context-dependent—and FOXP3/PD-

Table S4: Pearson correlations between M2 macrophage HLA-ABC groups and T-cell populations.

	T_{Reg}			T_{Ex}		
	r	p	stars	r	p	stars
M2 Mac HLA-ABC^{hi}	0.328	8.40×10^{-11}	****	0.365	3.87×10^{-13}	****
M2 Mac HLA-ABC^{lo}	0.169	0.00103	**	-0.010	0.842	
M2 Mac	0.437	0	****	0.335	3.21×10^{-11}	****

1 can be transiently expressed—this FOXP3/PD-1-based separation is operational rather than definitive. We applied it to a limited set of cells to obtain an initial view of how these T-cell subsets associate with macrophage phenotypes.

4. M2 Macrophages divided based on HLA-ABC level in TNBC IMC data

In treatment-naive samples from [34], which belong to triple-negative breast cancer (TNBC), we divided (M2) macrophages (the only cell type distinctly annotated as macrophages) into two groups based on their HLA-ABC levels and explored the correlation of each group with T_{Reg}&T_{Ex}. As shown in Fig. S19 (a), only HLA-ABC$^{\text{hi}}$ macrophages exhibit a significant correlation with T_{Ex}. While the correlation of HLA-ABC$^{\text{lo}}$ macrophages with T_{Reg} is significant, it is lower than that of HLA-ABC$^{\text{hi}}$ with T_{Reg}.

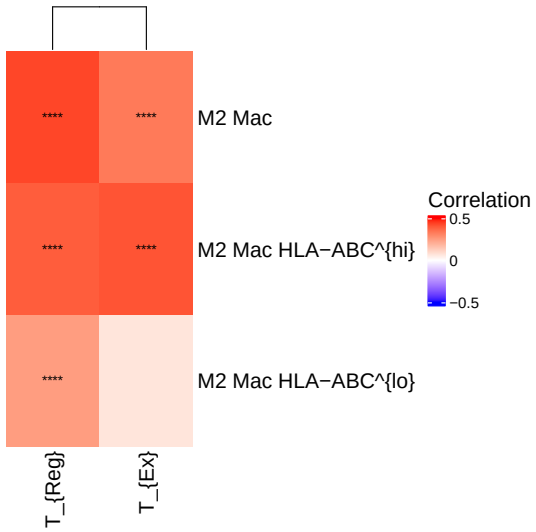


Figure S19: Correlation of macrophages subsets with T_{Reg} and T_{Ex}, see table S4 for corresponding p values from two-sided Pearson correlation tests. Source data are provided as a Source Data file.

5. DEA and PEA for macrophages subtypes

carried out differential expression analysis (DEA) and pathway enrichment analysis (PEA) on macrophage subsets from our single-cell RNA sequencing data. These analyses aimed to determine which macrophage cluster is enriched for the interferon (IFN) response signature observed in HLA-ABC$^{\text{hi}}$ macrophages.

