

Insights into Breast Cancer from deconvolution of bulk RNA Sequencing Data

By Yves Greatti, Reem Shamma, and Lingzhu Shen

Abstract

Single-cell RNA sequencing (scRNA-seq) has enabled the identification of new cell subtypes and gene expression patterns within tumors. However, the cost and technical complexity of scRNA-seq still makes it impractical for large-scale clinical studies. Therefore, a promising approach is to use computational methods to deconvolve the cell-type composition of bulk RNA sequencing data, which can provide insights into the molecular mechanisms underlying the development and progression of cancer.

In this study, we applied a single-cell RNA deconvolution method to bulk RNA sequencing data from the Cancer Genome Atlas (TCGA) breast cancer (BRCA) dataset to identify cell-type-specific gene expression signatures associated with overall and disease-free survival. We used the Single-Cell Expression Atlas (SCEA) database to generate a reference gene expression matrix for 9 different breast cell types, including luminal and basal epithelial cells, myoepithelial cells, and immune cells. We also used a dataset from an existing publication in the literature (Gray et al.) that identifies cells related to breast cancer at both the transcriptomic and proteomic levels such as mammary epithelial cells (MEC), alveolar (AV), Hormone Sensing (HS), basal (BA), and stromal cells (fibroblasts, vascular/lymphatic cells, and immune cells) [1]. We then applied the Multi-subject Single-cell Deconvolution (MuSiC2) algorithm to estimate the relative proportions of these cell types in the bulk RNA sequencing data [2].

We identified cell-type-specific gene expression signatures associated with overall survival in breast cancer (BRCA) patients; expression of genes associated with B or T cells was positively associated with overall survival. These findings suggest that the immune response to BRCA tumors may play an important role in patient survival.

Our study demonstrates the potential of single cell RNA deconvolution methods to identify cell-type-specific gene expression signatures associated with clinical outcomes in large-scale clinical datasets. This approach may lead to the development of more effective diagnostic and therapeutic strategies for BRCA patients.

Commented [RA1]: fine?

Commented [YG2R1]: @Reem Abu-Shamma yes perfect

Commented [RA3]: move out of abstract?

Commented [YG4R3]: I am ok to do it but I see the abstract as a summary.

Commented [RA5]: expand in discussion?

Commented [YG6R5]: yes sure

Commented [YG7R5]: Are you going to do it?

Commented [RA8R5]: yes. Some of these comments are reminders to myself

Introduction

Cancer cells produce cytokines and chemokines that attract a diverse population of immune cells, including macrophages, neutrophils, and lymphocytes. The impact of these tumor-infiltrating immune cells has been debated. Some groups have shown that tumor-infiltrating immune cells may physically destroy tumor cells, thereby reducing tumor burden and improving clinical prognosis [8]. However, persistent activation of the immune system and failure of the inflammatory response to resolve may lead to chronic inflammation, which promotes tumor growth [7]. This inflammation promotes genomic instability, epigenetic modifications, and upregulation of cancer anti-apoptotic pathways, highlighting potential mechanisms of inflammation in promoting tumor growth and possibly metastasis [10].

Recent studies have shown that accounting for the heterogeneity of immune cell infiltration can result in more sensitive survival analyses and more accurate tumor subtype predictions [3,4]. Ongoing research is focused on the role of infiltrating lymphocytes and other immune cells in the tumor microenvironment (TME). Myeloid cells such as macrophages, monocytes, dendritic cells, neutrophils, basophils, and eosinophils are frequently found in the tissue of various tumors. In malignant tumors, levels of infiltrating immune cells are associated with tumor growth and cancer progression [5, 6].

Bulk RNA sequencing measures the average gene expression across all cells within a sample, and therefore cannot distinguish between different cell types or states. On the other hand, scRNA-seq enables researchers to identify and profile the transcriptome of individual cells, allowing for the characterization of cell types and their heterogeneity within a sample. By comparing bulk RNA expression data to scRNA-seq data from the same or similar tissues, deconvolution algorithms estimate the proportions of different cell types present in the bulk sample.

Breast cancer (BRCA) is one of the most common cancers among women worldwide. Despite advances in treatment, the prognosis for patients with BRCA remains highly variable. Recent studies have demonstrated that the heterogeneity of tumor cells and the TME can significantly impact patient outcomes, with greater heterogeneity corresponding to less immune cell infiltration, less activation of the immune response, and worse survival in breast cancer [9]. Identifying the cell-type-specific molecular mechanisms is needed to improve our understanding of the development and progression of BRCA tumors, and ultimately in enhancing diagnostic and therapeutic strategies.

The molecular subtypes of breast cancer depend on the genes the cancer cells express. The main molecular subtypes of invasive breast cancer are as follows [10]:

- Luminal A breast cancer: estrogen receptor (ER)-positive and progesterone receptor-positive, human epidermal growth factor receptor 2 (HER2) negative and has low levels of the protein Ki-67.

- Luminal B breast cancer: estrogen receptor-positive and HER2-negative, and either have high levels of Ki-67 or is progesterone receptor-negative
- HER2-enriched breast cancer: estrogen receptor-negative, progesterone receptor-negative, and HER2-positive
- Triple-negative breast cancer (TNBC) or basal-like breast cancer: lacks estrogen and progesterone receptors, lacks HER2 expression, is more prevalent in individuals with a BRCA1 mutation, and is the most aggressive subtype

Methods

The population data for this study was sourced from the Cancer Genome Atlas (TCGA) project, a collaborative effort between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to systematically analyze and catalog genomic and molecular data from various types of cancer. The TCGA data on BRCA includes information on DNA mutations, gene expression, epigenetic changes such as DNA methylation, and clinical data related to cancer survival and demographics. The TCGA-BRCA project consists of data from 1,111 cancer patients and 113 disease-free control patients.

RNA sequence data selected for this study was of the "Primary" and "Solid Tissue Normal" categories. Since MuSiC2 performs its own normalization, unstranded data was only considered and *TPM normalized* data were disregarded. The median age of the cohort was 58 years, and most patients were white (75.6%). The two most common subtypes of BRCA were BRCA_LumA (50.9%) and BRCA_LumB (20.1%), with most patients at stage IIA (32.9%), stage IIB (23.6%), and stage IIIA (14.4%) (Table 1). Data was collected using TCGAAbiolinks and TCGAWorkflow packages in R. To ensure consistency across the data, the ENSEMBL Id genes present in the TCGA dataset were converted into gene symbols using the genomic centric EnsDb.Hsapiens.v79 package. Any genes that were unresolved or duplicated were subsequently removed from the expression count matrix, to prevent any discrepancies or confounding factors in the downstream analysis.

Normal Cohort Statistics

Age

Count	113
Mean	57.33
Std	14.58
Min.	30
25%	45
50%	56
75%	66
Max	90

Commented [RA9]: we can omit this

Commented [YG10R9]: Eleanor mentioned something similar in her report

Commented [RA11]: what do we mean by 'unstranded' here?

Commented [YG12R11]: When you download the data you have different choices. The other choices are TPM normalized and we don't need them since MUSic does its own normalization. Also Eleanor recommended to use this version of the raw counts.

