EN.585. 788: Foundations of Computational Biology and Bioinformatics – Project Rough Draft

Authors: Yves Greatti, Reem Abu Shamma, Lungzhu 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 (MuSiC) algorithm to estimate the relative proportions of these cell types in the bulk RNA sequencing data [2].

Commented [RA1]: move out of abstract?

Commented [YG2R1]: I am ok to do it but I see the abstract as a summary.

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.

Commented [RA3]: expand in discussion?

Commented [YG4R3]: yes sure

Commented [YG5R3]: Are you going to do it?

Commented [RA6R3]: yes. Some of these comments are reminders to myself

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.

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 [11]:

- 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 MuSiC 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 TCGAbiolinks 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.

Commented [RA7]: we can omit this

Commented [YG8R7]: Eleanor mentioned something similar in her report

Commented [RA9]: what do we mean by 'unstranded' here?

Commented [YG10R9]: 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.

Normal Cohort Statistics		
Age		
	Count	113
	Mean	57.33
	Std	14.58
	Min.	30
	25%	45
	50%	56
	75%	66
	Max	90
Race		
	Asian	1 (0.9%)
	Black or African American	6 (5%)
	White	105 (92.9%)
	Not Reported	1 (0.9%)

Tumor Cohort Statistics		
Age		
	Count	1,111
	Mean	58.42
	Std	13.21
	Min.	26
	25%	49
	50%	58
	75%	67
	Max	90
Race		
	American Indian or Alaska Native	1 (0.01%)
	Asian	60 (0.6%)
	Black or African American	182 (18.3%)
	White	751 (75.6%)

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 [12]. 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). 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.
- 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 [13]. They obtained the samples under various conditions including different hormonal stages, preneoplastic BRCA1+/- tissue, different cancer subtypes (4 TNBCs, 4 BRCA1 TNBCs, 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

Commented [RA11]: when finalizing, let's put these screenshots into a table

Commented [YG12R11]: yes I am going to do it

Commented [RA13R11]: Thank you ! Looks great

Commented [RA14]: is this information necessary/relevant to our downstream analysis? If not, we can leave out

Commented [YG15R14]: 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 [RA16R14]: 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.

	B cells Naive
CAFs	CAFs MSC iCAF-like
	CAFs myCAF-like
Cancer Epithelial	Cancer Basal SC
	Cancer Cycling
	Cancer Her2 SC
	Cancer LumA SC
	Cancer LumB SC
Endothelial	ACKR1
	CXCL12
	Endothelial
	Lymphatic LYVE1
	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) [12].

Major Type	Minor Type
AV	AP
	BAa
	BAb
	BAx
	BL
	Has
	Hsb
	HSx
BA	AP

	BAa, BAb, BAx, BL
	HSb, HSx
	HSx
Fibroblast	F1, f2, F3, Fx
	I1 Myeloid cell
	VL3 Pericyte
HS	AP
	BAb, BAx, BL
	Has, HSb, HSx
Immune	F3, Fx Fibroblast
	I1 Myeloid cell
	I2 NK cell
	I3 T cell
	I4 B cell
	I5 Plasma cell
	VL2 Vascular endothelial
	VL3 Pericyte
Vascular and lymphatic	F3, Fx Fibroblast
	VL1 Lymphatic endothelial
	VL2 Vascular endothelial
	VL3 Pericyte

Table 3: Identification of major and minor cell types from Pan et al. (dataset GSE161529) [13].

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.

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.

Commented [RA17]: 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 [YG18R17]: 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.

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.

Commented [RA19]: add to methods

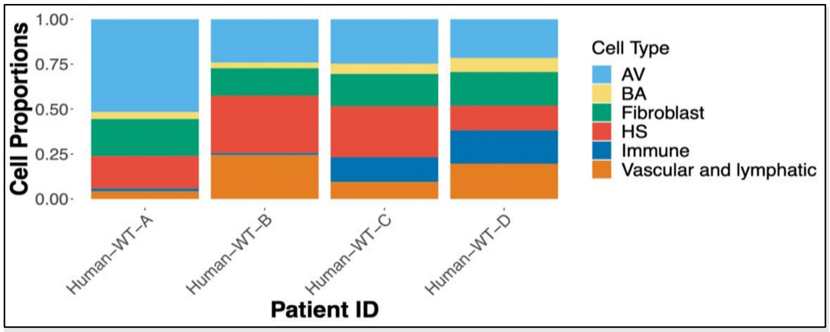
Commented [YG20R19]: @Reem Abu-Shamma done

Results

Deconvolution of Immune Cells From RNA-Seq Data

Using MUSiC 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.