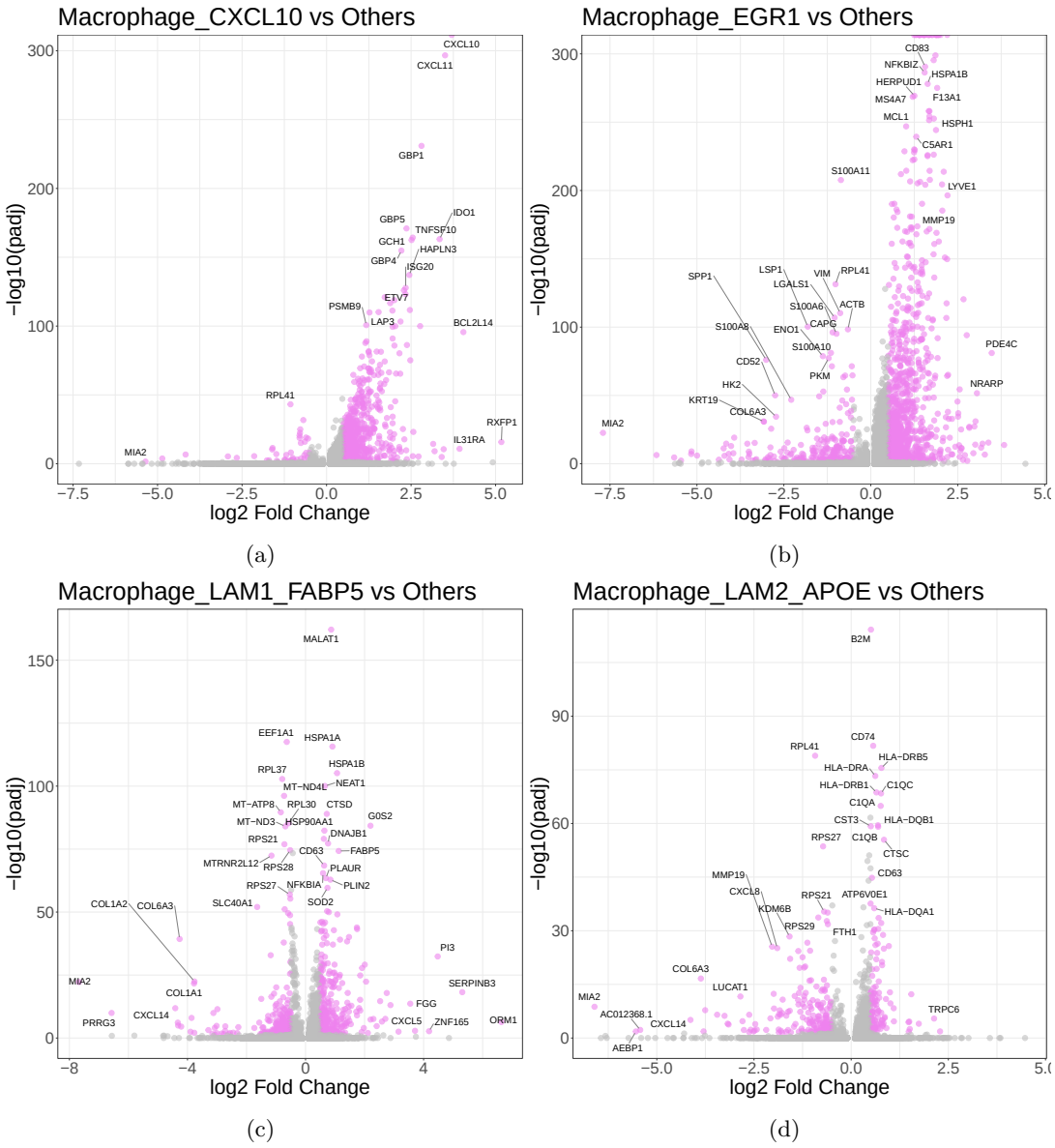


Figure S20: Volcano plots for CXCL10$^+$ (a), EGR1$^+$ (b), FABP5$^+$ (c) and APOE$^+$ (d) subsets of macrophages vs other subsets. Source data are provided as a Source Data file.

