Spatial Profiling of Tumor Microenvironments in Breast Cancer

Master thesis

Niek Brouwer

Delft University of Technology

TU Delft

Spatial Profiling of Tumor Microenvironments in Breast Cancer

by

Niek Brouwer

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday September 4, 2023 at 10:00 AM.

Programme:	Computer Science, Bioinformatics specialization	
Faculty:	Electrical Engineering, Mathematics and Computer Science	
Student number:	4567293	
Project duration:	February 6, 2023 – September 4, 2023	
Thesis committee:	Prof. Dr. Lodewyk F.A. Wessels	TU Delft & NKI Amsterdam, supervisor
	Dr. Daniël J. Vis	NKI Amsterdam, supervisor
	Dr. Joana S. de Pinho Gonçalves	TU Delft
	Dr. Cynthia C.S. Liem	TU Delft

TU Delft

Preface

This thesis report provides an overview of the research conducted over seven months as the concluding component of my master's programme in computer science at the Delft University of Technology. The project was carried out at the Netherlands Cancer Institute (NKI), a leading cancer research institute in Amsterdam. During the project, we studied the spatial distribution of cells in breast tumor tissue to identify patterns that have associations with tumor behavior. To do so, a novel approach was used that has recently been developed by researchers at NKI of which Alberto Gil-Jimenez has been the main contributor. The application of the method in a new setting and with new objectives has contributed to the validation and enhancement of this cutting-edge spatial analysis method.

From the very beginning of this project, I have felt welcome at the NKI and it has been motivating to work in such an innovative environment. I want to thank Dr. Daniël Vis for the daily supervision of this project which has pushed it to an ever-improving form, dragging me along with it. His involvement has always guaranteed clarity and is one of the reasons that this project progressed smoothly. Finally, I want to thank Prof. Dr. Lodewyk Wessels who offered me the opportunity to perform this research at the NKI and whose contribution and guidance I have always valued highly.

Niek Brouwer
Delft, August 2023

Summary

The tumor composition of breast cancer determines how tumors behave. Yet, there is a limited understanding of the arrangement of tumor cells in relation to cells in the tumor microenvironment (TME). In this research, we have characterized distance relationships between 324 cell-type pairs in 749 tissue samples of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study using Weibull distribution estimations, summarizing comprehensive spatial relationships with two parameters. The research showcased the first application of the method to a dataset of this substantial size and a dataset acquired with imaging mass cytometry. We identified distinct spatial relationships among breast cancer subtypes, particularly for basal, HER2-enriched, and luminal A tumors. The spatial relationships indicate attractive and repulsive interactions between different cell types and define cellular arrangements regardless of cellular abundance.

Moreover, several spatial relationships had significant associations with survival outcomes. Both findings could improve patient stratification and prognosis and emphasize the wealth of information that spatial analyses can retrieve. The results also confirm that Weibull distribution estimations are a suitable and effective method to summarize distance distributions. The application to other cohorts could lead to new insights into the tumor composition of different cancer types. Finally, the spatial profiling method was used to characterize neighborhoods and revealed distinct spatial relationships consistent with neighborhood characteristics but also provided new hallmarks.

Contents

Preface i

Summary ii

Nomenclature v

1 Introduction 1
1.1 Research Objectives . . . 2
1.2 Research Questions . . . 2
1.3 Thesis Outline . . . 2

2 Background 3
2.1 Multiplex Tissue Imaging . . . 3
2.1.1 Microscopy-based methods . . . 3
2.1.2 Mass-spectrometry-based methods . . . 4
2.2 Cell-type Classification . . . 4
2.2.1 Semi-supervised cell-type classification . . . 5
2.2.2 Unsupervised cell-type classification . . . 5
2.3 Spatial Profiling Techniques . . . 5
2.3.1 Cellular interactions . . . 5
2.3.2 Cellular neighborhoods . . . 6
2.3.3 Distance relationships . . . 6

3 Methodology 8
3.1 Dataset Overview . . . 8
3.2 Cell-type Classification . . . 8
3.3 Cell-type Quantification . . . 9
3.4 Spatial Relationship Analysis . . . 10
3.5 Feature Associations . . . 12
3.5.1 Recurrent cell types in spatial relationships . . . 12
3.5.2 Effect of cell-type fractions on spatial relationships . . . 12
3.6 Subtype Predictions . . . 13
3.7 Survival Predictions . . . 13
3.8 Neighborhood Identification . . . 13
3.9 Region Segmentation . . . 14

4 Results 15
4.1 Cellular Composition of Breast Cancer Subtypes . . . 15
4.1.1 Epithelial cell types . . . 15
4.1.2 TME cell types . . . 17
4.1.3 Section Summary . . . 18
4.2 Spatial Relationships between Cell Types . . . 20
4.2.1 Parameter estimations . . . 20
4.2.2 Comparison to spatial analysis of NABUCCO trial . . . 23
4.2.3 Section Summary . . . 24
4.3 Predicting Breast Cancer Subtypes . . . 25
4.3.1 Predictions based on Danenberg and alternative cell-type fractions . . . 25
4.3.2 Predictions based on epithelial cell-type fractions . . . 25
4.3.3 Predictions based On TME cell-type fractions . . . 26
4.3.4 Predictions based on spatial relationships . . . 27
4.3.5 Predictions based on all features . . . 28
4.3.6 Section Summary . . . 29

4.4 Spatial Characterization of Breast Cancer Subtypes . . . 30
4.4.1 Spatial characterization of basal tumors . . . 31
4.4.2 Spatial characterization of HER2-enriched tumors . . . 35
4.4.3 Spatial characterization of luminal A tumors . . . 38
4.4.4 Spatial characterization of luminal B tumors . . . 41
4.4.5 Spatial characterization of normal-like tumors . . . 42
4.4.6 Section Summary . . . 44
4.5 Predicting Patient Survival . . . 46
4.5.1 Predicting patient survival . . . 46
4.5.2 Predicting patient survival per ER status . . . 48
4.5.3 Section Summary . . . 51
4.6 Neighborhoods and Spatial Relationships . . . 52
4.6.1 Spatial relationships in vascular stroma . . . 53
4.6.2 Spatial relationships in active stroma . . . 55
4.6.3 Spatial relationships in $CD8^+$ and macrophage neighborhoods . . . 57
4.6.4 Spatial relationships in TLS-like neighborhoods . . . 58
4.6.5 Spatial relationships in APC-enriched neighborhoods . . . 60
4.6.6 Section Summary . . . 61

5 Discussion **62**
5.1 Conclusions . . . 62
5.2 Limitations . . . 64
5.3 Recommendations and Future Work . . . 64

References **66**

A Protein Panel **68**

B Cell-type Descriptions **69**

C Weibull Parameter Estimations NABUCCO trial **70**

Nomenclature

Abbreviations

Abbreviation	Definition
PAM50	Prediction Analysis of Microarray using 50 classifier genes
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
CK	Cytokeratin
ER	Estrogen Receptor
PR	Progesterone Receptor
HER	Human Epidermal growth factor Receptor 2
TME	Tumor Microenvironment
TIL	Tumor-Infiltrating Lymphocyte
TLS	Tertiary Lymphoid Structure
mIF	Multiplexed ImmunoFluorescence
cyTOF	Cytometry by Time-Of-Flight
t-SNE	t-distributed Stochastic Neighbor Embedding
SOM	Self-Organizing Map
1-NN	First Nearest Neighbor
PDF	Probability Density Function
MLE	Maximum Likelihood Estimation
NLME	Non-Linear Mixed Effect
LASSO	Least Absolute Shrinkage and Selection Operator
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
FPR	False Positive Rate

1

Introduction

Breast cancer is a global health challenge. It ranks as the second most commonly diagnosed cancer type and is the second leading cause of cancer mortality among women worldwide, accounting for 682,000 deaths in 2020 globally (Sung et al., 2021). Geographical variations in incidence rates are evident, with economically developed nations displaying disproportionately high diagnosis and mortality rates. However, developing countries in South America, Africa, and Asia are experiencing a rapid increase in breast cancer incidence. If current trends continue, it is projected that by 2040, the global burden of breast cancer will exceed 3 million new cases, primarily due to population growth and aging (Arnold et al., 2022). In the past decades, extensive research has led to the development of effective breast cancer treatments that significantly increased survival rates. Nevertheless, besides its evident impact on the lives of patients and their surroundings, breast cancer has the highest treatment cost of any cancer, accounting for 13 % of all cancer treatment costs in the Netherlands (RIVM, 2019). Understanding the behavior of breast cancer is essential to fight the increasing incidence rate and to identify new targeted treatments that treat the disease more effectively.

In clinical settings, breast cancer' aggressiveness and stage are assessed from features such as histopathology, tumor size, and nodal status. Breast cancer tumorigenesis is driven by diverse underlying genetic alterations and biological events, which pose challenges to accurately predicting tumor behavior solely based on conventional clinical parameters alone (Jackson et al., 2020). Recent studies aim to refine breast cancer classification by using high-throughput studies that expose the genetic variation of tumors. As a result, PAM50 (Prediction Analysis of Microarray using 50 classifier genes) has been developed as a standardized method to classify tumors into subtypes luminal A, luminal B, HER2-enriched, basal, and normal-like (Parker et al., 2009). The subtypes are associated with distinct survival outcomes and treatment sensitivities. Additionally, the molecular characterization of tumor cells has provided opportunities for developing new treatments targeting molecular aberrations driving individual tumor growth.