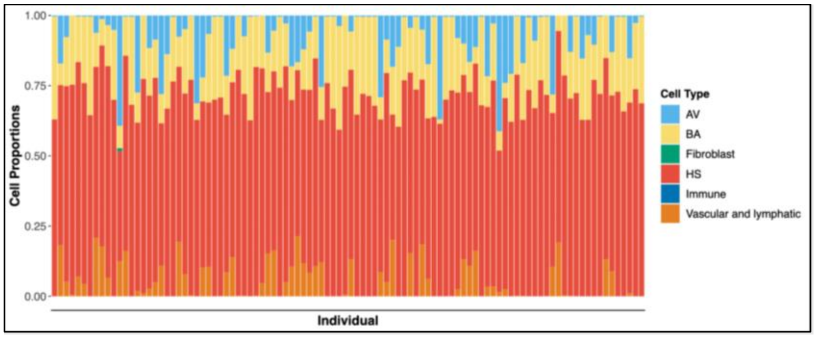


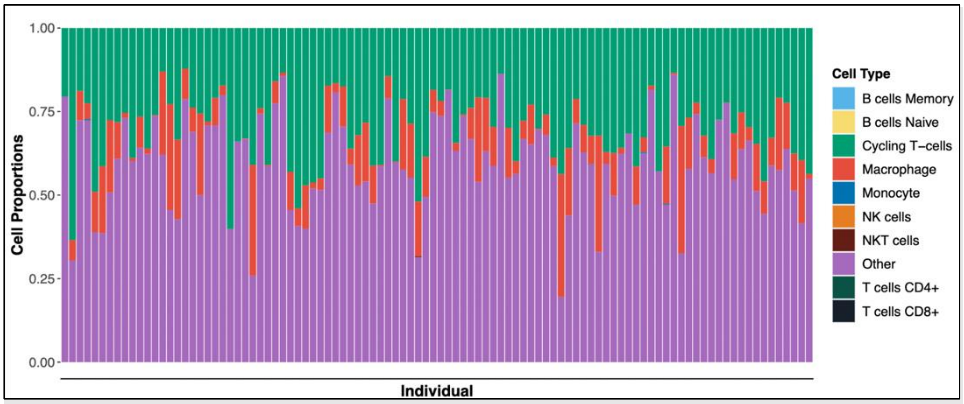
Figure 1: Cell type proportions from single-cell deconvolution using MUSiC of normal patients. **Fig 1A.** Cell type proportions from four individuals from GSE1611529 (top). **Fig 1B.** Cell types of normal patients in our TCGA BRCA cohort (bottom).

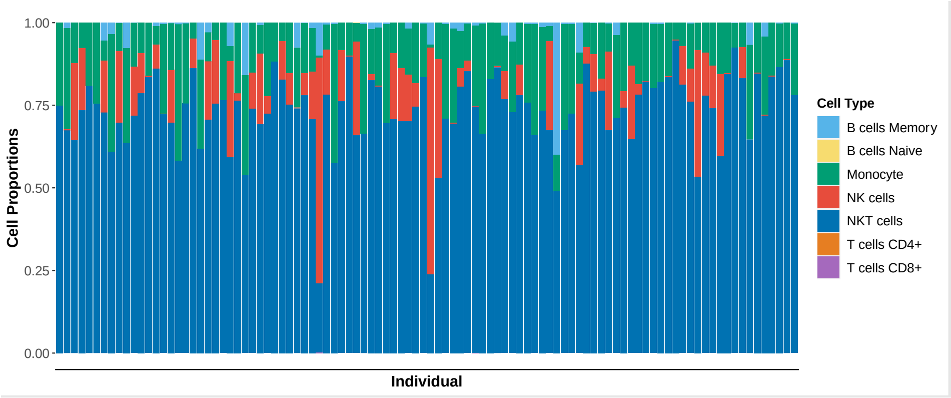
Commented [RA21]: confirm

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 ratio allowed us to observe a substantial presence 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 [RA22]: why did we exclude macrophages too? Aren't they immune cells?