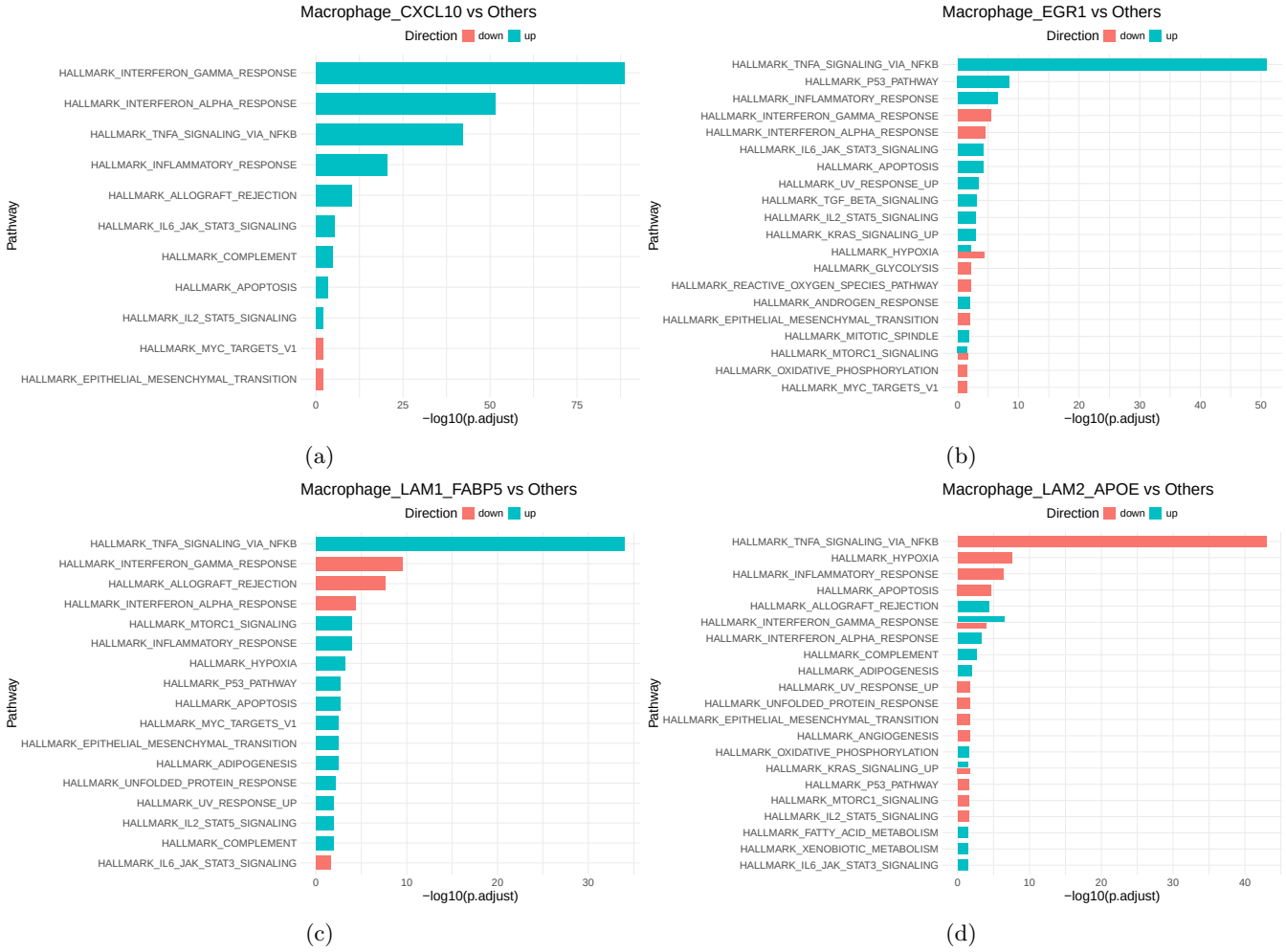


Figure S21: PEA for CXCL10$^+$ (a), EGR1$^+$ (b), FABP5$^+$ (c) and APOE$^+$ (d) subsets of macrophages vs other subsets. Source data are provided as a Source Data file.

N. HLA-ABC^{hi} ECs and immunosuppressive TME in scRNA-seq data

To investigate whether the frequency of HLA-ABC^{hi} ECs is associated with the frequency of TEx and TReg cells [35], we divided the ECs in scRNA-seq data into two groups, HLA-ABC^{lo} and HLA-ABC^{hi} ECs, and calculated the correlation of both groups with these T cells. However, as shown in Fig. S22, there is no significant correlation between the frequencies of EC subsets and the frequencies of TEx and TReg cells.