Tumor cells do not operate in isolation and engage in many interactions with the tumor microenvironment (TME) consisting of the cells surrounding tumors. The TME is a complex composition of fibroblasts, immune, cancer, and endothelial cells. It plays a crucial role in tumor development, as it engages in reciprocal interactions with tumor cells by secretion of proteins, cytokines, and growth factors. These interactions, combined with inherent genetic alterations in tumor cells, determine the tumor's growth characteristics, morphology, and invasiveness (Mittal et al., 2018).

Crucial tumor and TME cell interactions have been discovered thanks to technological developments in multiplex imaging and omics techniques (Elhanani et al., 2023). Omics methods measure the expression of genes (transcriptomics), proteins (proteomics), or small metabolites (metabolomics), revealing the extensive diversity of tumors and TME cells. The measurements are retrieved from tumor samples while preserving the tissue organization. Many recurrent spatial relationships have been identified from spatially-resolved tumor tissue, which has helped to understand the mechanisms that explain unique aspects of tumor behavior. The assessment of distances between tumor and immune cells has led to the discovery of associations between high levels of tumor-infiltrating lymphocytes (TILs) and better survival outcomes (Savas et al., 2016). In addition to distance relationships, spatial analyses have revealed specialized multicellular neighborhoods in the TME linked to distinct tumor phenotypes.

The presence of immune cells in tertiary lymphoid structures (TLSs), for example, is associated with enhanced immune response and improved survival rates (Cabrita et al., 2020).

The identification of significant spatial relationships in breast cancer is still limited. Spatial analyses require datasets large enough to account for the characteristic heterogeneity of breast cancer. A pioneering study by Danenberg et al., 2022 examined 749 biopsy samples of patients recruited to the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohort, identifying recurrent TME neighborhoods with distinct enrichment patterns among breast cancer subtypes and associations with survival outcomes. In this research, we use the same dataset to investigate the molecular properties and distance relationships of tumor and TME cells. The study aims to provide a systematical characterization of breast cancer tumors and complement the results of Danenberg et al. by identifying unique spatial properties of neighborhoods.

1.1. Research Objectives

The comprehensive analysis of the composition of breast cancer tumors aims to find associations between cellular diversity, spatial relationships, and tumor behavior. To achieve this goal, we have the following objectives:

1. Assess the diversity and abundance of cell types in breast cancer tumors and TMEs.
2. Quantify spatial relationships between cell types.
3. Identify associations between cell type abundance, spatial relationships, and breast cancer subtypes.
4. Identify associations between cell type abundance, spatial relationships, and patient survival.
5. Identify associations between cell type abundance, spatial relationships, and cellular neighborhoods.

1.2. Research Questions

The research objectives are achieved by answering the following research questions:

1. What cell types can we distinguish from protein expression profiles in breast cancer tumors and TMEs?
2. What are the spatial relationships of cell types throughout breast cancer tumors and TMEs?
3. What cell types and spatial relationships are associated with breast cancer subtypes?
4. What cell types and spatial relationships are associated with patient survival?
5. What cell types and spatial relationships are associated with cellular neighborhoods?

1.3. Thesis Outline

The remainder of this report is structured as follows. Chapter 2 aims to assist readers in comprehending the research's context by providing an overview of relevant prior work. It explains various imaging techniques and classification methods employed to acquire datasets suitable for spatial profiling. Additionally, the chapter outlines state-of-the-art methodologies for characterizing spatial arrangements. Chapter 3 provides an overview of the methods employed in this study. It describes how data was acquired and preprocessed, and how spatial features were extracted. It also explains the methods used to identify associations between spatial arrangements and tumor phenotypes. Chapter 4 details the analyses' outcomes. It begins by describing the diversity of cell types within breast cancer subtypes. Subsequent sections present and interpret the spatial relationships among these cell types. Associations between spatial relationships, tumor subtypes, and survival probability are then demonstrated. The final section applies spatial analyses to TME neighborhoods, offering insights into the spatial relationships within these recurring structures. Finally, Chapter 5 discusses this research' key insights and contributions, offering recommendations and suggestions for future work.

2

Background

This chapter explains relevant existing methods and previous studies. First, it briefly describes multiplex imaging techniques, explaining how images of tumor tissues are obtained. The explanation aids in understanding the construction of the dataset used in this research. Secondly, it describes single-cell classification methods to clarify how cell types are obtained and how the stratification can be obtained. Finally, this chapter provides an overview of spatial profiling techniques that describe the spatial properties of tumor tissues. It discusses the results of earlier spatial characterization of breast cancer and highlights parts of the organization that have not been studied thoroughly yet.

2.1. Multiplex Tissue Imaging

In the past decades, many new techniques have been developed to measure the genome, transcriptome, proteome, and metabolome of single cells. Antibody-based multiplex imaging techniques are widely used to measure proteins in situ while retaining the spatial distribution of cells. Antibodies are molecules that bind to very specific proteins. Due to this great precision, marked antibodies reveal the presence of distinct proteins in tissue slides. Antibodies are tagged with fluorophores, enzymes, DNA oligos, or metals. Multiplex imaging techniques that detect these markers are broadly categorized into microscopy-based and mass spectrometry-based methods. The resulting images provide a wealth of information but require dedicated analyses to reveal cell types, cellular morphology, interactions, and multicellular structures.

2.1.1. Microscopy-based methods

Through light microscopes, luminescent antibody markers are detected. However, the number of detectable tags is limited because of the spectral overlap. Cyclic microscopy methods perform an iterative process in which tissues are stained multiple times with different antibody panels. A tagged antibody is added, measured, and removed from the tissue in each cycle. An example of a cyclic method is multiplex immunofluorescence staining (mIF), which adds fluorophore tags to the tissue in up to six staining iterations (Vos et al., 2021) (Figure 1). Staining iterations are labor-intensive and damaging to the tissue slides whereby the number of cycles is restricted (Tsujikawa et al., 2017). The limited number of detectable tags is the biggest drawback of microscopy-based methods.

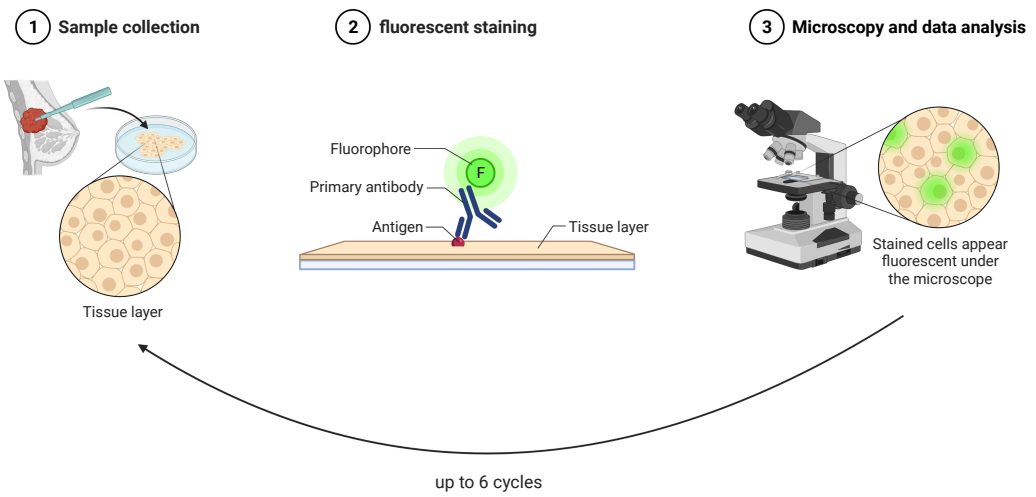

Figure 1: Overview of multiplex immunofluorescence protocol. Tumor biopsies are collected (1) and stained with fluorophore-tagged antibodies (2). The level of fluorescence is picked up by a light microscope and mapped back to the single cells (3). The procedure can be repeated up to six times. Created with BioRender.com.

2.1.2. Mass-spectrometry-based methods

Mass-spectrometry-based imaging techniques detect antibodies through unique metal isotope labels. The labeled antibodies are added to tissue slides in a single reaction and picked up by a mass spectrometer once bound. The tissue is divided into a fine grid, and the presence of isotopes is read pixel by pixel, restricting the tissue slides to smaller sizes compared to microscopy-based techniques. Moreover, mass-spectrometer-based quantification is destructive; each slide can only be used once. The methods differ in the way that isotopes are extracted and the type of mass spectrometer used. The dataset used in this research is generated with cytometry by time-of-flight (cyTOF) (Giesen et al., 2014). CyTOF measures the mass-to-charge ratio of up to 50 isotopes based on the time it travels through the detector (Figure 2).

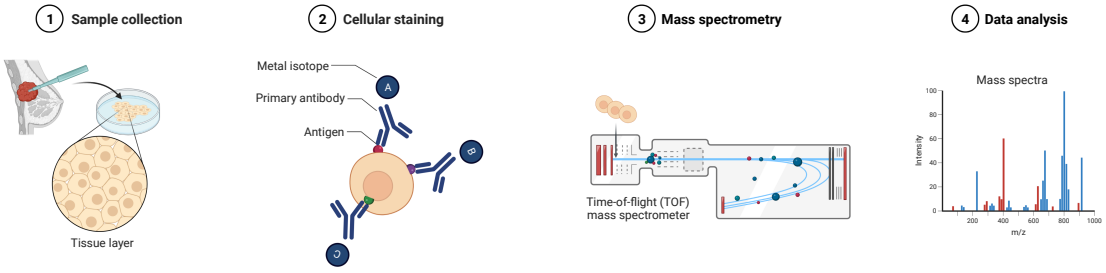

Figure 2: Overview multiplex mass spectrometry using a time-of-flight mass spectrometer (cyTOF). Tumor biopsies are collected (1) and stained with up to 50 metal-labeled antibodies (2). Cells are injected as droplets into an argon flow chamber and exposed to a plasma torch, vaporizing cells into an ionized cloud. Ions are separated by their mass-to-charge ratio (3). The protein expression levels are interpreted from resulting mass spectra and mapped back to single cells. Created with BioRender.com.