Commented [YG23R22]: 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




1.00
0.75
0.50
0.25
0.00
Cell Proportions
Individual
Cell Type
B cells Memory
B cells Naive
Monocyte
NK cells
NKT cells
T cells CD4+
T cells CD8+

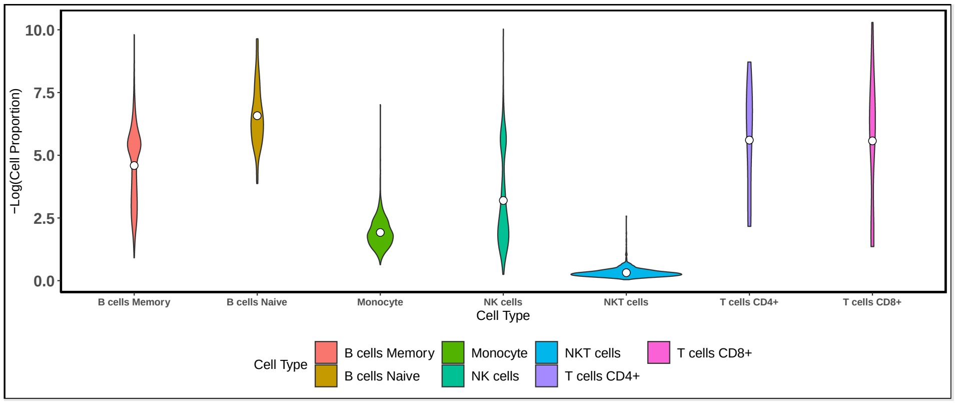
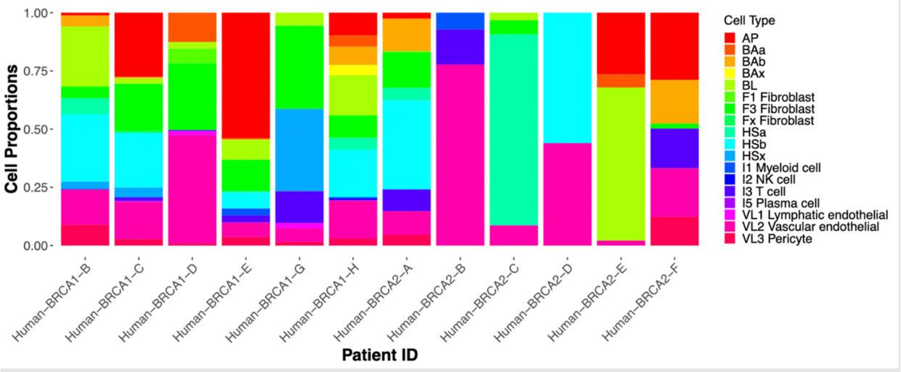


Figure 2: Cell type proportions from single-cell deconvolution using MUSiC of tumor patients. **Fig 2A.** Cell type proportions of tumor patients in the TCGA BRA cohort before excluding overrepresented cell types (top). **Fig 2B-C.** Cell type proportions in TCGA BRA cohort after excluding overrepresented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

Commented [RA24]: confirm

Commented [RA25]: confirm

The same MUSiC 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 cancer patients (Fig. 3B), although their median levels were lower compared to other cell subtypes such as Has, HSx, and BL (Fig 3C).