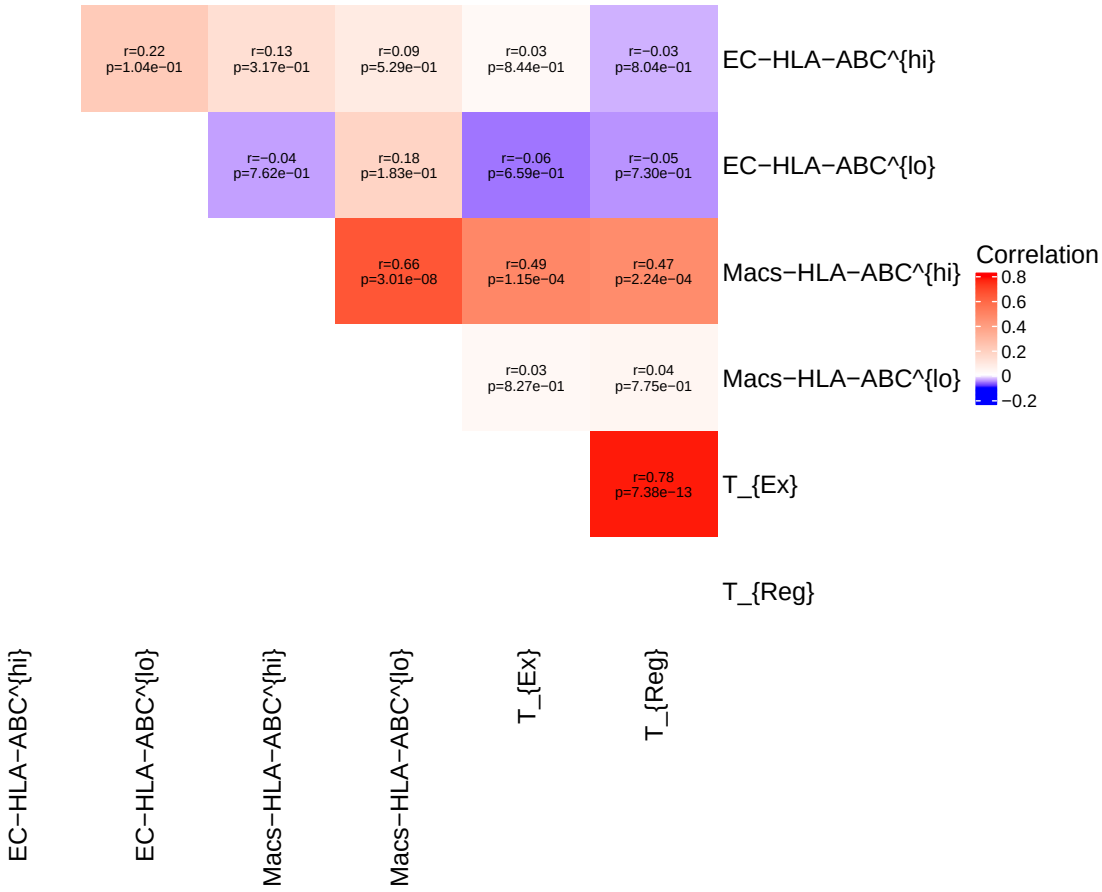


Figure S22: Correlation of HLA-ABC$^{hi/lo}$ ECs and macrophages with TReg and TEx. Corresponding p values from two-sided Pearson correlation tests. Source data are provided as a Source Data file.