2.2. Cell-type Classification

With the expanding application of multiplex imaging techniques, many methods have been developed to cluster and classify single cells based on high-dimensional datasets. In general, there are two method types: semi-supervised and unsupervised. Previously, cell types were manually identified from plots of all protein combinations or low-dimensional representations of the datasets. Dimensionality reduction techniques such as t-SNE (t-distributed Stochastic Neighbor Embedding) reduce high-dimensional images to a 2-dimensional representation (Van der Maaten and Hinton, 2008).

2.2.1. Semi-supervised cell-type classification

Manual classification is labor-intensive and requires many plots to get correct annotations of cluster boundaries and cell types. Still, the resulting cell classification is very accurate and is used by semi-supervised methods as a ground truth. The predetermined manual labels are used to define unique cell type characteristics that are generalized to the remaining cells (Abdelaal et al., 2019).

2.2.2. Unsupervised cell-type classification

Unsupervised classification techniques identify and classify groups of cells with common modalities. Many different techniques exist and here we only describe the methods that were used by Danenberg et al. (2022).

Phenograph (Levine et al., 2015) identifies subpopulations in high-dimensional single-cell data by representing it as a graph in which cells are connected if they are phenotypically similar. Subsequently, the Louvain method (Blondel et al., 2008) is used for community detection. The Louvain algorithm iteratively adds nodes together to yield the largest increase in the overall modularity of the graph. Modularity is the fraction of edges within clusters minus the expected fraction of edges if they were distributed randomly. The resulting clusters are visualized with t-SNE.

FlowSOM (Van Gassen et al., 2015) is a method that identifies clusters while also providing a discretized representation of the original high-dimensional space by using a self-organizing map (SOM) (Kohonen, 1990). A SOM is an unsupervised machine-learning technique that learns a lower-dimensional representation for a high-dimensional input dataset. The representation projects data points in proximity if they have similar variable values. FlowSOM finds many more clusters than cell types, and the resulting groups are merged into recurrent cell types using manual and automatic approaches.

2.3. Spatial Profiling Techniques

Spatial profiling techniques describe cell' spatial arrangement, indicating how cells interact and cooperate. In general, three types of spatial relationships are distinguished in the distribution of cells: cellular interactions, cellular neighborhoods, and distance relationships.

2.3.1. Cellular interactions

Cell types that interact directly through the exchange of proteins and small molecules are often found as neighbors. Therefore, methods identifying enriched or depleted interactions quantify how often cells are adjacent. The neighbor counts are compared to permutation experiments, which represent the neighbor counts in random distributions of cells (Figure 3).

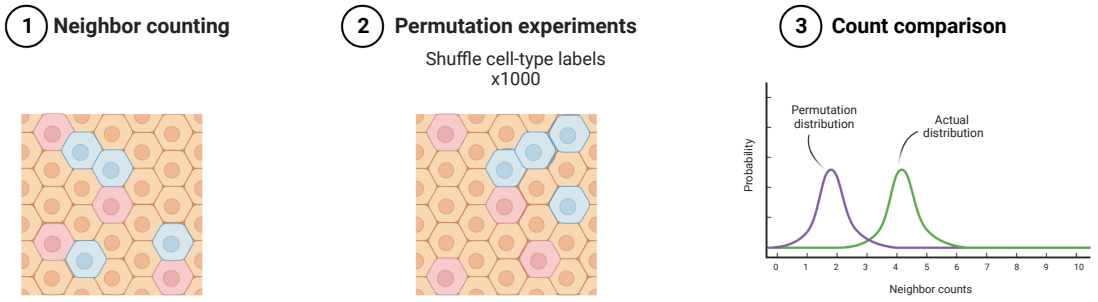

Figure 3: Interaction detection through permutation experiments. For each cell-type combination, the number of adjacent cells (1) is compared to the numbers in the same tissue slide where cell-type labels are randomly shuffled (2). The count distributions are compared with statistical tests (3). Created with BioRender.com.

Cellular interactions have been studied extensively in breast cancer. Similar to other cancer types, high levels of TILs in breast tumors are associated with improved survival outcomes. TILs are quantified by counting the number of lymphocytes near tumor cells (Cheung et al., 2021). Moreover, interaction detection analyses have found that tumor growth is suppressed if $CD8^+$ T cells are near antigen-binding

cells (Patwa et al., 2021). Finally, advanced tumors often contain high fractions of macrophages that are surrounded by proliferating and hypoxic tumor cells (Schapiro et al., 2017).

2.3.2. Cellular neighborhoods

Cellular neighborhoods are regions of more than two cells with recurrent compositions or spatial characteristics. In literature, neighborhoods are also called niches, microenvironments, compartments, or communities. One approach to detect neighborhoods is by using a sliding window to separate the tissue into regions systematically. The cell-type compositions of each window are used to cluster regions and to identify groups with recurrent properties (Schürch et al., 2020). Alternatively, community-detection algorithms are used to detect dense cell regions from a graph representation of the tissue (Blondel et al., 2008, Toth et al., 2022). Similar to the sliding-window approach, the detected communities are characterized and clustered to find dense cell groups with recurrent natures (Figure 4).

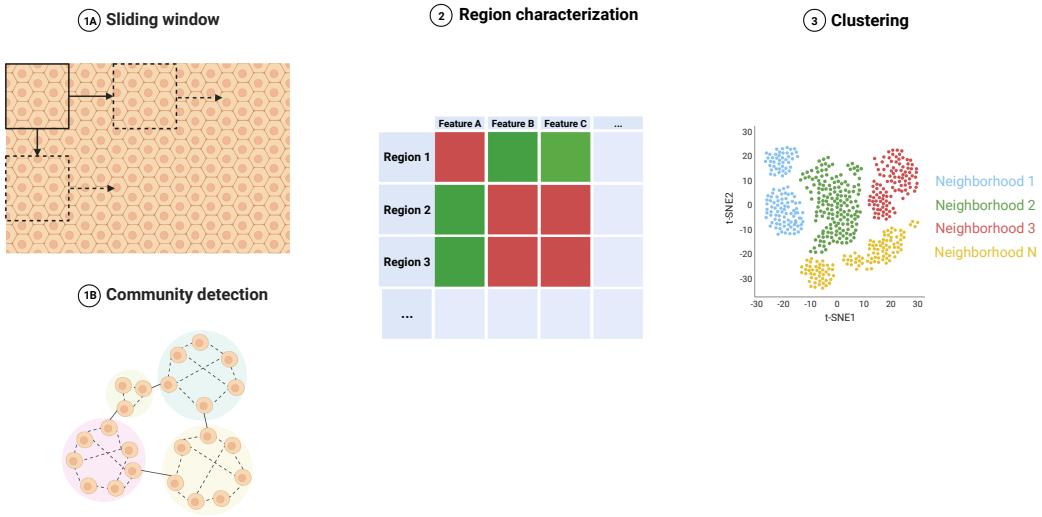

Figure 4: Overview of neighborhood detection methods. Regions are either systematically separated with a sliding window (1A) or identified with community-detection algorithms (1A). The resulting regions are characterized by phenotypic and spatial properties (2) and recurrent neighborhoods are identified by clustering regions based on these properties. Created with BioRender.com.

Neighborhood detection methods have been applied to the METABRIC cohort and have identified multiple recurrent neighborhoods that are associated with survival outcomes (Ali et al., 2020, Danenberg et al., 2022). In these studies, neighborhoods were characterized by proportions and vertex degrees of cell types. Although the properties provide some information about the cellular composition of neighborhoods, a comprehensive understanding of their function and effect on tumor behavior still needs to be provided. In this research, the characterization of neighborhoods is extended by providing a systematic approach to profile the cell-type content and distance relationships between cells.

2.3.3. Distance relationships

Distance relationships between cell types are measured with first nearest neighbor (1-NN) distances. The 1-NN distance from a reference cell to a target cell type is the length to its closest target cell. The 1-NN distance distribution is made up of the 1-NN distances from all reference cells to a target cell type. Cell types that are tightly packed are represented by a distance distribution with a sharp peak, while segregated cell types produce a 1-NN distance distribution with a larger mean and variance. Distance distributions are summarized with metrics such as the median (Parra et al., 2021) or mean (Ma et al., 2022) to represent spatial relationships with a single number (Figure 5). Both metrics reduce distance distributions to a single value that does not account for aspects of the distribution such as variances and amplitudes. The fact that distance distributions with distinct behaviors are often represented with the same value is a major shortcoming of these metrics.