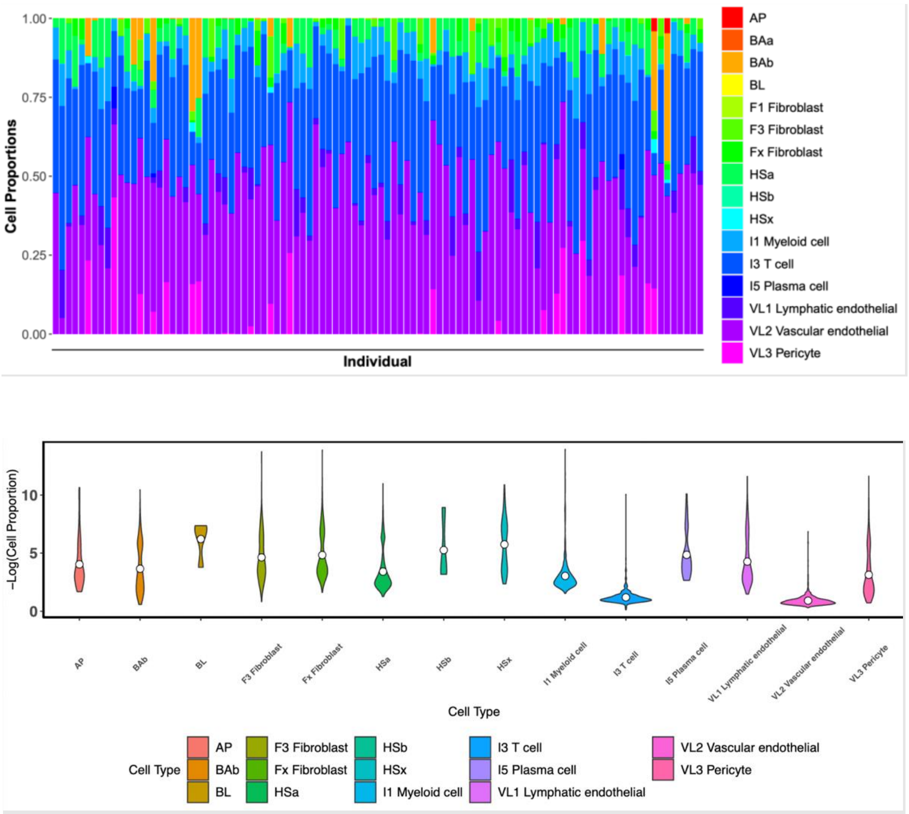


Figure 3: Cell subtype proportions from single-cell deconvolution using MUSiC of BRCA patients. **Fig 3A.** Cell subtypes of BRCA tumor patients from GSE161529. **Fig 3B-C.** TCGA BRCA cohort cell subtype proportions represented as stacked bar chart (middle) and violin plot (bottom) across individuals.

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).

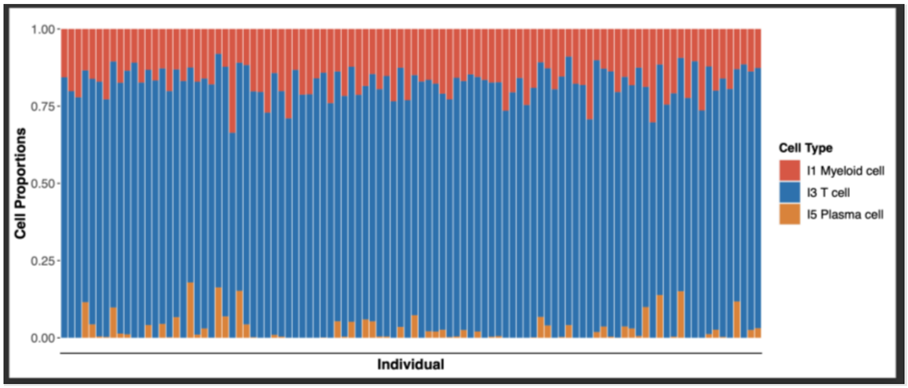


Figure 4: Immune cell proportions in TCGA BRA cohort.

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.

Cell Fractions Clinical Outcome Correlation

To investigate potential clinical implications of differences in immune cell type and subtype proportions across normal and BCRA patients, we conducted an analysis of overall survival (OS) and disease-free survival (DFS).

Our analysis of the GSE177078 cohort suggests that tumor patients with high levels of memory B cells had significantly greater OS and DFS when compared to patients with low levels of memory B cells (Figure 5). This difference was not observed for varying levels of naïve B cells (Figure 5). Levels of CD8+ T cells (Figure 6) and NKT cells (Figure 7) also correlated with improved DFS outcomes, when compared to patients with lower proportions of the same cell types.