O. Cell-cell communication assay

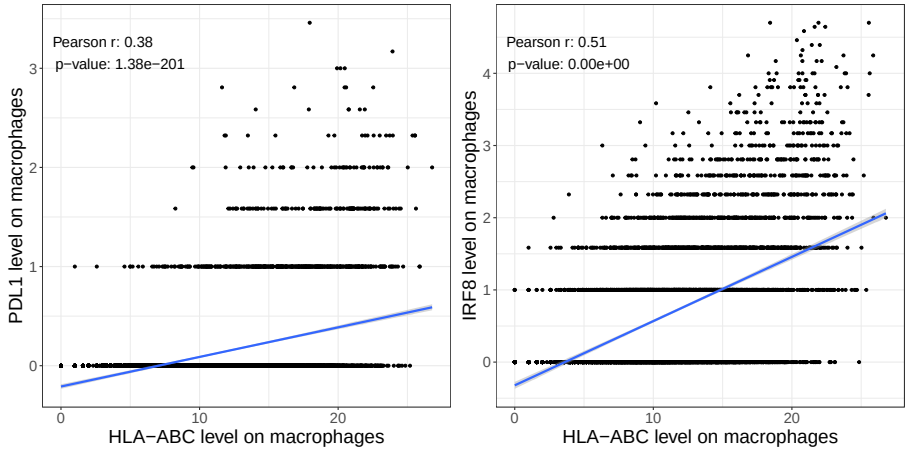
Computational cell-cell communication assays are used to study and predict interactions between different cell types within a tissue or organism. These assays help understand how cells influence each other's behavior through signaling molecules. NicheNet [36] is a method that predicts ligand–target links between interacting cells by combining their expression data with prior knowledge of signaling and gene regulatory networks.

For interaction between TEx and macrophages, TEx were set as sender cells while macrophages were set as receiver cells. To run NicheNet we again followed ligand activity geneset vignette. We selected the top 50 most variant genes on macrophages as target genes of interest in receiver cells.

Figure S23: Interaction of T_{Ex} and macrophages. Source data are provided as a Source Data file.

P. HLA-ABC and Other Genes in Macrophages

We also explored the correlation between the level of HLA-ABC and some other markers that were previously shown to be involved in immunosuppressive macrophages. In particular, we plotted HLA-ABC levels vs. PDL1 and IRF8. As Figs. S24 (a) and (b) show, both of these markers are highly correlated with HLA-ABC.



(a) (b)

Figure S24: a) Correlation of HLA-ABC vs PDL1 on macrophages in scRNA-seq. b) IRF8 vs HLA-ABC on macrophages in scRNA-seq data. Source data are provided as a Source Data file.

To further contextualize the HLA-ABC–associated macrophage state, we next examined genes linked to macrophage

functional polarization and immune modulation. Recent work has highlighted the balance between CXCL9 and SPP1 expression as a key axis distinguishing immunologically active versus tissue-remodeling macrophage programs [37]. We assessed the relationship between HLA-ABC expression and CXCL9, SPP1, and their ratio as shown in Fig. S25.