Tumor Cohort Statistics		
Age	Count	1,111
	Mean	58.42
	Std	13.21
	Min.	26
	25%	49
	50%	58
	75%	67
	Max	90
Tumor Stage	Stage I	16.8%
	Stage II	57.1%
	Stage III	23.1%
	Stage IV	3.1%
BRCA Subtype	Basal	17.4%
	Her2	8%
	LumA	50.9%
	LumB	20.1%
	Normal	3.7%

Table 1: Characteristics of normal (top) and tumor (bottom) cohorts obtained from TCGA for further analysis

For scRNA-seq data, two studies and their datasets were considered:

- Wu et al. (GSE176078) provided a more detailed understanding of the cellular and molecular heterogeneity within breast tumors [11]. The researchers performed scRNA-Seq (Chromium, 10X Genomics) on 26 primary tumors from three major subtypes of breast cancer (11 ER+, 5 HER2+, and 10 TNBC) and identified 9 major cell types, 29 minor cell types and 49 cell subtypes (Table 2). The study also found that macrophages with high expression of fatty acid metabolic genes FABP5 (LAM1), as well as macrophages that clustered around high levels of CXC chemokines 10 (CXCL10-hi) are key sources of immunosuppressive molecules within the human breast TME. Spatial analysis revealed the proximity of these macrophages to lymphocytes expressing programmed cell death 1 protein (PD-1+ lymphocytes).
- The second study and its corresponding dataset, Pal et al. (GSE161529), presents an extensive single-cell transcriptome map of over 430,000 cells (Table 3), from 52 patients [12]. They obtained the samples under various conditions including different hormonal stages, preneoplastic BRCA1+/- tissue, different cancer subtypes (4 TNBCs, 4 BRCA1 TNBCs,

Commented [RA13]: when finalizing, let's put these screenshots into a table

Commented [YG14R13]: yes I am going to do it

Commented [RA15R13]: Thank you ! Looks great

Commented [RA16]: is this information necessary/relevant to our downstream analysis? If not, we can leave out

Commented [YG17R16]: yes the definition of the cohorts are important. Eleanor for ex. added "The TCGA-obtained methylation beta values were from the Illumina Human Methylation 450 array. Differences in global methylation were assessed with a Wilcoxon rank-sum test." and "clinical data for the Glioblastoma Multiforme (a brain cancer) and Acute Myeloid Leukemia cancer subsets. Enhancer regions were downloaded from the FANTOM5 database for the associated organs: brain and blood (Andersson et al., 2014)."

Commented [RA18R16]: I can see how: ' They also identified that the LAM1 gene signature is strongly correlated with poor patient survival in large patient datasets, emphasizing the crucial role of these cells in the development and progression of breast cancer' ... describes the cohort, but I am not sure what the relevance of this part is: 'The study also found that macrophages with high expression of fatty acid metabolic genes FABP5 (LAM1), as well as macrophages that clustered around high levels of CXC chemokines 10 (CXCL10-hi) are key sources of immunosuppressive molecules within the human breast TME. Spatial analysis revealed the proximity of these macrophages to lymphocytes expressing programmed cell death 1 protein (PD-1+ lymphocytes)' . We can keep it for now but consider removing extraneous info when we revise the final version.

Commented [YG19R16]: @Reem Abu-Shamma this is what I will keep but you decide.

6 HER+ tumors), as well as matching tumor and involved axillary lymph node pairs. The data was downloaded using the GEOquery package.

Major Type	Minor Type
B-Cells	B Cells Memory, Naive
CAFs	MSC iCAF-like myCAF-like
Cancer Epithelial	Basal SC Cycling Her2 LumA, LumB
Endothelial	ACKR1 CXCL12 Endothelial Lymphatic RGS5
Normal Epithelial	Luminal Progenitors Mature Luminal Myoepithelial
Myeloid	Cycling Myeloid DCs Macrophage Monocyte
PVL	Cycling PVL PVL Differentiated PVL Immature
T-cells	Cycling T-cells NK cells NKT cells CD4+ CD8+

Table 2: Identification of major and minor cell types from Wu et al. (dataset GSE176078) [11].

Major Type	Minor Type
AV	AP
BA	BAa, BAb, BAX, BL
Fibroblast	F1, F2, F3, Fx
HS	HS (a, b, x)
Immune	I1 Myeloid cell I2 NK cell I3 T cell

Vascular and lymphatic	I4 B cell
	I5 Plasma cell
	VL2 Vascular endothelial
	VL3 Pericyte
	VL1 Lymphatic endothelial
	VL2 Vascular endothelial
	VL3 Pericyte

Table 3: Identification of major and minor cell types from Pan et al. (dataset GSE161529) [12].

Our study also aimed to determine the impact of various immune cell types on the immune response. To do this, we utilized ESTIMATE immune scores for TCGA BRCA patients and cross-validated our findings using an XGBoost regression model with optimized hyperparameters via the XGBoost Python API. Specifically, we trained the model on 75% of the available data and assessed its performance on the remaining 25%. The independent variables were the immune cell type proportions of the different TCGA tumor samples obtained by deconvolution and the outcome variable was the immune score for this patient as provided by the ESTIMATE R package. Cross-validation of the model and careful tuning of its parameters and using SHAP python package methods provided some insight into the importance of each immune cell type of the immune score.

To identify oncogenes from the estimated cell proportions we used the PROGENy R package, and the Python package, *decoupler*. PROGENy, performs gene set enrichment analysis (GSEA) and pathway analysis of gene expression data. The decoupler Python package utilizes statistical methods such as Weighed Sum (WMEAN) or Univariate Linear Model (ULM) and a prior knowledge on gene regulatory networks to predict the activity of transcription factors and pathways within a sample population.

We then investigated the potential correlation between cellular fractions and clinical outcomes in the TCGA BRCA cohort. To this end, we conducted survival analyses using TCGA clinical data obtained through the cBioPortal and the cBioPortalData R package. Specifically, we utilized a median-point strategy to divide patients into low and high cell type proportions. We then performed Kaplan-Meier survival analyses with a log-rank test using a Cox's proportional-hazard model from the Python package, *lifelines*. We then computed the hazard ratio with a 95% confidence interval and corresponding p-values, and generated Kaplan-Meier curves using the Kaplan Meier Estimator function in the Python package, *scikit-survival*. Overall, these analyses allowed us to assess any potential associations between cellular alterations and clinical outcomes, including overall survival (OS) and disease-free survival (DFS) of patients in the TCGA BRCA cohort.

Results

Commented [RA20]: add to methods

Commented [YG21R20]: @Reem Abu-Shamma done

Commented [RA22]: what do we mean by 'population samples' here? A population of cells? so like bulk RNA seq? from what I can find online, it seems like the input for this package is -omics data

Commented [YG23R22]: Bulk RNA gives gene raw counts for individuals compared to raw counts by cell. These tools work at individual level. They expect a matrix in x-axis genes and in y-axis sample Ids.

Deconvolution of Immune Cells From RNA-Seq Data

Using MUSiC2 for single-cell deconvolution, we were able to estimate the proportions of different immune cell subpopulations within each patient's tumor.

We used scRNA-seq data of normal individuals from GSE1611529 to deconvolute the bulk RNA seq data from the 113 disease-free control patients of the TCGA BRCA cohort. Analysis revealed that the disease-free control patients in our cohort had high proportions of HS, AV, and vascular and lymphatic cells (Fig. 1B), which is consistent with what was observed in normal GSE1611529 patients (Fig. 1A). While a statistical test such as MannWhitney-U comparing the two cohorts' cell type proportions would not return a meaningful result due to the small sample size of only 4 normal patients, these visual findings provide a rough validation of our method.