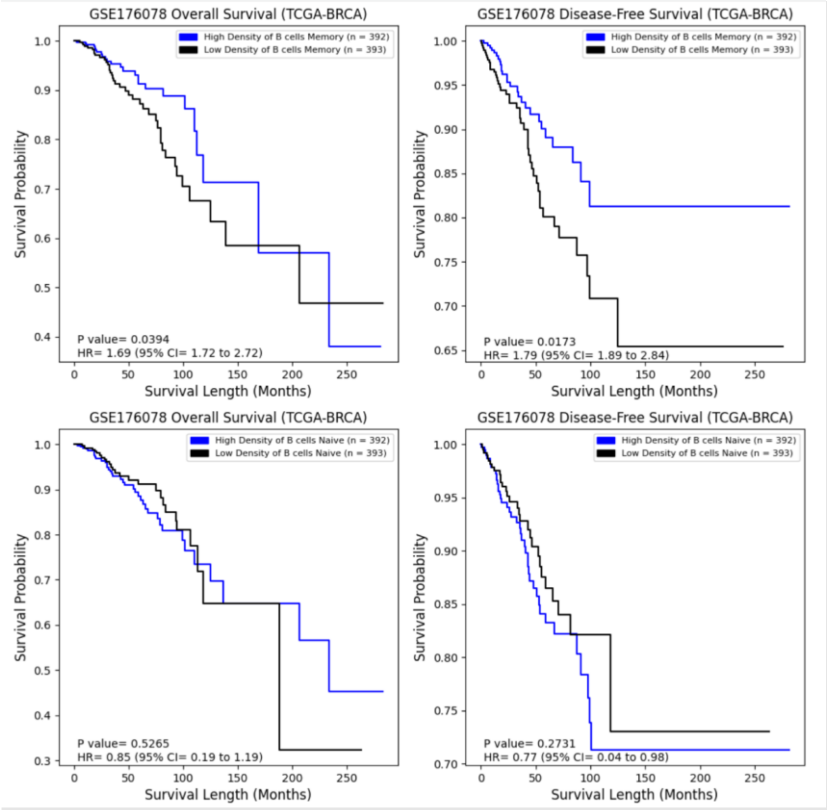


Figure 5: Difference in overall survival (left panels) and disease-free survival (right panels) across tumor patients in the GSE176078 tumor cohort showing different levels of memory B cells (top) and naïve B cells (bottom).

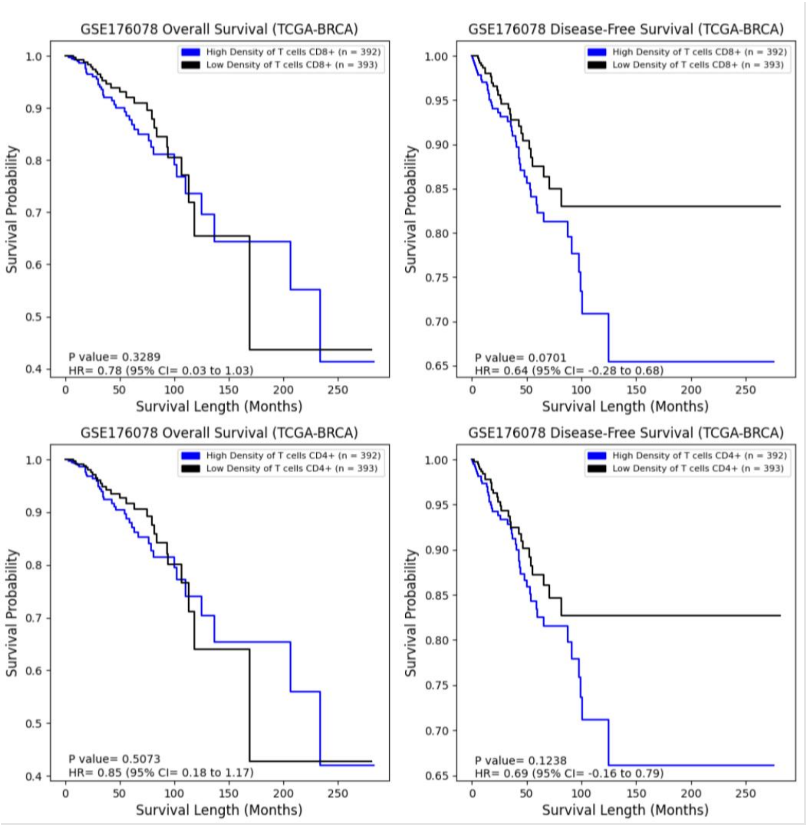
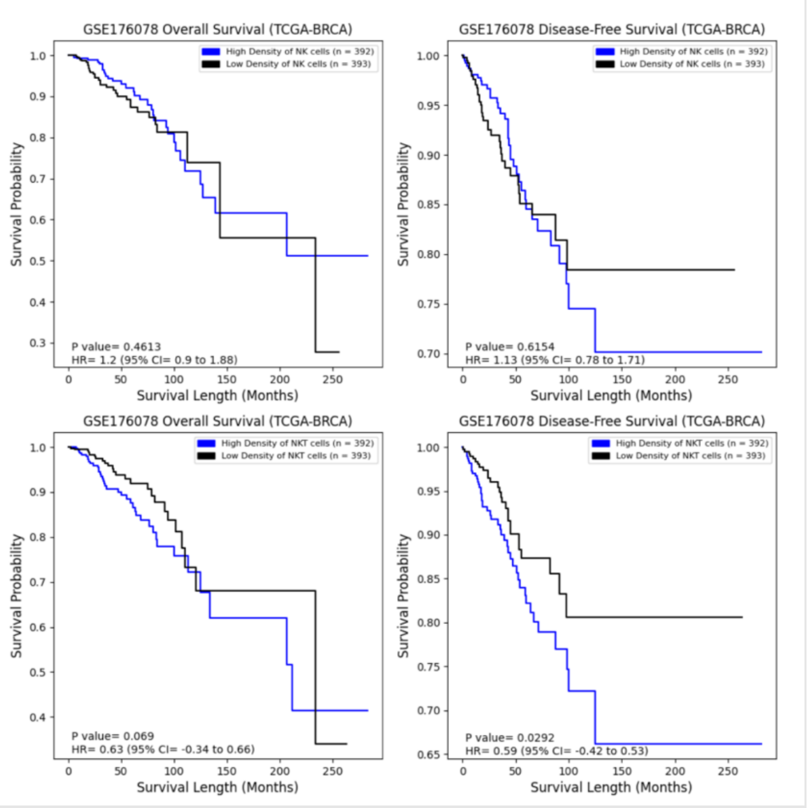