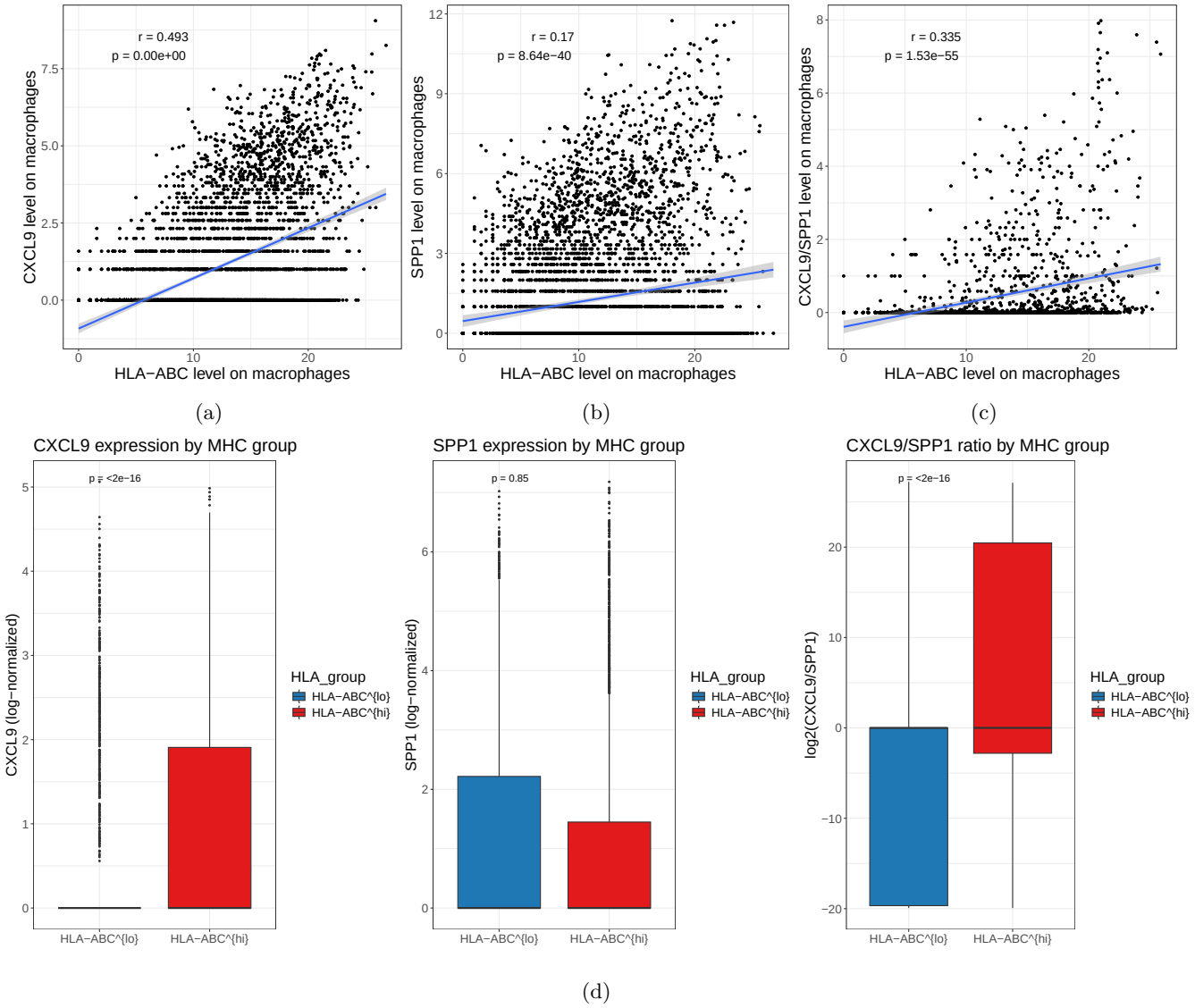


Figure S25: Correlation of HLA-ABC vs CXCL9 (a), SPP1 (b) and CXCL9/SPP1 (c) on macrophages in scRNA-seq. c) CXCL9, SPP1 and CXCL9/SPP1 values across macrophage subsets. Box plots show the median and 25th–75th percentiles in d. Source data are provided as a Source Data file.

Together, these analyses demonstrate that HLA-ABC^{hi} macrophages are characterized by elevated CXCL9 expression and a significantly higher CXCL9/SPP1 ratio.

[1] F. Avila Cobos, J. Vandesompele, P. Mestdagh, and K. De Preter, Computational deconvolution of transcriptomics data from mixed cell populations, Bioinformatics **34**, 1969 (2018).

[2] Y. Azimzade, M. Haugen, X. Tekpli, C. B. Steen, T. Fleischer, D. Kilburn, H. Ma, E. V. Egeland, G. Mills, O. Engebraaten, *et al.*, Explainable machine learning reveals the role of the breast tumor microenvironment in neoadjuvant chemotherapy outcome, bioRxiv , 2023 (2023).

[3] S. Z. Wu, G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy, N. Bartonicek, *et al.*, A single-cell and spatially resolved atlas of human breast cancers, Nature Genetics **53**, 1334 (2021).

[4] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, *et al.*, Determining cell type abundance and expression from bulk tissues with digital cytometry, Nature Biotechnology **37**, 773 (2019).

[5] L. Silwal-Pandit, S. Nord, H. von der Lippe Gythfeldt, E. K. Møller, T. Fleischer, E. Rødland, M. Krohn, E. Borgen, Ø. Garred, T. Olsen, *et al.*, The longitudinal transcriptional response to neoadjuvant chemotherapy with and without bevacizumab in breast cancermolecular response in bevacizumab-treated breast tumors, Clinical Cancer Research **23**, 4662 (2017).