An alternative approach to summarize distance distributions is by fitting a Weibull distribution. The

behavior of the Weibull distribution depends on the shape (k) and scale (λ) parameter. The Weibull distribution fits a wide range of distribution behaviors by altering the parameters and distributions are uniquely represented by parameter combination. The following formula gives the probability density function (PDF) of the Weibull distribution:

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{2.1}$$

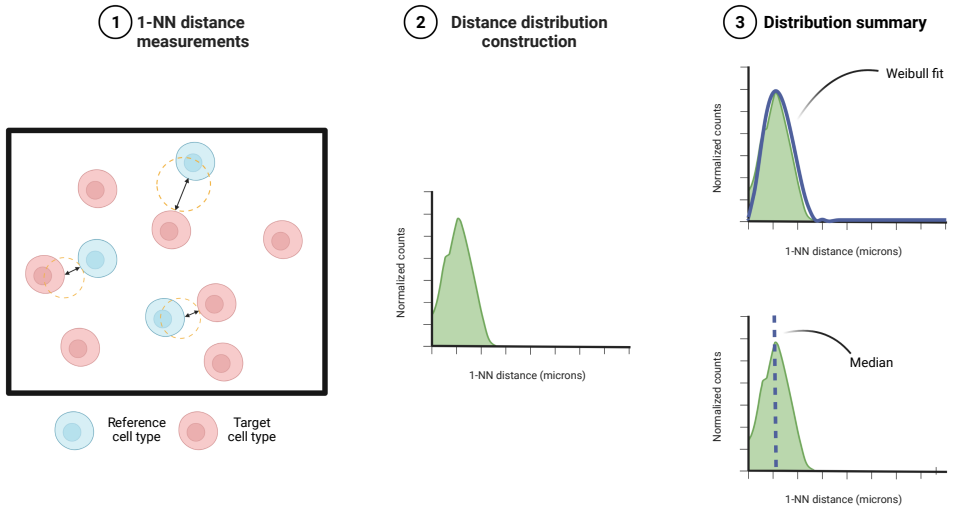

Figure 5: Overview of distance relationship analysis. The 1-NN distances from a reference cell type to a target cell type are the lengths between all reference cells and their closest target cells (1). The 1-NN distances are plotted as a distribution (2) and summarized using metrics such as medians or by fitting a Weibull distribution to its behavior. The Weibull distribution is uniquely described with the shape and scale parameters. Created with BioRender.com.

Few studies have investigated distance relationships of cell types in breast cancer tissue. The Weibull estimation method is a new approach, and the results of the first application suggest that Weibull parameters efficiently represent behaviors of 1-NN distance distributions. The study identified significant associations between distance relationships and treatment response of urothelial cancer by comparing the Weibull parameters among 24 patients. The dataset used in this study was small and acquired with mIF. The effectiveness and applicability in larger datasets with more comprehensive cell types are yet to be determined. In this research, we investigate whether the method can identify associations with breast cancer subtypes and survival from a larger dataset obtained with imaging mass cytometry.

3

Methodology

This chapter describes all methods that were used to reach the research objectives. It begins by briefly describing the data acquisition and preprocessing steps used by Danenberg et al. (2022). Subsequently, the chapter describes the quantification of cell-type abundance and spatial relationships within the dataset. Furthermore, it explains all statistical tests used to measure associations between the retrieved features, breast cancer subtypes, and survival probabilities. Finally, it illustrates the identification of TME neighborhoods and describes subsequent spatial analyses. By comprehensively describing the methods used in this research, this chapter aims to provide a clear and systematic foundation for the subsequent interpretations of the results.

3.1. Dataset Overview

The dataset used in this research consists of 749 fresh frozen breast tissue biopsies originating from 693 patients recruited to the METABRIC cohort. Clinical data in the public domain was linked to the samples (Rueda et al., 2019). METABRIC assessed the hormone receptor status of 681 patients from the expression of ER, PR, and HER2 measured with immunohistochemical analysis. Additionally, 638 patients were classified into PAM50 subtypes. Survival data is available for 565 patients.

For each tissue slide, protein expression profiles of single cells were constructed using CyTOF. Cells were stained with 37 metal-labeled antibodies (Appendix A) and metal labels of antibodies bound to the interior and surface of single cells were picked up by a time-of-flight mass spectrometer. The metal counts (corresponding to the bound antibody abundance) were reported in 37 image layers.

The cell-type classification combined an automated approach with the manual curation of a pathologist. First, cells were classified into epithelial and non-epithelial by fitting two complementary methods. A two-component Gaussian mixture was fitted to the pan-cytokeratin (pan-CK) counts, and a region classifier was trained on all cytokeratin counts (pan-CK, CK8-18, and CK5). A pathologist evaluated the predictions of both classifiers using cell morphology and expression of cytokeratins as a guide. After the segmentation of epithelial and non-epithelial, single cells in both regions were clustered based on protein expression profiles using FlowSOM and Phenograph. Both methods were explained in the previous chapter. Clusters were manually merged and labeled in 32 cell types (16 epithelial and 16 TME) based on similar median protein expression values (Figure 6).

3.2. Cell-type Classification

An alternative cell-type classification was developed by manually merging cell types with similar median protein expression values. The 32 Danenberg cell types were reduced to 18 alternative cell types (8 epithelial, 10 TME) (Figure 6). To guide the interpretation of significant findings that are later described in this report, a brief overview of the functionalities of different cell types is provided in Appendix B.

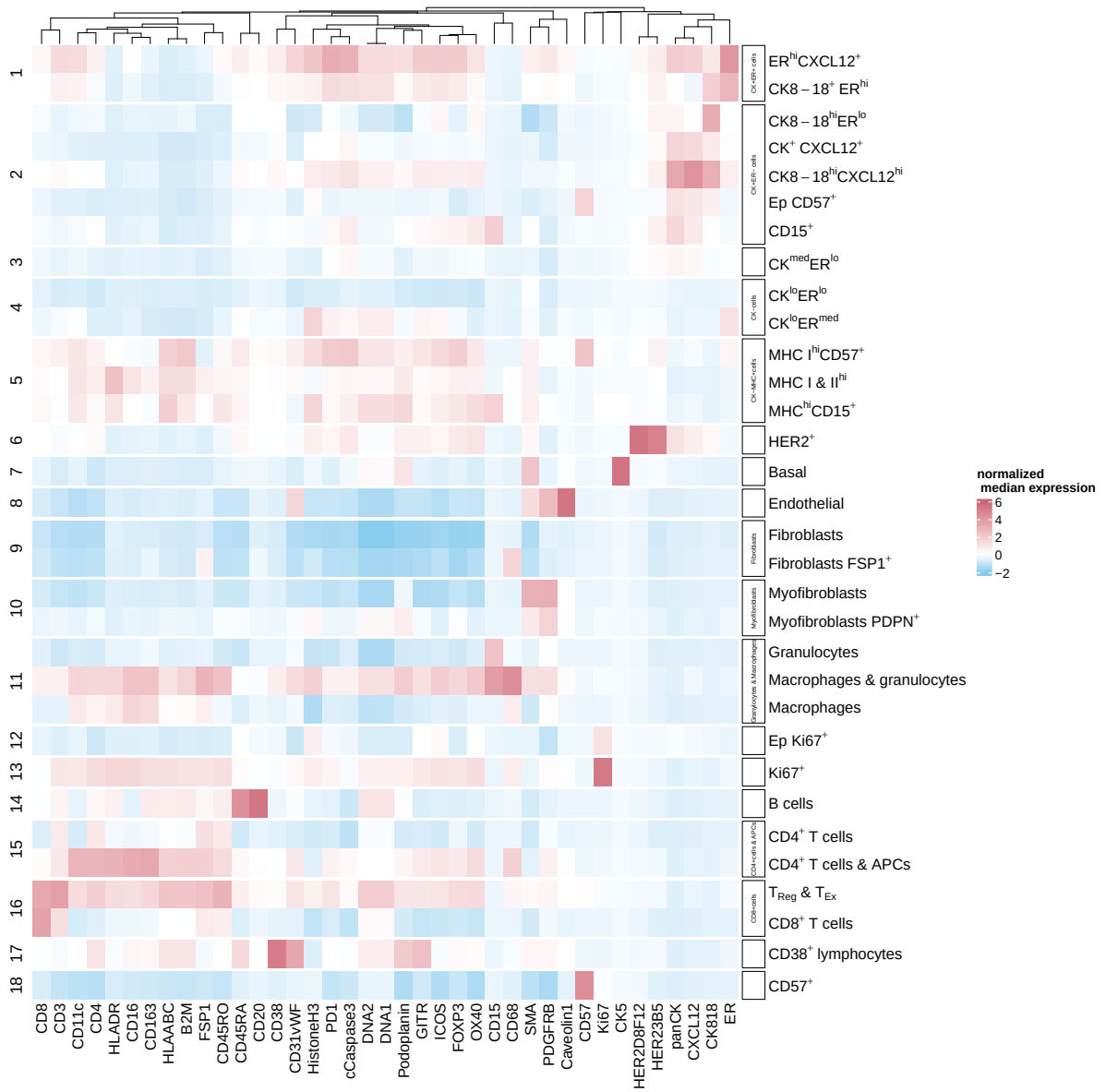

Figure 6: Median protein expression profiles of Danenberg cell types. Boxes indicate cell types that were merged and mention the new labels if applicable.

3.3. Cell-type Quantification

Each tissue slide's abundance of cell types was quantified with normalized fractions. First, the total epithelial and TME cell counts were computed. Cell-type fractions were calculated with the following formula:

$$Fraction_i = \begin{cases} count_i / count_{epithelial} & \text{if } i \in epithelial\ types \\ count_i / count_{TME} & \text{if } i \in TME\ types \end{cases} \tag{3.1}$$

Cell-type densities are an alternative approach to describe the number of cells in a tissue slide. Densities are defined as the cell-type count per unit area and computed with the following formula:

$$Density_i = \begin{cases} count_i / area_{epithelial} & \text{if } i \in epithelial\ types \\ count_i / area_{TME} & \text{if } i \in TME\ types \end{cases} \tag{3.2}$$

Cell-type fractions and densities largely correlate (Figure 7) suggesting that both metrics capture the same information. Outlier points correspond to TME cell types with low counts in samples with a

very small TME region whereby the number of cells per square millimeter of TME is excessively large. In further analyses, only cell-type fractions were used to quantify abundance, which is consistent with the quantification used in Danenberg et al. (2022) research.