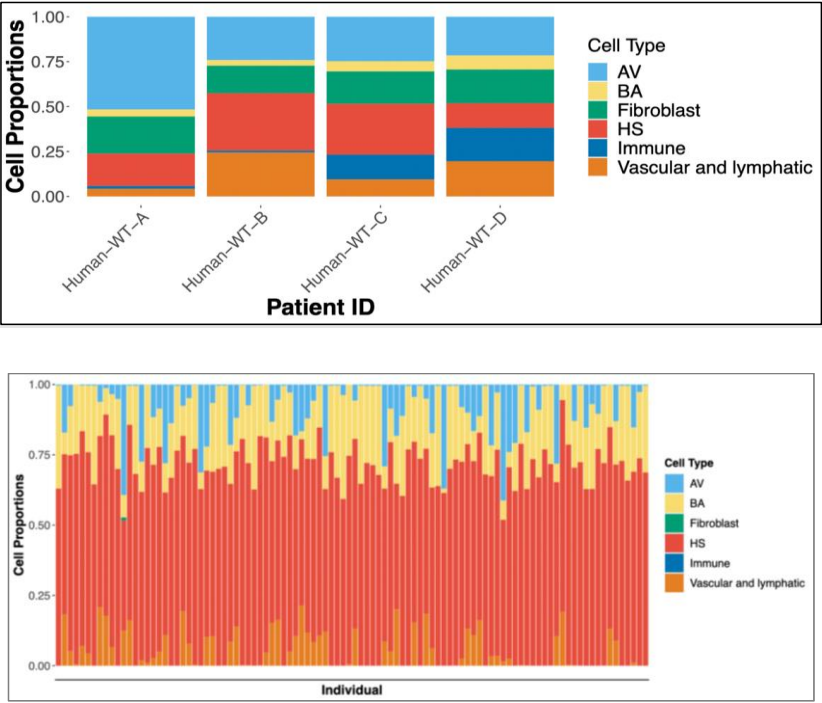


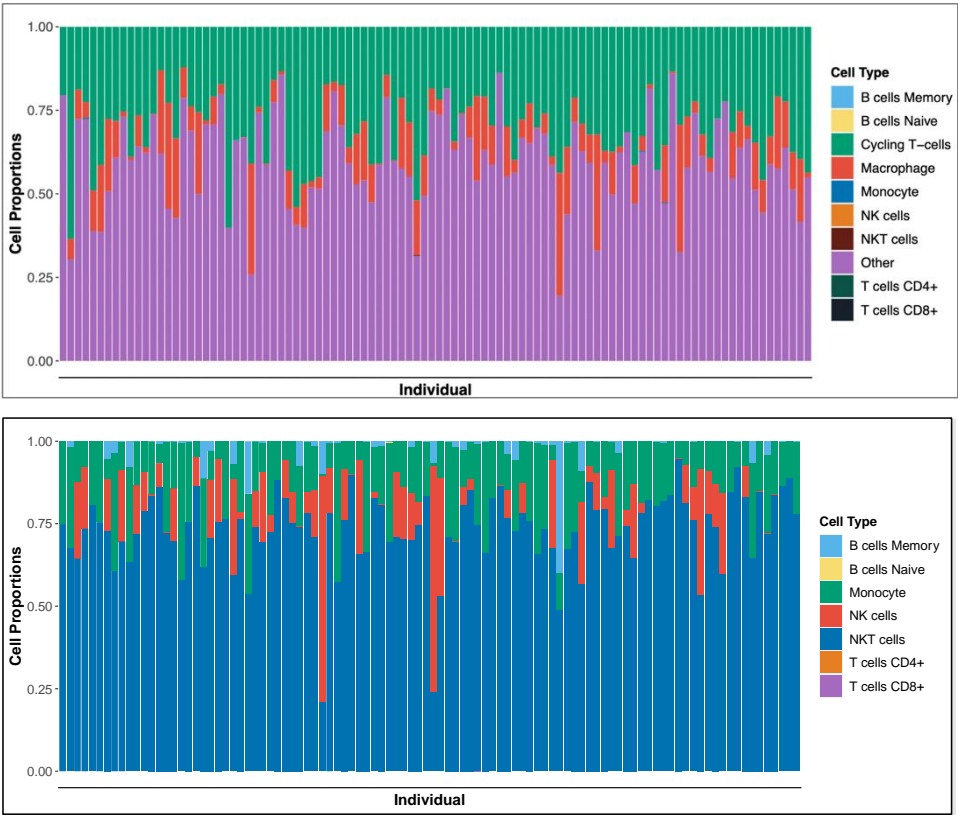
Figure 1: Cell type proportions from single-cell deconvolution using MUSiC2 of normal patients. **Fig 1A.** Cell type proportions from four individuals from the GSE1611529 dataset (top). **Fig 1B.** Cell types of normal patients in our TCGA BRCA cohort deconvoluted using GSE1611529 (bottom).

Commented [RA24]: confirm

In GSE17078 tumor patients, a significant presence of immune cells, such as macrophages, monocytes, and T cells (CD4+ and CD8+), was observed, with similar proportions found in the deconvolution output of TCGA breast cancer patients (Fig. 2A, and table 1 and 2 in appendix). However, upon excluding the most overrepresented cell types, macrophages and non-immune cells, the adjusted ratios allowed us to observe a substantial amount of NK and NKT cells, as well as memory B cells to a lesser extent (Fig. 2B-C). These findings highlight the complex and diverse nature of the immune cell composition within breast cancer tumors.

Commented [RA25]: why did we exclude macrophages too? Aren't they immune cells?

Commented [YG26R25]: They were omnipresent when including them, they were over-represented, it's all about ratios. To get an understanding of the other cells, I had to filter them out



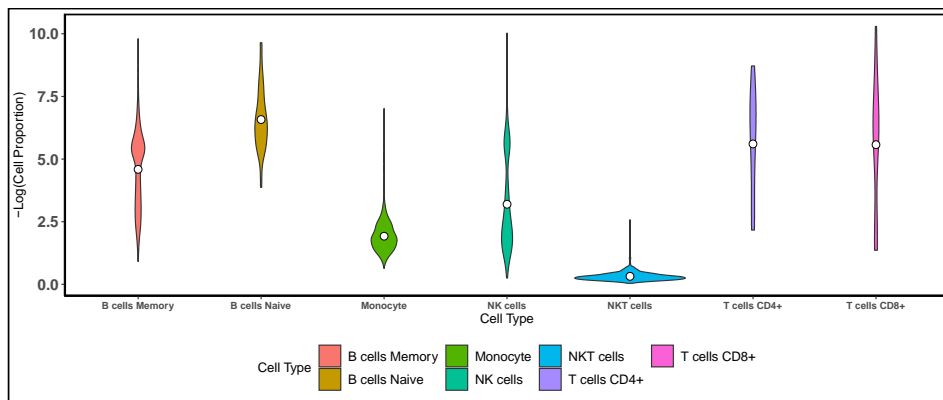
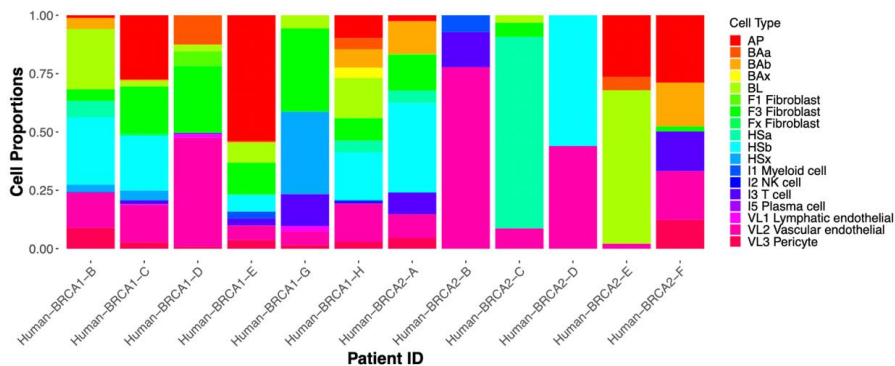


Figure 2: Cell type proportions from single-cell deconvolution using MUSiC2 of tumor patients. **Fig 2A.** Cell type proportions of tumor patients in the TCGA BRA cohort before excluding overrepresented cell types (top). **Fig 2B-C.** Cell type proportions in TCGA BRA cohort after excluding overrepresented cell types as stacked bar plots (middle) and violin plot (bottom).

Commented [RA27]: confirm

Commented [RA28]: confirm

The same MUSiC2 deconvolution technique was applied to analyze cell subtypes in tumor patients. The VL2 vascular endothelial and immune I3 T cell subtypes were predominant in both GSE161529 (Fig 3A) and TCGA (Fig. 3B) cancer patients, although their median levels were lower compared to other cell subtypes such as HSa, HSx, and BL (Fig 3C).