Figure 6: Difference in overall survival (left panels) and disease-free survival (right panels) across tumor patients in the GSE176078 tumor cohort showing different levels of CD8+ T cells (top) and CD4+ T cells (bottom).



GSE176078 Overall Survival (TCGA-BRCA)

GSE176078 Disease-Free Survival (TCGA-BRCA)

GSE176078 Overall Survival (TCGA-BRCA)

GSE176078 Disease-Free Survival (TCGA-BRCA)

Figure 7: Difference in overall survival (left panels) and disease-free survival (right panels) across tumor patients in the GSE176078 tumor cohort showing different levels of NK cells (top) and NKT cells (bottom).

Interestingly, we did not observe the same trends for the immune cells when using the GSE161529 dataset and corresponding single-cell bulk-RNA deconvolution (Appx. Figure 1). Additionally, we found that vascular and lymphatic cells, as well as alveolar cells, did not significantly impact survival outcomes (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 [RA26]: want to include Immune_cell_survival_logranktest.png image in the appendix? or we can leave it like this

Commented [RA27R26]: Also for next sentence, can put VA_AV_cell_survival_logranktest.png in the appendix

Commented [YG28R26]: Not sure why you would want it? See papers critical figures are within the text and not in the appendix.

Commented [YG29R26]: oh sorry I misunderstood let me find it.

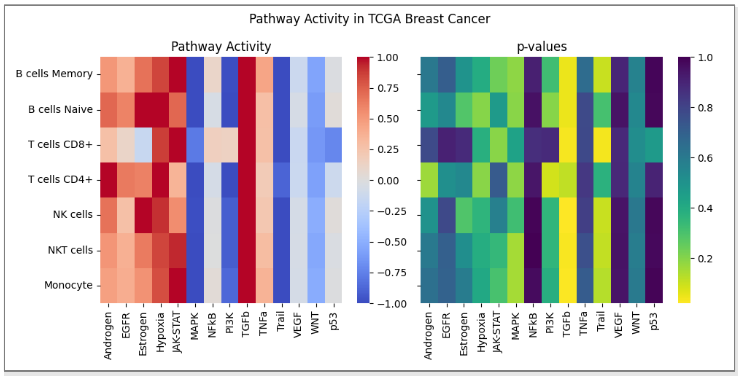
Commented [YG30R26]: @Reem Abu-Shamma Done!

Commented [RA31R26]: Thank you!

Pathway Analysis

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 TGFb, but lower activity in pathways that induce apoptosis, such as Trail. It is noteworthy that these same cell types have a significant reduction in the activity of the MAPK pathway, which is known to promote cell growth and proliferation (Fig. 8). 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 TGFb 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 8: middle and bottom plots, and Appx. Figure 2). Interestingly the same immune cell types in normal patients are more active 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 [RA32]: all cells in the body have MAPK activity; is it particularly high for these cells?

Commented [YG33R32]: Compared to the other pathways, a quite significant p value

Commented [RA34R32]: but blue? so underexpressed?

Commented [YG35R32]: yes correct

Commented [RA36]: do we have a figure/table for this?

Commented [YG37R36]: For the list of 14 pathways, yes it is in GH

Commented [YG38R36]: @Reem Abu-Shamma I have added the table in appx.

Commented [RA39R36]: Thank you!