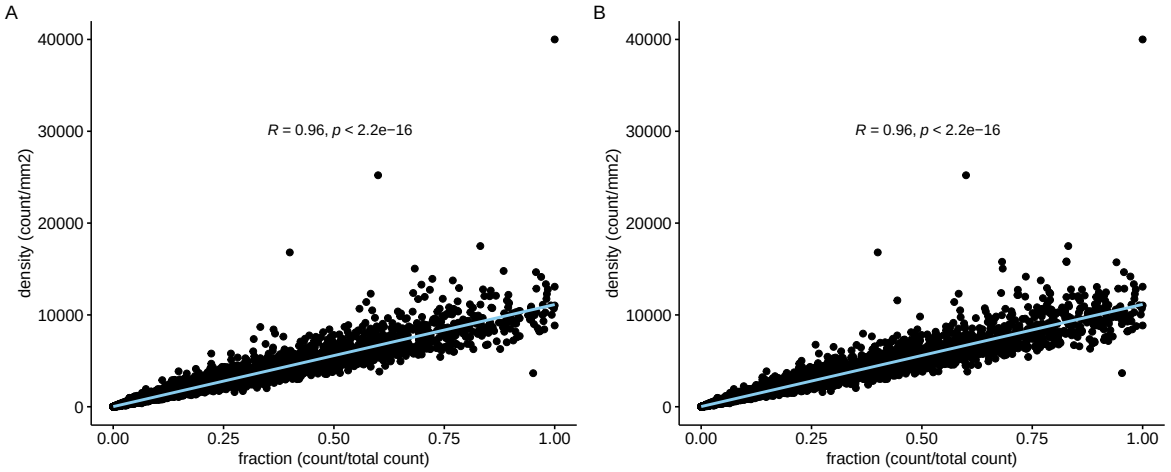

Figure 7: Fractions and densities of Danenberg cell types (A) and alternative cell types (B). The correlation between the two features is measured with the Pearson correlation coefficient. The p-value is the probability that the two variables are not correlated.

3.4. Spatial Relationship Analysis

The spatial relationship between a reference and target cell type is defined here as the distribution of 1-NN distances from the reference cell type to the target cell type. The spatial relationship from a cell type to itself is called a self-self relationship. Cell-type combinations with fewer than 20 reference cells and 5 target cells were not considered because the distance distributions of these combinations are too sensitive to the effect of single cells.

The 1-NN distances were measured using *nndist* of the *spatstat v3.0-5* package. All distances were plotted as a histogram which was smoothened by a 5-micron sliding window summing the frequencies within the window for each micron. Finally, the histogram was normalized so that the area under the curve (AUC) equaled 1.

The 1-NN distance histograms were summarized by fitting a Weibull distribution to the PDF. Fitting the Weibull distribution was a two-step process consisting of an initialization and estimation step. First, an initial estimate of the distribution parameters for each image and each cell-type combination was obtained with maximum likelihood estimation (MLE) using *fitdist* of the *fitdistrplus v1.1-11* package. MLE finds the parameters of the Weibull distribution that best fit the observed data under the assumption that the total probability of observing the data (the joint probability) is the product of observing the data points individually. The probability of observing a single data point is given by the formula 2.1. The following formula gives the joint probability function:

$$P(x; k, \lambda) = \prod_{i=1}^{n} P(x_i; k, \lambda) \tag{3.3}$$

Here, x_i is an individual observation, and k and λ are the Weibull parameters.

The partial derivates of the joint probability distribution for both parameters are used to calculate the maximum likelihood estimates. The optimal parameters are found by setting the function to zero and rearranging the equation to make the parameters of interest the subject of the equation.

Weibull distributions were fitted on all samples using a nonlinear mixed-effect (NLME) model for the estimation step. NLME models incorporate a fixed effect, which is the same for all samples, and a random effect that assumes all individual samples are sampled from a normal distribution with zero mean and unknown variance.

Some 1-NN distance distributions could not be fitted as Weibull distributions (Figure 8). The distributions often belonged to cell-type combinations with a handful of cells (Figure 8AB). If convergence

of the NLME model was not achieved on all samples, we removed samples with fewer than 10 cells (both reference and target cell count). If convergence was again not achieved, this was repeated with samples less than 20 and 50. For some combinations, the 1-NN distance distribution has a multimodal behavior due to the natural positioning of cells (Figure 8CD). Therefore 100 % convergence was not reached.

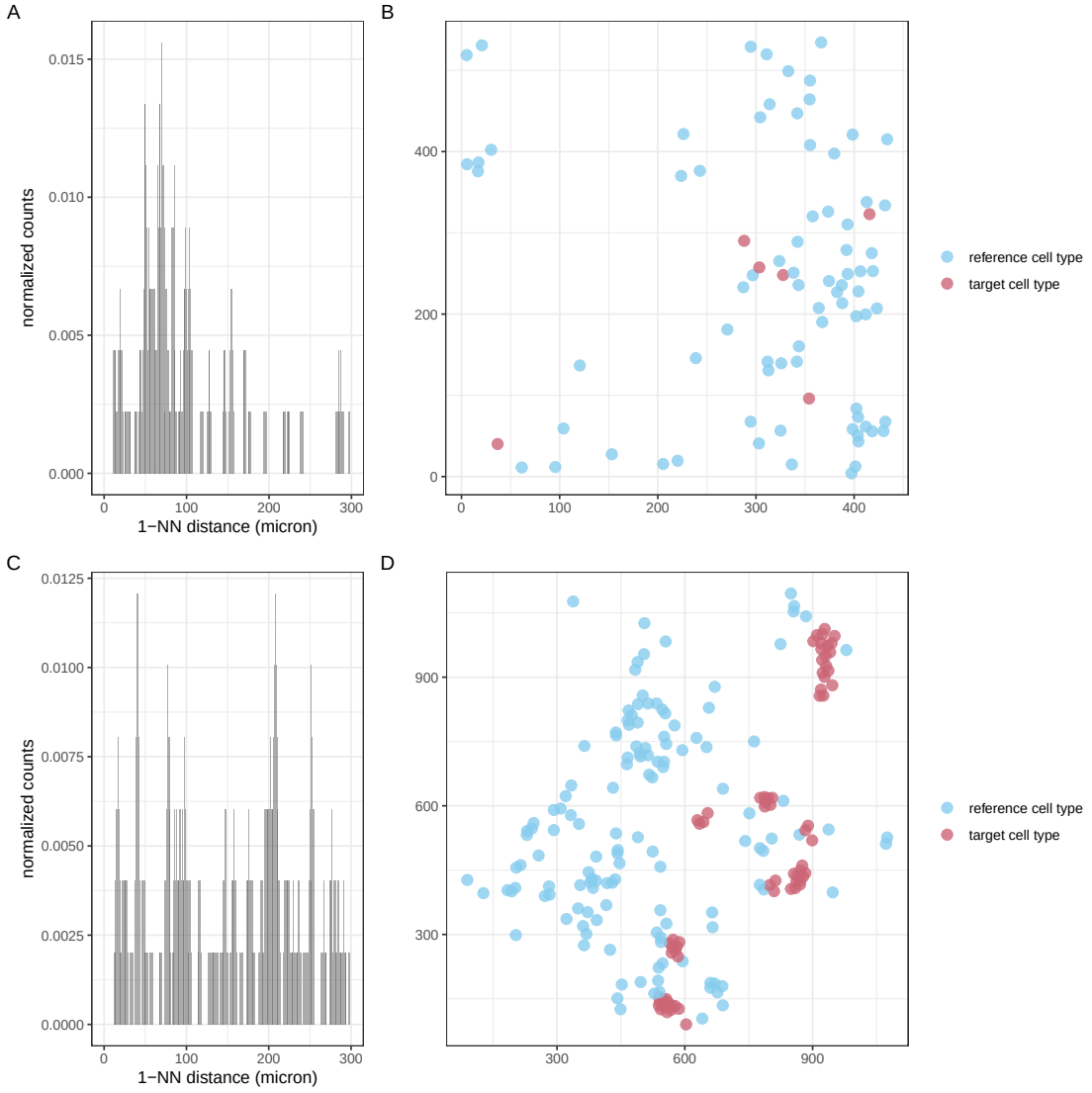

Figure 8: Two 1-NN distance distributions with multimodal behavior (AC). Distributions belong to combinations with low cell counts (B), or where the target cell type is clustered in small groups(D).

The NLME model finds robust parameter estimations for cell-type combinations because parameter estimations are based on the distance distributions of all samples. The model assumes that distance distributions of a cell-type combination are similar across all samples. Distance distributions with behaviors deviating from the other samples are also fitted with a Weibull distribution matching the general behavior of the other samples. Those distributions are removed from the population by computing the goodness-of-fit of the Weibull distribution. The goodness-of-fit was calculated with the following formula:

$$\Delta_M = |\lambda(\ln 2)^{1/k} - M(v_{distances})| \tag{3.4}$$

Here, the first term is the median of the Weibull distribution. The second term is the median of the actual distances. If the difference exceeds 80 microns, the spatial relationship is labeled as an outlier.

3.5. Feature Associations

Association strengths between features and samples of a subtype were calculated with a two-sample t-test as implemented by *t.test* of the *stats v4.2.0* package. A two-sample t-test measures whether the means of the subtype samples and all other samples are significantly different with the following formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.5}$$

Here, $\bar{x}$ is the mean, s is the variance of this vector and n is the number of elements.