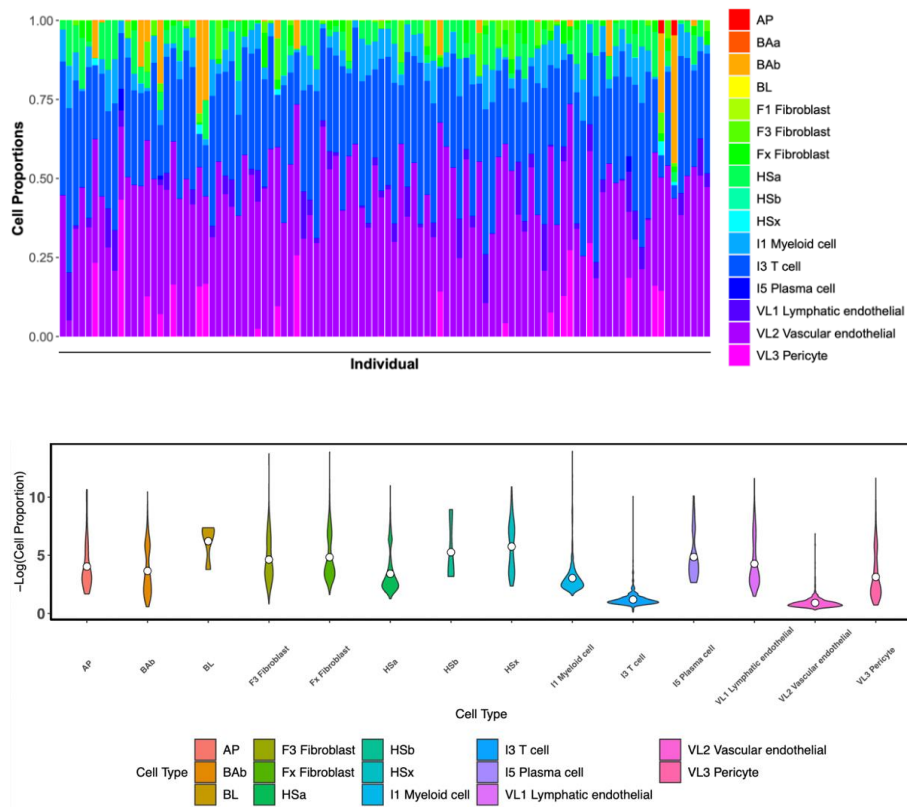


Figure 3: Cell subtype proportions from single-cell deconvolution using MUSiC2 of BRCA patients. **Fig 3A.** Cell subtypes of BRCA tumor patients from scRNA seq data of GSE161529. **Fig 3B-C.** TCGA BRCA cohort cell subtype proportions represented as stacked bar chart (middle) and violin plot (bottom) across individuals.

When computing the proportions of immune cells only (T cells, myeloid cells, and plasma cells) among the TCGA BRCA cohort, we found that T cells constituted the majority of immune cells (Fig. 4).

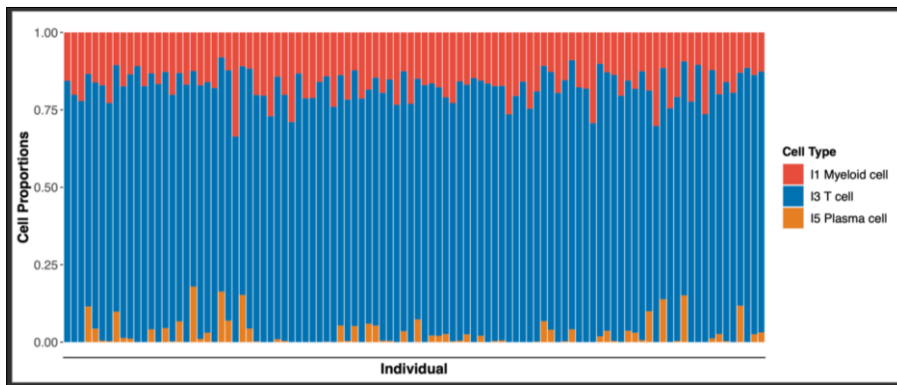


Figure 4: Immune cell proportions in TCGA BRA cohort.

These results suggest that while some immune cell subtypes are more prominent in breast cancer tumors, there is still significant heterogeneity in the immune cell composition across tumors.

Evaluation of Immune Cells in the Immune Response

Immune scores quantitatively measure the impact of immune cells on the immune response. Most of the ESTIMATE immune score distribution is similar between normal and tumor patient (Mann-Whitney-U test, $p\text{-val}=0.556$). However, some tumor patients have higher absolute immune scores compared to normal patients.

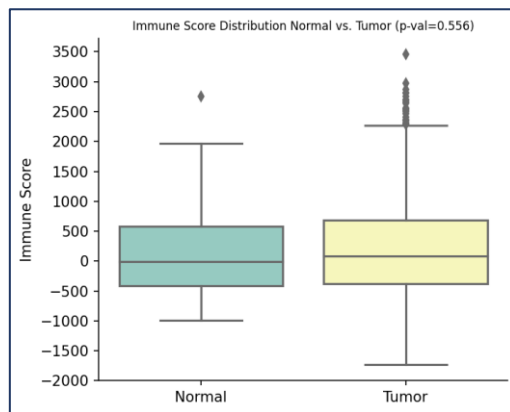


Figure 5: Immune score distribution across normal and tumor patients.

Commented [RA29]: this is for all immune cell types, considered together, correct?

Commented [YG30R29]: @Reem Abu-Shamma this is the immune score computed by ESTIMATE <https://bioinformatics.mdanderson.org/estimate/rpackage.html>

Commented [YG31R29]: @Reem Abu-Shamma ESTIMATE did not create these scores with a specific list of immune cell types.

After obtaining immune cell type proportions by deconvolution using MUSiC2, we developed a decision-tree based regression model ($R^2 = 0.647$, RMSE = 494.589) to predict the ESTIMATE immune score of the tumor patients. To determine the influence of each cell type on the immune score, we calculated the model coefficients and their p-values (Table 1). Higher proportions of CD8+ T cells and NKT cells reduce the immune score, indicating that they contribute less to immune response, while higher proportions of monocytes and NK cells increase immune scores and have a higher impact on immune response.

Additionally, we utilized Shapley values to estimate the contribution of each cell type to the immune system. Our findings revealed that monocytes, NK, and NKT cells were the most impactful contributors to the immune score among tumor patients (Figure 6).

Cell Type	Coefficient	P-value
B Cell Memory	2.773	0.728
B cells Naive	20.458	0.167
T cells CD8+	-106.369	< 0.001
T cells CD4+	42.236	0.230
NK cells	25.289	0.007
NKT cells	-744.481	< 0.001
Monocytes	1155.076	< 0.001

Table 1: Model coefficients and p-values

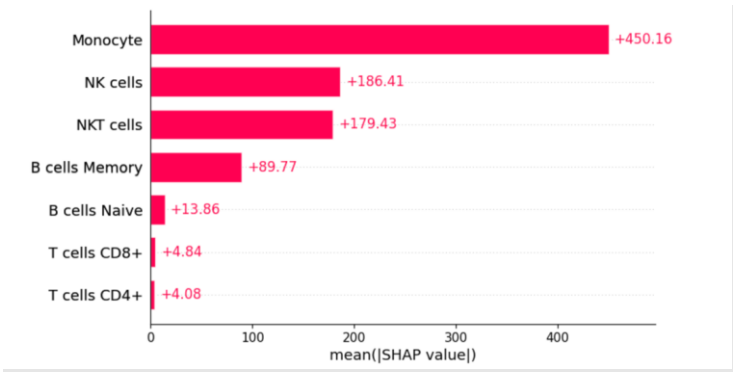


Figure 6: Distribution of mean absolute of SHAP values for each immune cell type.

Commented [RA32]: this is my interpretation of the results based on the sign of the coefficient values (+ vs -). Could you confirm that my understanding is correct please

Commented [YG33R32]: @Reem Abu-Shamma yes it is correct.

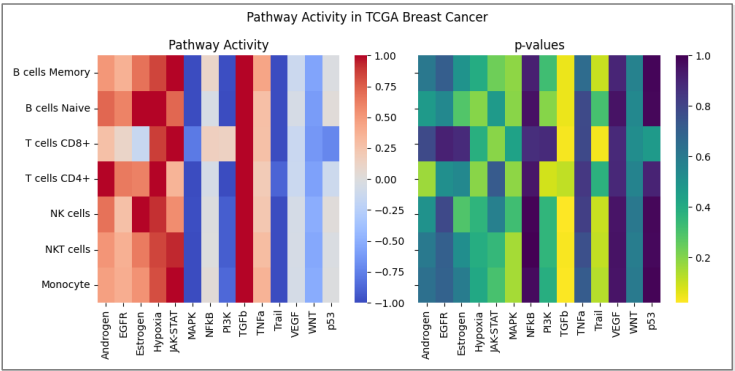
Commented [RA34]: As a control, would it be possible to run this same pathway analysis in normal tissues (i.e. using the GSE161529 set)? I think it would really help us see any differences that would be biologically relevant

Pathway Analysis across cell types in normal and BRCA patients

Within the context of the 14 cancer pathways investigated in our study, immune cells including B, T, and NK cells exhibited a trend of higher activity in pathways that regulate immune responses, such as TGF- β , but lower activity in pathways that induce apoptosis, such as Trail. Immune cells also show a significant reduction in the activity of the MAPK pathway, which is known to promote cell growth and proliferation (Fig. 7).