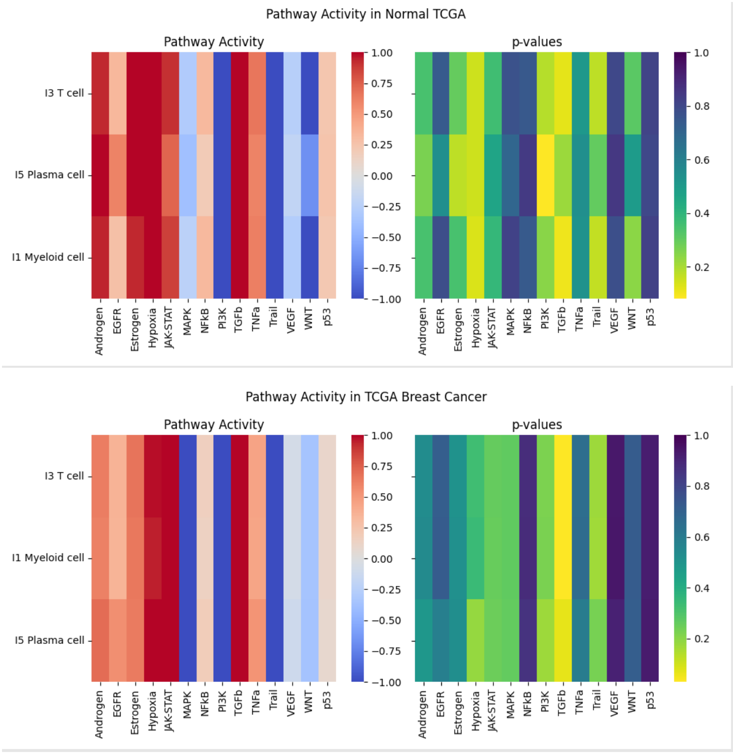
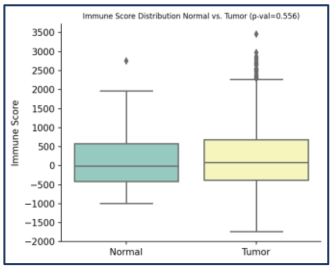


Figure 8: Pathway inference performed on the TCGA BRCA cohorts using the results obtained from deconvolution analysis (top). Immune cell activity levels across different pathways for TCGA normal and tumor patients (middle and bottom).

Commented [RA40]: 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 [YG41R40]: @Reem Abu-Shamma done!

Importance evaluation of Immune Cell Type in the Immune Response



Most of the immune score distribution is similar between normal and tumor patient (Mann-Whitney-U test p-val=0.556). However, some tumor patients have higher absolute immune scores compared to normal patients.

After obtaining immune cell type proportions from the MUSiC algorithm, we developed a decision-tree based regression model (R2 = 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 4). 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 9).

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 4: Model coefficients and p-values

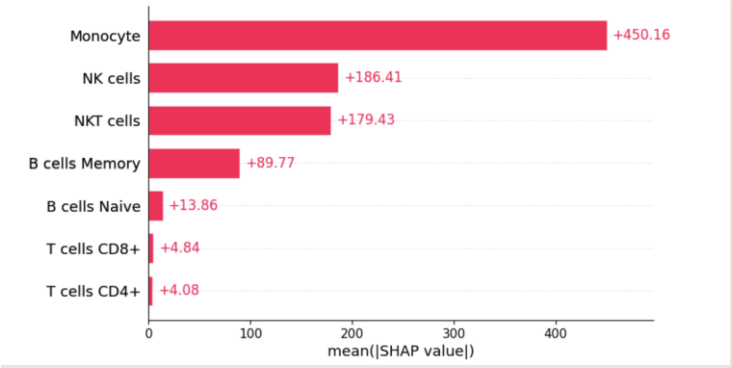


Figure 9: Distribution of mean absolute of SHAP values for each immune cell type.

Commented [RA42]: 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

Discussion

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. 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
11. 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.
12. 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

13. 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

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

	Min.	0
	25%	0
	50%	0.010
	75%	0.093
Immune		
	Mean	0
	Std	< 0.001
	Min.	0
	25%	0
	50%	< 0.001
	75%	< 0.001
BA		
	Mean	0.193
	Std	0.086
	Min.	0.014
	25%	0.133
	50%	0.189
	75%	0.252

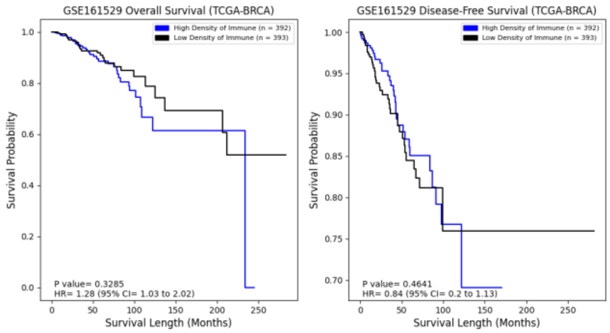
TCGA BRCA (1,111 patients) **(source** GSE176078 – 26 patients)		
AV		
	Mean	0.003
	Std	0.025
	Min.	0
	25%	0
	50%	0
	75%	0
	Max.	0.366
Fibroblast		
	Mean	0.057
	Std	0.063
	Min.	0
	25%	0.007
	50%	0.038
	75%	0.086
	Max.	0.533
HS		
	Mean	0

