## 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 make 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 the dataset from "A Human Breast Atlas Integrating Single-Cell Proteomics and Transcriptomics" which identifies cells related to breast cancer at both the transcriptomic and proteomic levels such as mammary epithelial cell (MEC); (alveolar (AV), Hormone Sensing (HS) and basal (BA)); and stromal cells (fibroblasts, vascular/lymphatic cells, and immune cells).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.

We identified cell-type-specific gene expression signatures associated with overall survival in BRCA patients. Specifically, we found that the 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 can provide insights into the molecular mechanisms underlying cancer development and progression and 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, although other cell types may also be present. However, persistent activation of the immune system and failure of the inflammatory response to resolve can lead to chronic inflammation, which promotes tumor growth.

The intricate interplay between tumor and immune cells in the microenvironment leads to the production of a wide variety of cytokines and growth factors that foster tumor cell proliferation, survival, and metastasis. The complex nature of this communication highlights the significant impact that immune cells have on the tumor microenvironment, with both pro-tumoral and anti-cancer roles.

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 [2,3]. Ongoing research is focused on the role of infiltrating lymphocytes and other immune cells in the tumor microenvironment.

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 [6, 8].

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 provide an estimation of 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 tumor microenvironment can have a significant impact on patient outcomes. Therefore, identifying the cell-type-specific molecular mechanisms underlying the development and progression of BRCA tumors is essential for the development of effective diagnostic and therapeutic strategies.

The molecular subtype of breast cancer is based on the genes the cancer cells express.  And there are five main molecular subtypes of invasive breast cancer:

- Luminal A breast cancer is estrogen receptor-positive and progesterone receptor-positive, HER2 negative, and has low levels of the protein Ki-67.
- Luminal B breast cancer is estrogen receptor-positive and HER2-negative, and either has high levels of Ki-67 or is progesterone receptor-negative
- HER2-enriched breast cancer is estrogen receptor-negative and progesterone receptor-negative and HER2-postive.
- Triple-negative or basal-like breast cancer lacks estrogen and progesterone receptors, as well HER2 expression, and is more prevalent in individuals with a BRCA1 mutation, and the most aggressive subtype.

## **Methods**

The population data for this study was sourced from the Cancer Genome Atlas (TCGA) project, which was 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. Specifically, the TCGA project collected data on breast invasive carcinoma (BRCA), which 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 patients. RNA sequence data of various types is available, but for this study, we downloaded and used the "Primary" and "Solid Tissue Normal" data, as MuSiC performs its own normalization and *unstranded* data was only considered. The median age of the cohort was 58 years, and most patients were white (75.6%). The two most common types of cancer 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. 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.

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

- "A single-cell and spatially resolved atlas of human breast cancers" (GSE177078) was to provide a more detailed understanding of the cellular and molecular heterogeneity within breast tumors. The researchers of this study performed scRNA-Seq (Chromium, 10X Genomics) on 26 primary tumors from three major subtypes of breast cancer (11 ER+, 5 HER2+, and 10 TNBC). The study identified 9 major cell types, 29 minor cell types and 49 cell subtypes (Table 2, GSE176078_cell_hierarchy.html). The study found that LAMs and CXCL10hi macrophages are a key source of immunosuppressive molecules within the human breast tumor microenvironment (TME), and spatial analysis revealed their proximity to PD-1+ lymphocytes. Furthermore, 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 study "A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast " (GSE161529) presents an extensive single-cell transcriptome map of over 430,000 cells (Table 3, GSE161529_cell_hierarchy.html), from 52 patients; covering normal breast tissue at various hormonal stages, preneoplastic BRCA1+/- tissue, different subtypes of breast cancer (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.

We then conducted an investigation into 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 and performed Kaplan-Meier survival analyses with a log-rank test using a Cox's proportional hazard model from the Python package lifeline. We then computed the Hazard Ratio with a 95% Confidence Interval and corresponding p-values, and generated Kaplan-Meier curves using the Python scikit-survival package Kaplan Meier Estimator. 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, using estimated cell proportions we use PROGENy, an R package designed for gene set enrichment analysis (GSEA) and pathway analysis of gene expression data and the decoupler Python package. This Python package utilizes statistical methods such as Weighed Sum (WMEAN) or Univariate Linear Model (ULM) and prior knowledge about gene regulatory networks to predict the activity of transcription factors and pathways within a population samples.

## Results

### Deconvolution of Immune Cells From RNA-Seq Data

Using MUSiC as the algorithm for single-cell deconvolution, we were able to estimate the proportions of different immune cell subpopulations within each patient's tumor. Our analysis of the TCGA BRCA cohort revealed that normal patients had high proportions of HS, AV, and vascular and lymphatic cells, which were consistent with those found in normal TCGA patients (normal_cell_type_proportions_patients, Tcga_GSE161529_cell_type_normal_proportions_patients). In comparison, BRCA patients had a higher presence of immune cells (tumor_cell_type_proportions_patients), which were even more significantly present in breast cancer TCGA patients (Tcga_GSE161529_cell_type_tumor_proportions_patients, Tcga_GSE161529_cell_type_tumor_proportions_patients_violin).

Regarding cell subtypes, VL2 vascular endothelial cells and I3 T cells were predominant in both GSE161529 and TCGA cancer patients (tumor_cell_subtype_proportions_patients, Tcga_GSE161529_cell_subtype_tumor_proportions_patients), although their median levels were lower compared to other cell subtypes such as Has, HSx, and BL cell subtypes (Tcga_GSE161529_cell_subtype_tumor_proportions_patients_violin).

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.

In GSE17078 patients, a significant presence of immune cells, such as macrophages, monocytes, and T cells (CD4+ and CD8+), was observed, with similar proportions found in TCGA breast cancer patients (“With Others” plots). However, upon excluding macrophages and non-immune cells, we observed a substantial presence of NK and NKT cells, as well as B cells (memory B cells) to a lesser extent (“With_no_Other” plots). These findings highlight the complex and diverse nature of the immune cell composition within breast cancer tumors.

## Cell Fractions Clinical Outcome Correlation

Our findings suggest a significant correlation between high proportions of B cells Memory and improved overall survival (OS) and disease-free survival (DFS) outcomes, as well as between T cells CD8+ and NKT cells and improved DFS outcomes, when compared to patients with lower proportions of the same cell types. Interestingly, we did not observe the same trends for the immune cells when using the GSE161529 dataset and corresponding single-cell bulk-RNA deconvolution. Additionally, we found that vascular and lymphatic cells, as well as alveolar cells, did not significantly impact survival outcomes. 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.

## 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 also demonstrated involvement in the MAPK pathway, which is known to promote cell growth and proliferation. Each gene in PROGENy pathway has a weight corresponding to its up-regulation expression within a given pathway. Sorting these genes by weights we found that ID1, ID3, COM, PMEPA1, SMAD7 in the TGFb pathway and DUSP6, SPRY4, SPRY2, FOSL1, MMP1 in the MAPK pathway are prognostic markers in different cancers. Similar correlations were found with Vascular and lymphatic, BA, HS, Fibroblast and AV cells.

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.