Using a Benjamini-Hochberg correction, P-values are corrected for multiple testing per feature over all samples. The adjusted p-values are calculated with the following formula:

$$p_{adjusted} = \frac{i}{m} \cdot 0.05 \tag{3.6}$$

Here i is the rank of the original p-value and m is the total number of tests. Features were labeled significant if the adjusted p-value was smaller than 0.05.

Additionally, the log2 fold change of the means was computed with the following formula:

$$log2foldchange = \log_2\left(\frac{\bar{x}_1}{\bar{x}_2}\right) \tag{3.7}$$

The log2 fold change indicates the direction of change (increase or decrease).

3.5.1. Recurrent cell types in spatial relationships

The Fisher's exact test was used to identify recurrent reference or target cell types in significant associations. The statistical test determines whether there is a significant difference in category proportions between multiple groups from a contingency table. For all cell types the following contingency table was constructed:

	Associated relationships with cell type A as reference/target	Associated relationships without cell type A as reference/target
Subtype I	a	b
Not subtype I	c	d

The Fisher's exact test computes the probability that the contingency table is obtained on the null hypothesis. The null hypothesis states that the reference or target cell type occurs equally in association with the subtype and all other subtypes. Fisher's exact test is computed with the following formula:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3.8}$$

Fisher's exact test was implemented by *fisher.test* of the *stats v4.2.0* package. The null hypothesis is rejected when the p-value < 0.05 after multiple testing corrections.

3.5.2. Effect of cell-type fractions on spatial relationships

Cell-type fractions potentially affect spatial relationships. For abundant cell types, distances between cells are inherently smaller, while two rare cells are more likely to have large distances between them. We assessed the Pearson correlation coefficients for the Weibull parameters and cell-type fractions for each spatial relationship significantly associated with a subtype. The Pearson correlation coefficient captures the linear dependence of two random vectors. Spatial relationships for which the parameters are significantly correlated with the cell type fractions (p-value < 0.05) are impacted by the number of cells. Spatial relationships for which all correlations are not significant are independent of cell-type fractions.

The effect was visualized by plotting the Pearson correlation coefficients in two plots: one where the correlations of the shape parameters and target fractions were plotted against the correlations of the scale parameters and target fractions and another where the correlations between shape parameters and reference fractions were plotted against correlations of the scale parameters and reference fractions.

3.6. Subtype Predictions

Subtypes are predicted by a regularized logistic regression model as implemented by *cva.glmnet* of the *glmnet v4.1-7* package. The model assigns coefficients to features to find a linear combination that separates the two classes. Additionally, an elastic net model, which is a regularization method that combines ridge and lasso (least absolute shrinkage and selection operator) regularization encourages a grouping effect where strongly correlated features are either removed or retained together (Zou and Hastie, 2005). The optimal coefficients (β) are found with the following formula:

$$\beta = \underset{\beta}{\text{argmin}} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \alpha\lambda\|\beta\|_1 + (1-\alpha)\lambda\|\beta\|_2^2 \right) \tag{3.9}$$

The balance between lasso and ridge regression in *cva.glmnet* is set with the α parameter (0 only ridge and 1 only lasso) and the strength of the regularization penalty is set by the λ parameter. Both parameters are optimized with 10-fold cross-validation.

All predictions were run with 5-fold cross-validation in which the subtype proportions are stratified. The performance is reported with the AUC value of the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), at various threshold settings. AUC values of models with good performance are close to 1, while the AUC value of a random guess equals 0.5.

3.7. Survival Predictions

Associations between features and survival outcomes were quantified using a Cox regression model (Cox, 1972). The model is essentially a regression model that models the effect of features against the rate of deaths over time (called hazard rate) with the following formula:

$$h(t|\mathbf{X}) = h_0(t) \cdot \exp(\mathbf{X}^T \boldsymbol{\beta}) \tag{3.10}$$

Here, $h_0(t)$ is the baseline hazard function representing the hazard when all covariates are zero. The hazard ratios are the quantities $\exp(\beta_i)$. Hazard ratios greater than one indicate that the survival probability decreases as variable x_i increases. Similarly, if the hazard ratio is between zero and one, the survival probability increases, with an increase in the covariate.

The effect of individual features was estimated using *coxph* of the *survival v3.5-5* package. HER2 status was always added as a covariate.

The effect of individual continuous variables was assessed by converging the variable into a discrete form. Samples were assigned a zero if their value was equal to or below the median and a one otherwise. The functions were plotted with *ggsurvplot* of the *survminer v 0.4.9* package. The functions were compared using the log-rank test.

3.8. Neighborhood Identification

Dense cell regions were detected from a graph representation of the TME with the Walktrap algorithm (Latapy and Pons, 2004) as implemented in *cluster_walktrap* of the *igraph v1.4.2* package. Walktrap is an algorithm from graph theory that identifies communities in large networks via random walks. First, random walks are used to compute distances between nodes, and nodes are then assigned into groups with small intra- and larger inter-community distances via hierarchical clustering. The TME was transformed into a graph by representing each cell centroid as a node and connecting cells within a distance of 8 micrometers.

Subsequently, the resulting communities were characterized by connectivity profiles. The connectivity profile of a neighborhood is constructed by separating cells based on vertex degrees ranging from 1 to 9+. Cell-type proportions are computed for each vertex degree subset (Figure 9). The resulting connectivity profiles are 16 by 9 matrices with the TME cell types as rows and vertex degrees 1 to 9+ as columns. Neighborhoods are clustered using hierarchical clustering.

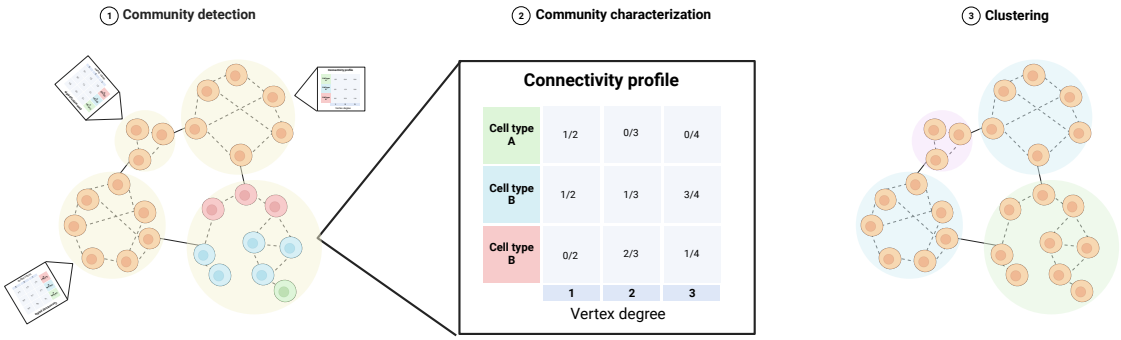

Figure 9: Overview of neighborhood detection by Danenberg et al. (2022). Dense cell regions are identified with a community-detection algorithm (1). For all dense regions, connectivity profiles are constructed that divide all cells in vertex degree bins and measure the proportion of cell types in each bin (2). The connectivity profiles are used to cluster regions with recurrent profiles into neighborhoods (3). Created with BioRender.com.

3.9. Region Segmentation

Parts of the tissue that contain many neighborhoods were separated from the rest of the tissue with a kernel density estimation (KDE) of the point patterns defined by cells belonging to a neighborhood ($KDE_{neighborhood}$) and all other cells (KDE_{other}). KDE is a non-parametric method that estimates the class-conditional marginal densities as a function of the coordinates. KDE was applied as implemented by *density* in the *stats v4.2.0* package. The smoothing bandwidth for the KDE was optimized using likelihood cross-validation as implemented by *bw.ppl* in the *spatstat v3.0-5* package. The bandwidth for the $KDE_{neighborhood}$ was amplified five times to ignore isolated neighborhoods and join tightly packed neighborhoods into one region. Cells are classified as 'neighborhood' if $KDE_{neighborhood} > KDE_{other}$ and as "other" otherwise (Figure 10).

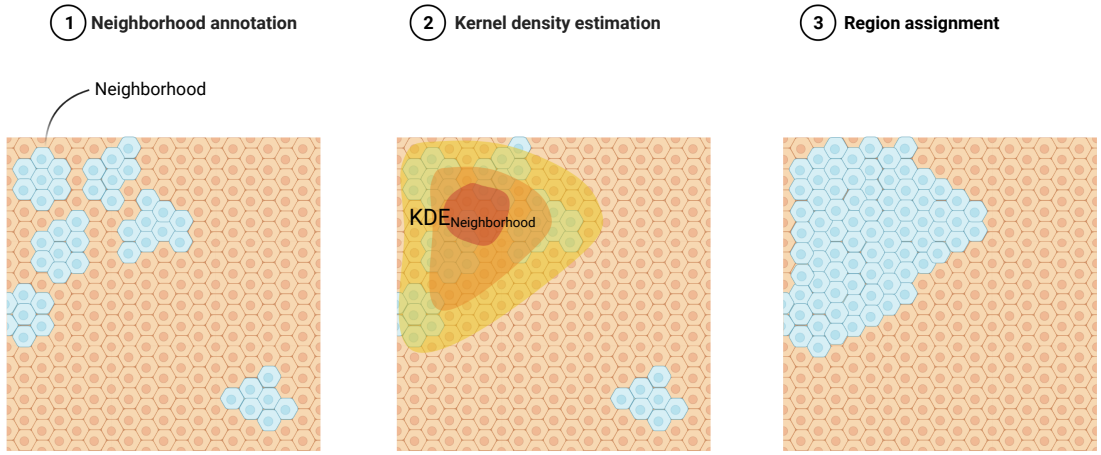

Figure 10: Region segmentation using kernel density estimation. Neighborhoods are annotated in the tissue (1) and two probability density functions are estimated (KDE_{other} is omitted here) (2). Individual cells are marked based on the function with the highest probability (3). Created with BioRender.com.