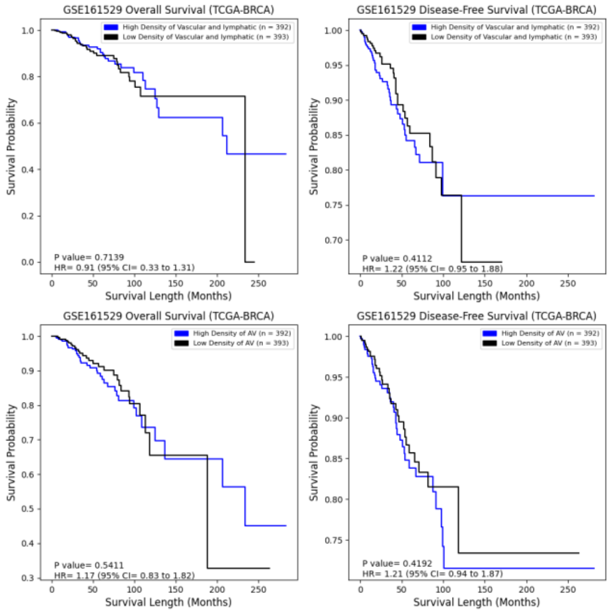
	Std	0.003
	Min.	0
	25%	0
	50%	< 0.001
	75%	0
	Max.	0.091
Vascular and lymphatic		
	Mean	0.515
	Std	0.146
	Min.	0
	25%	0.427
	50%	0.527
	75%	0.625
	Max.	0.840
Immune		
	Mean	0.405
	Std	0.156
	Min.	0
	25%	0.298
	50%	0.400
	75%	0.5106
	Max.	0.971
BA		
	Mean	0
	Std	0.057
	Min.	0
	25%	0
	50%	0.017
	75%	0
	Max.	0.462

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

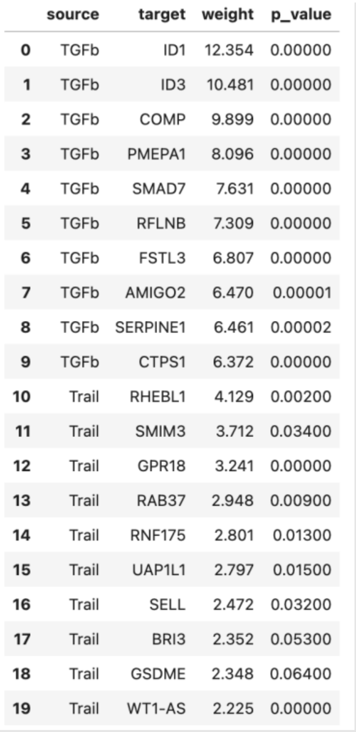
	AV	FIBROBLAST	HS	VASCULAR AND LYMPHATIC	IMMUNE	BA
NORMAL	↑	↓	↑	↓	↓	↑
TUMOR	↓	↑	↓	↑	↑	↓

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



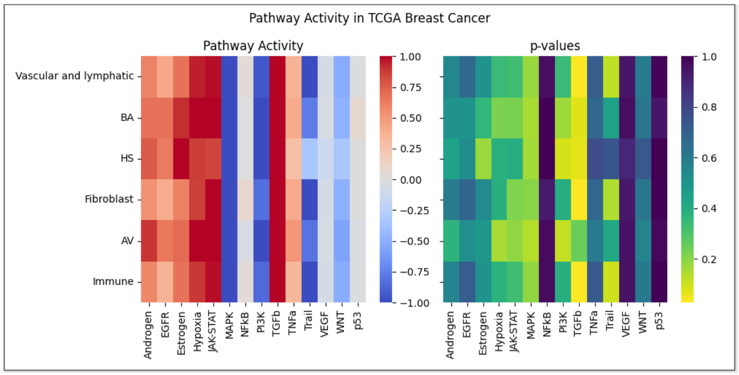


Appx. Fig.1: Analysis of immune cells (top), vascular and lymphatic alveolar cells (bottom) on the patient's overall survival and disease-free survival.



	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.