Each gene in PROGENy pathway has a weight representing its level within a given pathway (Appx. Table 3). Sorting these genes by weight shows that ID1, ID3, COM, PMEPA1, SMAD7 in the TGF- β pathway and RHEBL1, SMIM3, GPR18, RAB37, RNF175 in the Trail pathway are potential prognostic markers in different cancers. Similar correlations were found with the immune cell type proportions obtained from GSE161529 data deconvolution (Figure 7: middle and bottom plots, and Appx. Figure 2). Interestingly the same immune cell types in normal patients show higher activity in the Hypoxia pathway compared to tumor patients.

These observations highlight the multifaceted role of immune cells in cancer development and underscore the importance of considering the functional activity of these cells in the context of cancer pathways.



Commented [RA35]: all cells in the body have MAPK activity; is it particularly high for these cells?

Commented [YG36R35]: Compared to the other pathways, a quite significant p value

Commented [RA37R35]: but blue? so underexpressed?

Commented [YG38R35]: yes correct

Commented [RA39]: do we have a figure/table for this?

Commented [YG40R39]: For the list of 14 pathways, yes it is in GH

Commented [YG41R39]: @Reem Abu-Shamma I have added the table in appx.

Commented [RA42R39]: Thank you!

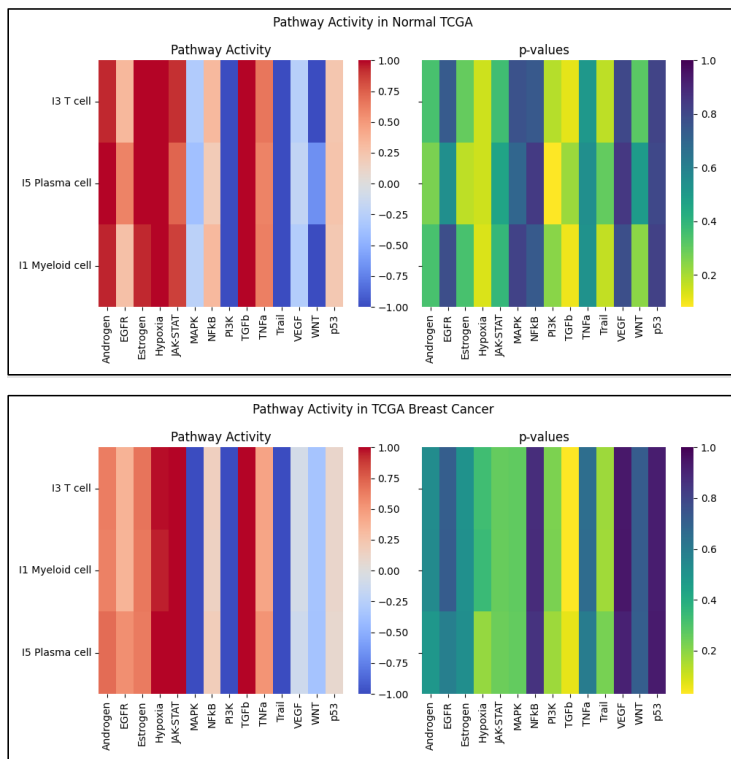


Figure 7: Pathway inference performed on the TCGA BRCA cohorts using the results obtained from deconvolution analysis (top). Immune cell activity levels across different pathways for TCGA normal and tumor patients (middle and bottom).

Correlation between cell type and clinical outcomes

To investigate potential clinical implications of differences in immune cell type and subtype proportions, we conducted an analysis of overall survival (OS) and disease-free survival (DFS) of the tumor patients in our BRCA TCGA cohort showing varying levels of cell type/subtypes.

Our analysis of the BRCA TCGA cohort with deconvolution using GSE177078 suggests that tumor patients with high levels of memory B cells had significantly lower OS and significantly higher DFS when compared to patients with low levels of memory B cells (Fig. 8). Significant differences in survival were not observed for varying levels of naïve B cells (Fig. 8). High levels of CD8+ T cells showed lower DFS that was near significant ($p\text{-val} = 0.07$) (Fig. 9). Similarly, high

Commented [RA43]: As a control, would it be possible to run this same pathway analysis in normal tissues (i.e. using the GSE161529 set)? I think it would really help us see any differences that would be biologically relevant

Commented [YG44R43]: @Reem Abu-Shamma done!

levels of NKT cells had significantly lower DFS, while low levels of NKT cells showed higher OS that was near significant (p-val = 0.07) (Fig. 10).

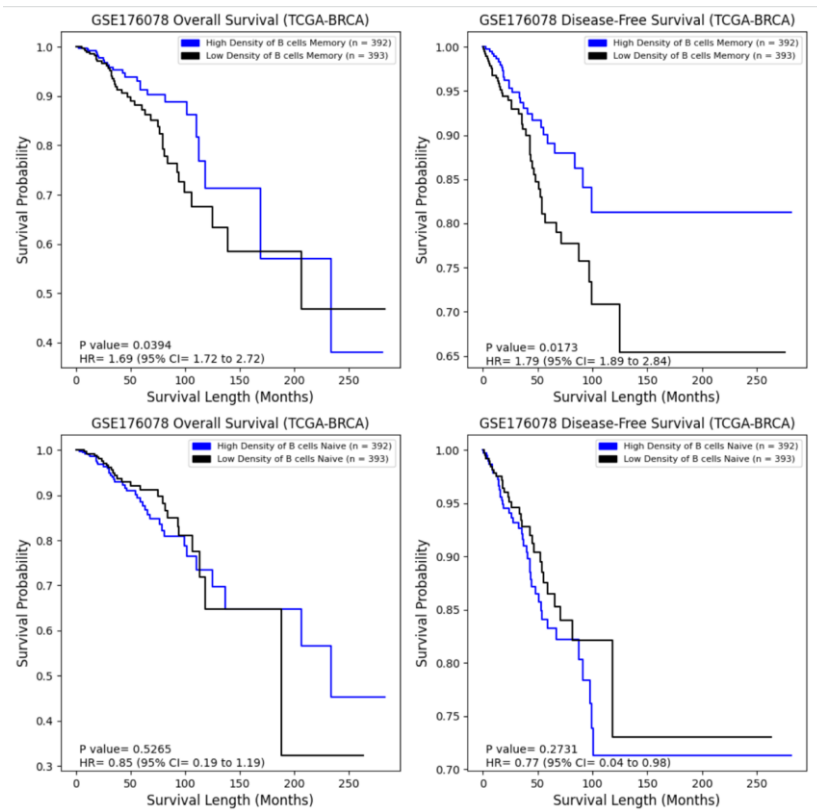


Figure 8: Difference in overall survival (left panels) and disease-free survival (right panels) across TCGA BRCA tumor patients (deconvoluted via GSE176078) showing different levels of memory B cells (top) and naïve B cells (bottom).

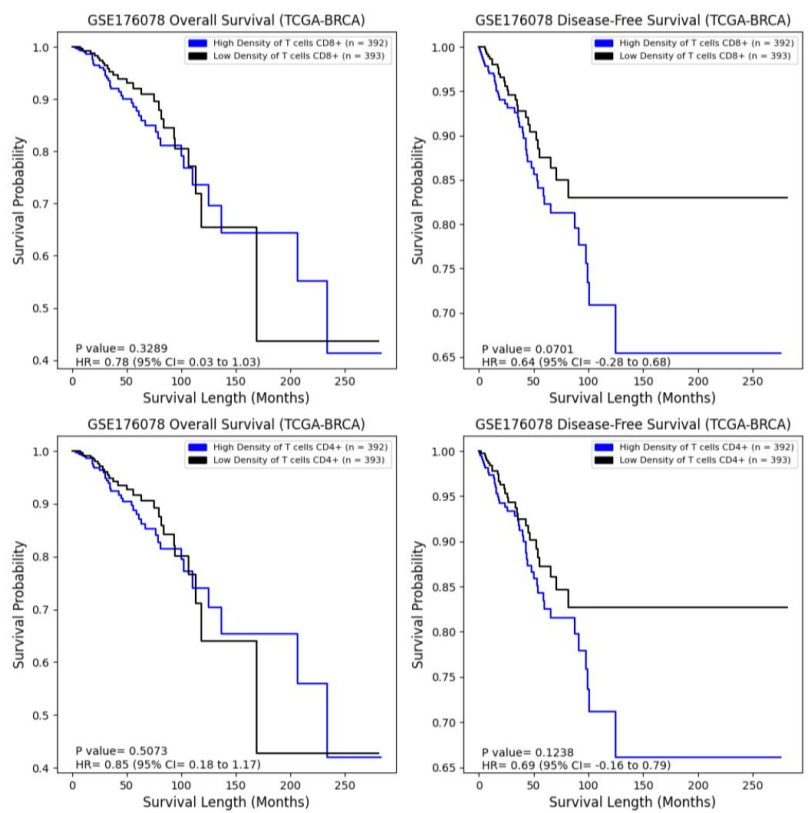


Figure 9: Difference in overall survival (left panels) and disease-free survival (right panels) across TCGA BRCA tumor patients (deconvoluted via GSE176078) showing different levels of CD8+ T cells (top) and CD4+ T cells (bottom).