4

Results

This chapter presents the conducted research and the resulting insights. It begins by identifying cell types and their abundance among tumor subtypes. Subsequently, it calculates and examines the spatial relationships between pairs of cell types. An interpretable prediction model quantitatively evaluates the predictive capabilities of cell-type fractions and spatial relationships. The model also identifies feature combinations that offer predictive value for each subtype. Furthermore, it assesses associations between spatial relationships and survival outcomes, offering insights into the potential prognostic significance of these spatial factors. Lastly, it explores the spatial relationships within TME neighborhoods, uncovering the spatial characteristics of these cellular regions and enhancing our understanding of the interactions and collaborations associated with local TME functions.

4.1. Cellular Composition of Breast Cancer Subtypes

PAM50 is a standardized method that classifies tumors into intrinsic molecular subtypes based on the expression of 50 genes within tumor cells. The classification does not consider heterogeneity in the composition of other epithelial and non-epithelial cells. In this section, we provide additional characterization of these subtypes by describing the cellular composition of tumor and TME tissue regions. Cell types are derived from the extensive protein expression profiles obtained with imaging mass cytometry.

4.1.1. Epithelial cell types

Cells in the epithelium are separated into tumor and non-tumor cells based on the expression of cytokeratins. Cytokeratins are found in the cytoskeleton of epithelial tissue and are used clinically to identify tumor cell activity (Barak et al., 2004). Tumor cells of each subtype express distinct cytokeratin types: basal tumors express cytokeratin 5; HER2-enriched and normal-like tumors express mostly pan-cytokeratin; luminal B tumors express predominantly cytokeratin 8-18, and luminal A tumors express both pan-cytokeratin and cytokeratin 8-18. Furthermore, subtype tumors have distinct hormone receptor profiles consistent with the immunohistochemical assessment: basal tumor cells do not contain receptor proteins; HER2-enriched tumor cells contain predominantly HER2 proteins; and luminal A, luminal B, and normal-like tumor cells contain mostly ER proteins (Figure 11).

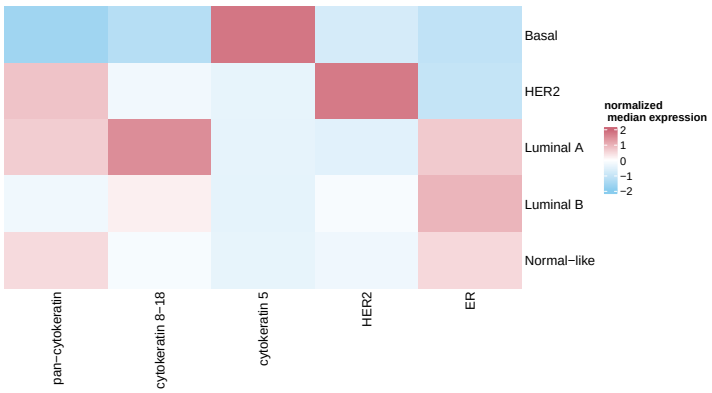

Figure 11: Normalized median expression of three tumor markers (pan-cytokeratin, cytokeratin 8-18, and cytokeratin5) and receptor proteins (ER and HER2) within tumor cells of PAM50 subtypes.

Nine tumor cell types were distinguished from the expression levels of ER, HER2, CD57, CD15, and cytokeratins. Additionally, the epithelial cells without high levels of cytokeratin were separated into seven epithelial cell types, including MHC-presenting cells, proliferating cells (Ki67$^+$ cells), and normal epithelial cells. The alternative classification merges the tumor cell types into the cell types basal, HER2$^+$, CK$^+$ER$^+$ and CK$^+$ER$^-$ cells, because the expression of HER2 and ER predominantly characterizes tumor subtypes.

Inspections of the normalized median fractions among subtypes revealed tumor cell-type compositions consistent with tumor cell expression profiles. Basal tumors contain an abundance of basal cells, HER2-enriched tumors contain mostly HER2$^+$ cells, luminal A, luminal B, and normal-like tumors contain similar numbers of ER$^+$ tumor cells. Normal-like tumors also contain high fractions of normal epithelial cells (Figure 12).

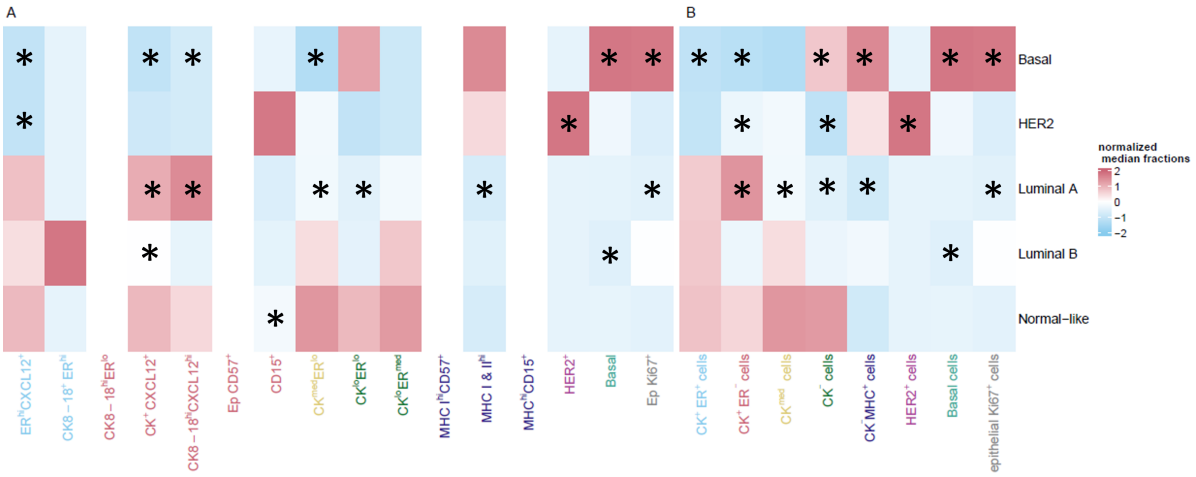

Figure 12: Normalized median fractions of the Danenberg epithelial cell types (A) and alternative epithelial cell types (B). The colors of the cell-type labels indicate what Danenberg cell types are merged in the alternative classification. Asterisks indicate a significant association between cell type fraction and subtype.

Association strengths between cell-type fractions and tumor subtypes were measured using two-sample t-tests, revealing significant associations for all subtypes (Figure 13). Basal tumors are associated with higher fractions of basal cells, MHC-presenting cells, normal epithelial cells, and proliferating cells. The Danenberg classification separated MHC-presenting cells into three cell types, that are not significantly associated with basal tumors. Finally, Basal tumors are associated with lower fractions of other cell types expressing cytokeratin.

HER2-enriched tumors are associated with higher fractions of HER2$^+$ cells and lower fractions of normal epithelial cells and tumor cells expressing ER proteins. Associations with lower fractions are only evident from the alternative cell types.

Luminal A tumors are associated with higher fractions of CK^+ER^- cells and lower fractions of normal epithelial cells, proliferating cells, and MHC-presenting cells.

Luminal B tumors are only associated with lower fractions of basal cells and $CK^+CXCL12^+$ cells. The generalization of $CK^+CXCL12^+$ cells to CK^+ER^- in the alternative classification resulted in the loss of its associations with luminal B tumors.

Finally, normal-like tumors are only associated with lower fractions of $CD15^+$ cells. $CD15^+$ cells were merged with other tumor cells without ER proteins (CK^+ER^- cells) in the alternative classification and normal-like tumors are not associated with this cell type. The median fractions suggest that normal-like tumors contain high fractions of normal epithelial cells, but the tumors are not significantly associated with CK^- cells.

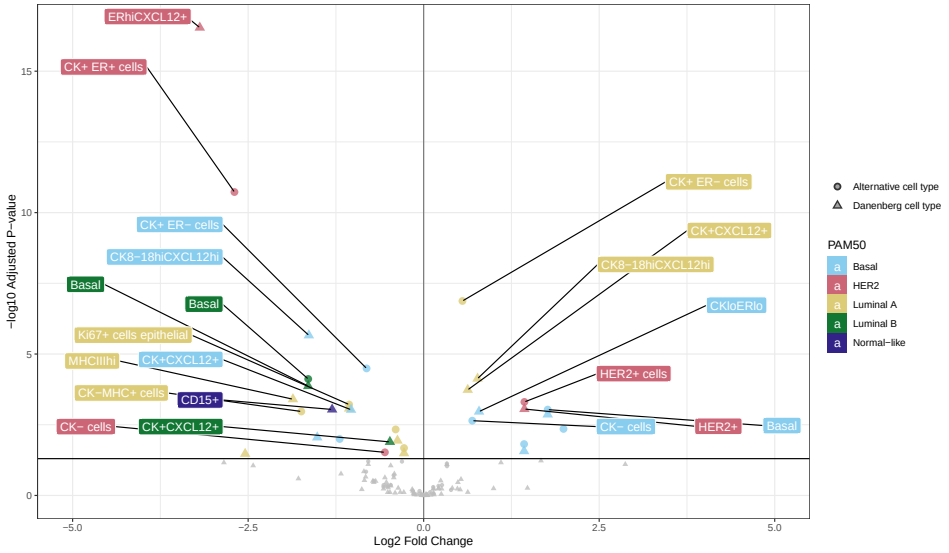

Figure 13: Associations between fractions of epithelial cell types and PAM50 subtypes. Associations are colored by subtype, and the shape indicates Danenberg and alternative cell types. The log2 fold change indicates the difference between the mean fractions. Labels indicate the most significant associations per subtype.