[6] L. A. Korde, L. Lusa, L. McShane, P. F. Lebowitz, L. Lukes, K. Camphausen, J. S. Parker, S. M. Swain, K. Hunter, and J. A. Zujewski, Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer, Breast Cancer Research and Treatment **119**, 685 (2010).

[7] P. F. Lebowitz, J. Eng-Wong, S. M. Swain, A. Berman, M. J. Merino, C. K. Chow, D. Venzon, F. Zia, D. Danforth, E. Liu, *et al.*, A phase ii trial of neoadjuvant docetaxel and capecitabine for locally advanced breast cancer, Clinical cancer research **10**, 6764 (2004).

[8] Y. Lin, S. Lin, M. Watson, K. M. Trinkaus, S. Kuo, M. J. Naughton, K. Weilbaecher, T. P. Fleming, and R. L. Aft, A gene expression signature that predicts the therapeutic response of the basal-like breast cancer to neoadjuvant chemotherapy, Breast Cancer Research and Treatment **123**, 691 (2010).

[9] V. Popovici, W. Chen, B. D. Gallas, C. Hatzis, W. Shi, F. W. Samuelson, Y. Nikolsky, M. Tsyganova, A. Ishkin, T. Nikolskaya, *et al.*, Effect of training-sample size and classification difficulty on the accuracy of genomic predictors, Breast Cancer Research **12**, 1 (2010).

[10] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T.-M. Chu, F. M. Goodsaid, L. Pusztai, *et al.*, The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models, Nature biotechnology **28**, 827 (2010).

[11] A. Tabchy, V. Valero, T. Vidaurre, A. Lluch, H. Gomez, M. Martin, Y. Qi, L. J. Barajas-Figueroa, E. Souchon, C. Coutant, *et al.*, Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer30-gene t/fac response predictor in breast cancer, Clinical Cancer Research **16**, 5351 (2010).

[12] T. Iwamoto, G. Bianchini, D. Booser, Y. Qi, C. Coutant, C. Ya-Hui Shiang, L. Santarpia, J. Matsuoka, G. N. Hortobagyi, W. F. Symmans, *et al.*, Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer, Journal of the National Cancer Institute **103**, 264 (2011).

[13] S. Glück, J. S. Ross, M. Royce, E. F. McKenna, C. M. Perou, E. Avisar, and L. Wu, Tp53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine±trastuzumab, Breast Cancer Research and Treatment **132**, 781 (2012).

[14] K. Shen, Y. Qi, N. Song, C. Tian, S. D. Rice, M. J. Gabrin, S. L. Brower, W. F. Symmans, J. A. O'Shaughnessy, and F. A. Holmes, Cell line derived multi-gene predictor of pathologic response to neoadjuvant chemotherapy in breast cancer: a validation study on us oncology 02-103 clinical trial, BMC Medical Genomics **5**, 1 (2012).

[15] J. A. Bauer, A. B. Chakravarthy, J. M. Rosenbluth, D. Mi, E. H. Seeley, N. De Matos Granja-Ingram, M. G. Olivares, M. C. Kelley, I. A. Mayer, I. M. Meszoely, *et al.*, Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiationmarkers of taxane sensitivity in breast cancer, Clinical Cancer Research **16**, 681 (2010).

[16] C. Hatzis, L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, *et al.*, A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer, JAMA **305**, 1873 (2011).

[17] M. J. M. Magbanua, D. M. Wolf, C. Yau, S. E. Davis, J. Crothers, A. Au, C. M. Haqq, C. Livasy, H. S. Rugo, L. Esserman, *et al.*, Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response, Breast Cancer Research **17**, 1 (2015).

[18] T. Miyake, T. Nakayama, Y. Naoi, N. Yamamoto, Y. Otani, S. J. Kim, K. Shimazu, A. Shimomura, N. Maruyama, Y. Tamaki, *et al.*, Gstp 1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in er-negative breast cancer, Cancer Science **103**, 913 (2012).