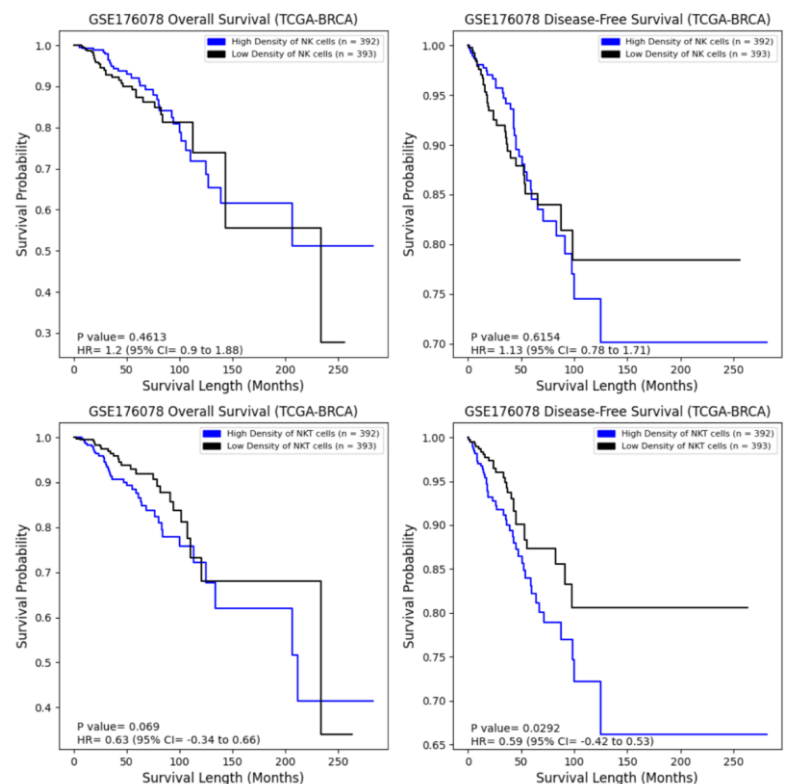


Figure 10: Difference in overall survival (left panels) and disease-free survival (right panels) across TCGA BRCA tumor patients (deconvoluted via GSE176078) showing different levels of NK cells (top) and NKT cells (bottom).

Interestingly, we did not observe significant differences in OS and DFS across varying levels of immune cells when conducting the same analysis using the GSE161529 dataset (Appx. Figure 1). Additionally, we found that differences in levels of vascular, lymphatic and alveolar cells did not significantly impact survival outcomes in GSE161529 (Appx. Figure 1). These results provide insight into the potential prognostic value of specific immune cell subpopulations in BRCA patients and underscore the importance of considering heterogeneity in immune cell composition when assessing clinical outcomes.

Commented [RA45]: want to include Immune_cell_survival_logranktest.png image in the appendix? or we can leave it like this

Commented [RA46R45]: Also for next sentence, can put VA_AV_cell_survival_logranktest.png in the appendix

Commented [YG47R45]: Not sure why you would want it? See papers critical figures are within the text and not in the appendix.

Commented [YG48R45]: oh sorry I misunderstood let me find it.

Commented [YG49R45]: @Reem Abu-Shamma Done!

Commented [RA50R45]: Thank you!

Discussion and Conclusion

In the following study, we used a deconvolution algorithm (MuSiC2) to interpret a myriad of bulk RNA-seq data, sampled from both tumor and normal/control patients in the TCGA BRCA cohort of patients. Two sc-RNA seq datasets were selected as references to deconvolute TCGA BRCA: GSE176078 and GSE161529. The former was applied to the bulk RNA-seq data from tumor samples in TCGA BRCA, and the latter was used on the normal/control patients within the same cohort. The MuSiC2 iterative algorithm relies on the assumption that when using bulk and single-cell reference samples from different clinical conditions, most cell-type-specific gene expression patterns will remain consistent across both. Therefore, isolating the few genes that show cell-type-specific differential expression between different clinical conditions in the scRNA reference datasets allows for accurate cell-type proportion estimates [2].

Our deconvolution results provided insight into the heterogeneous population of cell types and subtypes across the normal and tumor bulk RNA samples. Hormone sensing (HS) cells were the most populous cell type in the normal/control individuals, followed by basal (BA), vascular/lymphatic and alveolar (AV) cells (Fig. 1). These cells maintain the homeostatic functioning of breast tissue. For example, the hormone-sensing hormone receptor (HR+) cells in the luminal layer of epithelial ducts express steroid hormone receptors for estrogen, progesterone and prolactin, allowing them to participate in dynamic signaling pathways [13]. Meanwhile, BA cells in the outer basal layer of mammary glands have contractile features, allowing the movement of milk through the breast during lactation [14].

The relative proportion of immune cells was much higher in BRCA patients of the cohort as compared to normal patients. This is consistent with previous studies showing that the composition of immune cells in breast tissues progressively increases from normal to breast cancer [17]. Of the immune cell types present, cycling T-cells and macrophages predominated (Fig. 2), which is also consistent with literature findings [19]. T-cells are the most effective at triggering adaptive anti-tumor responses, although T-cell subtypes play varying roles [19]. CD8+ T cells are known to produce IL2 and IFN γ involved in tumor elimination and their presence is associated with good prognosis [19], although our findings did not support this (Fig. 6). CD4+ T cells produce a different subset of ILs, including IL4, IL5, IL13 IL21 and IL2, that correlate with tissue inflammation and have pro-tumor effects. This is consistent with the trend observed in our DFS analysis of CD4+ cells (Fig. 9), although our results were not statistically significant (p -val = 0.230). Macrophages are involved in tumor-associated inflammation, with a higher abundance of macrophages corresponding to poor prognosis. Cancer cells secreting colony-stimulating factor (CSF)-1 recruit tumor-associated macrophages that release epidermal growth factors (EGFs), modifying cancer cells to increase their migration and metastasis [19].

Our pathway analysis showed a significant upregulation of the TGF- β pathway across all immune cell types in deconvoluted tumor samples of the TCGA BRCA cohort (Fig. 7). In T-cells, TGF β inhibits the expression of transcription factors needed in T-cell differentiation, exerting

antiproliferative effects on T-cells [21]. Within the broader TME, high TGF- β expression is known to suppress the anti-tumor responses of Type 1 T helper cells [22]. TGF- β pathway activity blocks NK function in several different ways, including silencing NK surface receptors that facilitate recognition of stressed and transformed cells, or influencing NK differentiation to remove its cytotoxic activity [22]. The WNT pathway was more downregulated in normal compared to tumor samples. High WNT activity keeps peripheral T-cells in a more undifferentiated (i.e., proliferative) state and enhances their motility and migration [23]. This is consistent with our observations of higher T-cell proportions in tumor samples (Fig. 2, Appx. Table 1). Finally, MAPK activity was reduced in BRCA. The MAPK pathway increases the production of several inflammatory mediators, including the tumor-necrosis factor (TNF), IL-1 β and IL-6 [24].