[19] J. C. Liu, V. Voisin, G. D. Bader, T. Deng, L. Pusztai, W. F. Symmans, F. J. Esteva, S. E. Egan, and E. Zacksenhaus, Seventeen-gene signature from enriched her2/neu mammary tumor-initiating cells predicts clinical outcome for human her2+: Erα- breast cancer, Proceedings of the National Academy of Sciences **109**, 5832 (2012).

[20] A. Prat, G. Bianchini, M. Thomas, A. Belousov, M. C. Cheang, A. Koehler, P. Gómez, V. Semiglazov, W. Eiermann, S. Tjulandin, *et al.*, based pam50 subtype predictor identifies higher responses and improved survival outcomes in her2-positive breast cancer in the noah study, Clinical Cancer Research **20**, 511 (2014).

[21] W. E. Johnson, C. Li, and A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, Biostatistics **8**, 118 (2007).

[22] H. G. Russnes, O. C. Lingjærde, A.-L. Børresen-Dale, and C. Caldas, Breast cancer molecular stratification: from intrinsic subtypes to integrative clusters, The American Journal of Pathology **187**, 2152 (2017).

[23] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, Molecular portraits of human breast tumours, Nature **406**, 747 (2000).

[24] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, Evaluating the yield of medical tests, Jama **247**, 2543 (1982).

[25] M. Schmid, M. N. Wright, and A. Ziegler, On the use of harrell's c for clinical risk prediction via random survival forests, Expert Systems with Applications **63**, 450 (2016).

[26] G. Buyrukoğlu, Survival analysis in breast cancer: evaluating ensemble learning techniques for prediction, PeerJ Computer Science **10**, e2147 (2024).
[27] K. Evangeline I, S. A. Kirubha, and J. G. Precious, Survival analysis of breast cancer patients using machine learning models, Multimedia Tools and Applications **82**, 30909 (2023).
[28] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems **30** (2017).
[29] L. S. Shapley, Notes on the n-person game—ii: The value of an n-person game, (1951).
[30] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, *et al.*, Supervised risk predictor of breast cancer based on intrinsic subtypes, Journal of Clinical Oncology **27**, 1160 (2009).
[31] E. Danenberg, H. Bardwell, V. R. Zanotelli, E. Provenzano, S.-F. Chin, O. M. Rueda, A. Green, E. Rakha, S. Aparicio, I. O. Ellis, *et al.*, Breast tumor microenvironment structures are associated with genomic features and clinical outcome, Nature Genetics **54**, 660 (2022).
[32] J. Fox and S. Weisberg, Cox proportional-hazards regression for survival data, An R and S-PLUS companion to applied regression **2002** (2002).
[33] R. Denz and N. Timmesfeld, Visualizing the (causal) effect of a continuous variable on a time-to-event outcome, Epidemiology **34**, 652 (2023).
[34] X. Q. Wang, E. Danenberg, C.-S. Huang, D. Egle, M. Callari, B. Bermejo, M. Dugo, C. Zamagni, M. Thill, A. Anton, *et al.*, Spatial predictors of immunotherapy response in triple-negative breast cancer, Nature **621**, 868 (2023).
[35] S. Tietscher, J. Wagner, T. Anzeneder, C. Langwieder, M. Rees, B. Sobottka, N. de Souza, and B. Bodenmiller, A comprehensive single-cell map of t cell exhaustion-associated immune environments in human breast cancer, Nature Communications **14**, 98 (2023).
[36] R. Browaeys, W. Saelens, and Y. Saeys, Nichenet: modeling intercellular communication by linking ligands to target genes, Nature Methods **17**, 159 (2020).
[37] R. Bill, P. Wirapati, M. Messemaker, W. Roh, B. Zitti, F. Duval, M. Kiss, J. C. Park, T. M. Saal, J. Hoelzl, *et al.*, Cxcl9: Spp1 macrophage polarity identifies a network of cellular programs that control human cancers, Science **381**, 515 (2023).