Our survival analysis across BRCA patients with high versus low-density of immune cells showed conflicting results. In some cases, higher levels of immune cells (i.e., memory B cells) had a reduced probability of overall survival (OS), but an increased probability of disease-free survival (DFS) (Fig. 8). Meanwhile, high levels of NKT were associated with a significant reduction in DFS and a nearly significant increase in OS probability (Fig.10). These contradictions might represent the two opposing faces of the immune system's role in tumor development. On one hand, the immune system might prevent and control tumor progression through immunosurveillance, in which immune cells recognize and destroy cancer precursor cells [15, 16]. However, the immune system might also facilitate cancer progression through tumor-associated inflammation [7]. Particularly, B cells have been shown to have both positive and negative roles across different cancer types including BRCA; B cells may have anti-tumor effects by increasing T-cell functionality, thereby having favorable prognosis, but higher levels of B cells are also associated with worse prognosis in other studies [19,20]. Similarly, NK cells can have anti-tumor responses by releasing cytolytic granules that target and kill cancer cells, and by producing chemokines and cytokines that enhance the adaptive immune system's ability to target tumors. However, the types of tumor-related soluble factors in the microenvironment (e.g., IL-10) produced by different tumor-infiltrating immune cells (e.g., macrophage), may negatively affect NK cell's activity [19].

One limitation of our study is that the TCGA BRCA cohort pooled several different classes of BRCA together. The heterogeneity of immune cell compositions varies across cancer subtypes and even across individuals with the same cancer subtype [17, 18]. This heterogeneity creates a challenge in developing effective immunotherapies that restore and elicit anti-tumor immune responses by modulating specific tumor cells or antibodies. Diagnostic tools that evaluate patients through a more personalized-medicine approach might allow physicians to predict immunotherapy treatment responses on a per-individual basis. Single-cell analyses such as that used in this study may improve our understanding of interactions between tumor and immune cells in the TME and subsequently allow us to develop new drug targets for patients who do not respond to current immunotherapies [17, 25, 26].

Another limitation of our deconvolution approach in its potential clinical utility is that it does not provide information on the spatial distribution of immune cells within breast tissues. In normal breast tissue, both innate (e.g., NK cells) and adaptive (e.g. CD8+ and CD4+) immune cells are primarily present in epithelial tissues of breast ductal lobules. Throughout cancer progression, these same immune cell types, along with B cells and macrophages infiltrate the tumor parenchyma and stroma [17]. Therefore, coupling our single-cell deconvolution technique, which provides a quantitative measure of immune cell proportions, with histological analyses or molecular imaging that show the spatial distribution of cells, may provide a more comprehensive image of the unique TME composition of each patient, helping tailor individual treatment plans.

References

1. Gray, G. K., Li, C. M., Rosenbluth, J. M., Selfors, L. M., Girnius, N., Lin, J. R., Schackmann, R. C. J., Goh, W. L., Moore, K., Shapiro, H. K., Mei, S., D'Andrea, K., Nathanson, K. L., Sorger, P. K., Santagata, S., Regev, A., Garber, J. E., Dillon, D. A., & Brugge, J. S. (2022). A human breast atlas integrating single-cell proteomics and transcriptomics. *Developmental cell*, 57(11), 1400–1420.e7. <https://doi.org/10.1016/j.devcel.2022.05.003>
2. Fan, J., Lyu, Y., Zhang, Q., Wang, X., Li, M., & Xiao, R. (2022). MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Briefings in bioinformatics*, 23(6), bbac430. <https://doi.org/10.1093/bib/bbac430>
3. Huang, Ruichao, et al. "Combining Bulk RNA-Sequencing and Single-Cell RNA-Sequencing Data to Reveal the Immune Microenvironment and Metabolic Pattern of Osteosarcoma." *Frontiers in Genetics*, vol. 13, Oct. 2022, p. 976990. DOI.org (Crossref), <https://doi.org/10.3389/fgene.2022.976990>.
4. Lai, Wenwen, et al. "Integrated Analysis of Single-cell RNA-seq Dataset and Bulk RNA-seq Dataset Constructs a Prognostic Model for Predicting Survival in Human Glioblastoma." *Brain and Behavior*, vol. 12, no. 5, Apr. 2022, p. e2575. PubMed Central, <https://doi.org/10.1002/brb3.2575>.
5. Manoharan, Malini, et al. "A Computational Approach Identifies Immunogenic Features of Prognosis in Human Cancers." *Frontiers in Immunology*, vol. 9, Dec. 2018, p. 3017. PubMed Central, <https://doi.org/10.3389/fimmu.2018.03017>.
6. Qi, Zongtai, et al. "Single-Cell Deconvolution of Head and Neck Squamous Cell Carcinoma." *Cancers*, vol. 13, no. 6, Mar. 2021, p. 1230. DOI.org (Crossref), <https://doi.org/10.3390/cancers13061230>.
7. Gonzalez, H., Hagerling, C., & Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes & development*, 32(19-20), 1267–1284. <https://doi.org/10.1101/gad.314617.118>
8. Man, Y. G., Stojadinovic, A., Mason, J., Avital, I., Bilchik, A., Bruecher, B., Protic, M., Nissan, A., Izadjoo, M., Zhang, X., & Jewett, A. (2013). Tumor-infiltrating immune cells

- promoting tumor invasion and metastasis: existing theories. *Journal of Cancer*, 4(1), 84–95. <https://doi.org/10.7150/jca.5482>
9. McDonald, K. A., Kawaguchi, T., Qi, Q., Peng, X., Asaoka, M., Young, J., Opyrchal, M., Yan, L., Patnaik, S., Otsuji, E., & Takabe, K. (2019). Tumor Heterogeneity Correlates with Less Immune Response and Worse Survival in Breast Cancer Patients. *Annals of surgical oncology*, 26(7), 2191–2199. <https://doi.org/10.1245/s10434-019-07338-3>
 10. Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of Breast Cancer. In: Mayrovitz HN. editor. *Breast Cancer*. Brisbane (AU): Exon Publications. Online first 22 Jun 2022.
 11. Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., Wang, T., Larsson, L., Kaczorowski, D., Weisenfeld, N. I., Uytingco, C. R., Chew, J. G., Bent, Z. W., Chan, C. L., Gnanasambandapillai, V., Dutertre, C. A., ... Swarbrick, A. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9), 1334–1347. <https://doi.org/10.1038/s41588-021-00911-1>
 12. Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., Bryant, V. L., Penington, J. S., Di Stefano, L., Tubau Ribera, N., Wilcox, S., Mann, G. B., kConFab, Papenfuss, A. T., Lindeman, G. J., Smyth, G. K., & Visvader, J. E. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *The EMBO journal*, 40(11), e107333. <https://doi.org/10.15252/embj.2020107333>
 13. De Silva, D., Kunasegaran, K., Ghosh, S., & Pietersen, A. M. (2015). Transcriptome analysis of the hormone-sensing cells in mammary epithelial reveals dynamic changes in early pregnancy. *BMC developmental biology*, 15, 7. <https://doi.org/10.1186/s12861-015-0058-9>
 14. Gusterson, B., & Eaves, C. J. (2018). Basal-like Breast Cancers: From Pathology to Biology and Back Again. *Stem cell reports*, 10(6), 1676–1686. <https://doi.org/10.1016/j.stemcr.2018.04.023>
 15. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0280851#pone.0280851.ref008>
 16. Zitvogel, L., Tesniere, A. & Kroemer, G. Cancer despite immunosurveillance: immunoselection and immunosubversion. *Nat Rev Immunol* 6, 715–727 (2006). <https://doi.org/10.1038/nri1936>
 17. Goff, S. L., & Danforth, D. N. (2021). The Role of Immune Cells in Breast Tissue and Immunotherapy for the Treatment of Breast Cancer. *Clinical breast cancer*, 21(1), e63–e73. <https://doi.org/10.1016/j.clbc.2020.06.011>
 18. Bhat-Nakshatri P, Gao H, Sheng L, McGuire PC, Xuei X, Wan J, Liu Y, Althouse SK, Colter A, Sandusky G, Storniolo AM, Nakshatri H. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Rep Med*. 2021 Mar 16;2(3):100219. doi: 10.1016/j.xcrm.2021.100219. PMID: 33763657; PMCID: PMC7974552.

19. Gonzalez, H., Hagerling, C., & Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes & development*, 32(19-20), 1267–1284. <https://doi.org/10.1101/gad.314617.118>
20. Shen, M., Wang, J., & Ren, X. (2018). New Insights into Tumor-Infiltrating B Lymphocytes in Breast Cancer: Clinical Impacts and Regulatory Mechanisms. *Frontiers in immunology*, 9, 470. <https://doi.org/10.3389/fimmu.2018.00470>
21. Chen, C. H., Seguin-Devaux, C., Burke, N. A., Oriss, T. B., Watkins, S. C., Clipstone, N., & Ray, A. (2003). Transforming growth factor β blocks Tec kinase phosphorylation, Ca^{2+} influx, and NFATc translocation causing inhibition of T cell differentiation. *The Journal of experimental medicine*, 197(12), 1689-1699.
22. Batlle E, Massagué J. Transforming Growth Factor- β Signaling in Immunity and Cancer. *Immunity*. 2019 Apr 16;50(4):924-940. doi: 10.1016/j.immuni.2019.03.024. PMID: 30995507; PMCID: PMC7507121.
23. Staal, F., Luis, T. & Tiemessen, M. WNT signalling in the immune system: WNT is spreading its wings. *Nat Rev Immunol* 8, 581–593 (2008). <https://doi.org/10.1038/nri2360>
24. Liu, Y., Shepherd, E. & Nelin, L. MAPK phosphatases — regulating the immune response. *Nat Rev Immunol* 7, 202–212 (2007). <https://doi.org/10.1038/nri2035>
25. Debien, V., De Caluwé, A., Wang, X. *et al.* Immunotherapy in breast cancer: an overview of current strategies and perspectives. *npj Breast Cancer* 9, 7 (2023). <https://doi.org/10.1038/s41523-023-00508-3>
26. Galli, F., Aguilera, J.V., Palermo, B. *et al.* Relevance of immune cell and tumor microenvironment imaging in the new era of immunotherapy. *J Exp Clin Cancer Res* 39, 89 (2020). <https://doi.org/10.1186/s13046-020-01586-y>

Appendix

Statistical Results of Cell Types Proportions

Normal Cohort (113 patients)		
(source: GSE161529 – 4 patients)		
AV		
	Mean	0.081
	Std	0.098
	Min.	0
	25%	0.004
	50%	0.046
	75%	0.131
	Max.	0.413
Fibroblast		

HS	Mean	< 0.001
	Std	0.001
	Min.	0
	25%	0
	50%	< 0.001
	75%	0
	Max.	< 0.001
Vascular and lymphatic	Mean	0.675
	Std	0.080
	Min.	0.392
	25%	0.626
	50%	0.684
	75%	0.736
	Max.	0.828
Immune	Mean	0.050
	Std	0.063
	Min.	0
	25%	0
	50%	0.010
	75%	0.093
BA	Mean	0
	Std	< 0.001
	Min.	0
	25%	0
	50%	< 0.001
	75%	< 0.001
	Max.	0.252

TCGA BRCA (1,111 patients)**(source GSE161529 – 26 patients)**

AV

Mean	0.003
Std	0.025
Min.	0
25%	0
50%	0
75%	0
Max.	0.366

Fibroblast

Mean	0.057
Std	0.063
Min.	0
25%	0.007
50%	0.038
75%	0.086
Max.	0.533

HS

Mean	0
Std	0.003
Min.	0
25%	0
50%	< 0.001
75%	0
Max.	0.091

Vascular and lymphatic

Mean	0.515
Std	0.146
Min.	0
25%	0.427
50%	0.527
75%	0.625
Max.	0.840













Immune

Mean	0.405
Std	0.156
Min.	0
25%	0.298
50%	0.400
75%	0.5106
Max.	0.971

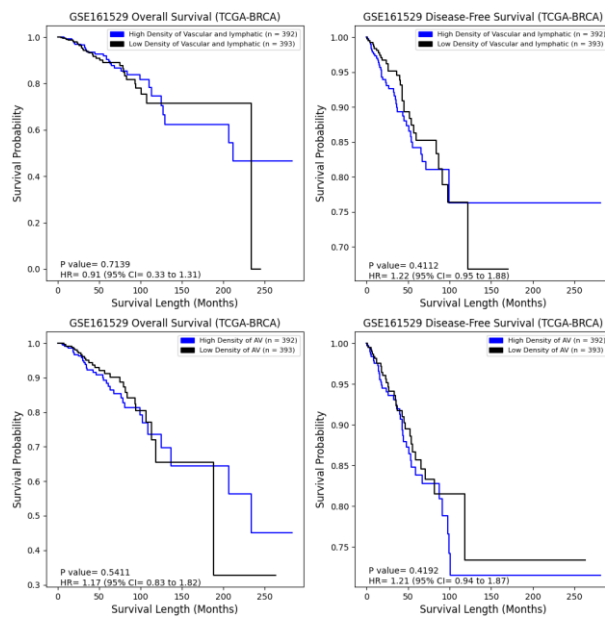
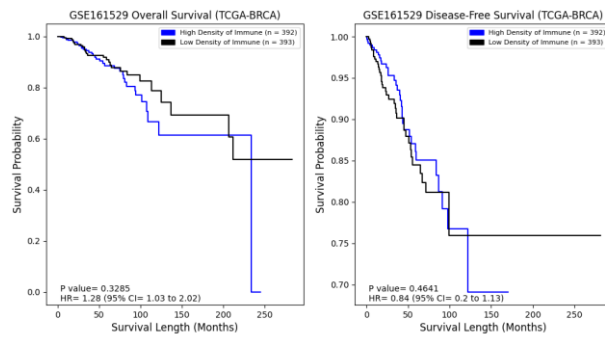
BA

Mean	0
Std	0.057
Min.	0
25%	0
50%	0.017
75%	0
Max.	0.462

Appx. Table 1: Statistical results of cell Types proportions Normal vs. Tumor patients

	AV	FIBROBLAST	HS	VASCULAR AND LYMPHATIC	IMMUNE	BA
NORMAL						
TUMOR						

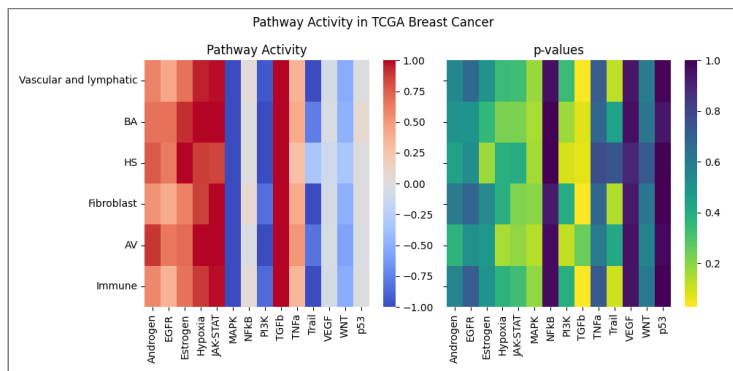
Appx. Table2: Statistical Trends of Cell Type Proportions Normal vs. Tumor



Appx. Fig.1: Analysis of immune cells (top), vascular and lymphatic, and alveolar cells (bottom) on the patient's overall survival and disease-free survival from GSE161529.

	source	target	weight	p_value
0	TGFb	ID1	12.354	0.00000
1	TGFb	ID3	10.481	0.00000
2	TGFb	COMP	9.899	0.00000
3	TGFb	PMEPA1	8.096	0.00000
4	TGFb	SMAD7	7.631	0.00000
5	TGFb	RFLNB	7.309	0.00000
6	TGFb	FSTL3	6.807	0.00000
7	TGFb	AMIGO2	6.470	0.00001
8	TGFb	SERPINE1	6.461	0.00002
9	TGFb	CTPS1	6.372	0.00000
10	Trail	RHEBL1	4.129	0.00200
11	Trail	SMIM3	3.712	0.03400
12	Trail	GPR18	3.241	0.00000
13	Trail	RAB37	2.948	0.00900
14	Trail	RNF175	2.801	0.01300
15	Trail	UAP1L1	2.797	0.01500
16	Trail	SELL	2.472	0.03200
17	Trail	BRI3	2.352	0.05300
18	Trail	GSDME	2.348	0.06400
19	Trail	WT1-AS	2.225	0.00000

Appx. Table3: Pathway model target genes, sorting by descending weight.



Appx. Fig.2: Pathway analysis for Vascular and lymphatic, BA, HS, Fibroblast and AV cells.