4.1.2. TME cell types

TME cells were distinguished from epithelial cells based on the evaluation of cytokeratin expression and cell morphology. The cells in the TME were classified into 16 cell types. Normalized median fractions of these cell types show varying prevalence across the subtypes (Figure 14).

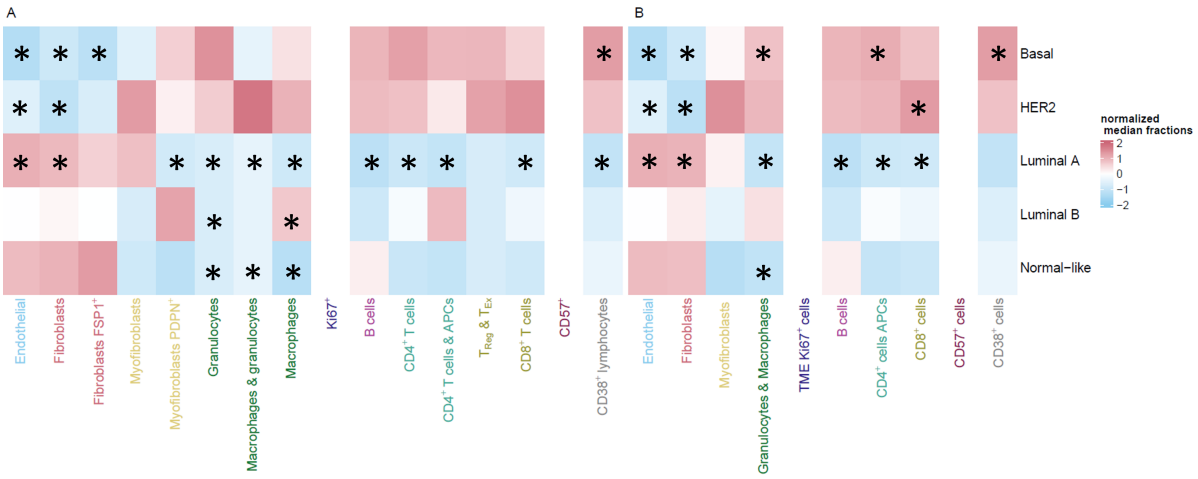

Figure 14: Normalized median fractions of the Danenberg TME cell types (A) and alternative epithelial cell types (B). The colors of the cell type labels indicate what Danenberg cell types are merged in the alternative classification. Asterisks indicate a significant association between cell type fraction and subtype.

Associations between cell-type fractions and tumor subtypes revealed large differences in the TME composition across all subtypes (Figure 15). Basal cells are associated with higher fractions of $CD38^+$ cells, $CD4^+$ T cells, and granulocytes & macrophages. There is no association with the separate fractions of granulocytes and macrophages. Furthermore, basal tumors are associated with lower fractions of fibroblasts and endothelial cells.

HER2-enriched tumors are associated with higher fractions of $CD4^+$ cells & APCs, but not with specific fractions of $CD4^+$ T cells and APCs. Additionally, HER-enriched tumors are associated with lower fractions of fibroblasts and endothelial cells, similar to basal tumors.

Luminal A tumors are associated with many cell-type fractions, including higher fractions of fibroblasts and endothelial cells and lower fractions of granulocytes, macrophages, B cells, $CD4^+$ T cells, and $CD8^+$ T cells.

Luminal B tumors are only associated with higher fractions of macrophages and lower fractions of granulocytes. In the alternative classification, both cell types were merged, but luminal B tumors are not associated with the fraction of this combined cell type.

Finally, normal-like tumors are associated with lower fractions of macrophages and granulocytes.

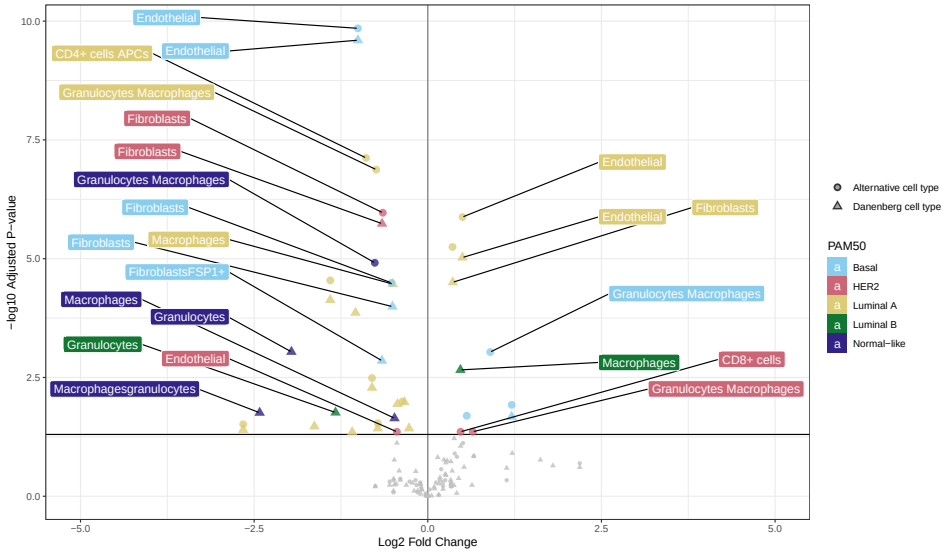

Figure 15: Associations between fractions of TME cell types and PAM50 subtypes. Associations are colored by subtype, and the shape indicates Danenberg and alternative cell types. The log2 fold change indicates the difference between the mean fractions. Labels indicate the most significant associations per subtype.

4.1.3. Section Summary

The characteristics of breast cancer tumor cells have been studied extensively, and intrinsic tumor types have been established based on histological and molecular properties. In our dataset, the hormone receptor expression of tumor cells is consistent with previous descriptions of PAM50 subtypes (Curtis et al., 2012). HER2-enriched tumors contain $HER2^+$ cells, basal tumors contain basal cells without ER and HER2, and tumor cells of luminal A, luminal B, and normal-like subtypes share a commonality, predominantly expressing ER.

After dissecting tumor compositions, we extended our subtype characterization to encompass the entire epithelium and TME. This broader investigation aimed to address Research Question 1:

> What cell types can we distinguish from protein expression profiles in breast cancer tumors and TMEs?

We established two cell-type classifications, employing the approach detailed in Danenberg et al. (2022) and an alternative strategy where rare cell types were grouped into larger entities.

The alternative classification revealed that merging or splitting cell types can unveil or eliminate associations with subtypes. For example, the broader categorization of MHC-presenting cells revealed associations with basal and luminal A tumors not apparent when analyzing individual cell-type fractions. Similarly, combining granulocyte and macrophage fractions unveiled associations with basal tumors not

detected when considering these cell types individually. Conversely, merging these cell types removed the associations of granulocyte and macrophage fractions with luminal B tumors.

By computing the associations between subtypes and all cell types we addressed the first part of Research Question 3:

What cell types and spatial relationships are associated with breast cancer subtypes?

The intrinsic differences of PAM50 subtypes find support in the presence of specific cell types. Basal tumors are characterized by higher fractions of MHC-presenting cells, proliferating cells, and normal epithelial cells in addition to the anticipated prevalence of basal cells and the absence of other tumor cell types. The TME of basal tumors contains high proportions of $CD38^+$ cells, $CD4^+$ cells & APCs and granulocytes & macrophages and low numbers of fibroblasts and endothelial cells. The TME of HER-enriched tumors is similar to basal tumors containing similar cell types and their distinction from other subtypes emerged due to the absence of fibroblasts and endothelial cells.

Luminal A tumors are typified by high proportions of tumor cells expressing CXCL12 and small proportions of normal epithelial cells. The TME of Luminal A tumors is restricted to high proportions of endothelial cells and fibroblasts.

Associations of cell-type fractions with luminal B and normal-like tumors were relatively limited. Previous research confirms the challenge of distinguishing unique properties of these subtypes due to the striking similarity between luminal B and luminal A (Kensler et al., 2019), as well as the substantial contamination of normal breast tissue in normal-like tumors (Nolan et al., 2023). Although median cell-type fractions initially suggested a distinct characterization for these subtypes, statistical tests did not find significant associations to confirm this impression.

4.2. Spatial Relationships between Cell Types

Associations between cell-type fractions and tumor subtypes showed that the tumor's and TME's composition differ across subtypes. Here, we further extend the characterization of both compartments by considering the distance relationships of all cell-type pairs. The relationships were captured with 1-NN distances, and the distributions were summarized with the Weibull parameters aiming to reflect the unique distribution behaviors with two values. This section discusses how the parameters were obtained and how they result from specific cell-type combinations. Additionally, the section evaluates how parameter pairs represent different spatial arrangements.

4.2.1. Parameter estimations

Initially, spatial relationships were estimated for all combinations of Danenberg cell types (n=32) in all samples (n=749). A cell-type combination was estimated successfully if a Weibull distribution was fitted to its 1-NN distance distribution. The number of distributions that were fitted was low because only 35 % of all existing combinations contained more than 20 reference and 5 target cells. Moreover, parameter estimations for 7 % of the remaining combinations did not converge. Finally, 28 % of the existing combinations were estimated, but many contained the same epithelial cells ($CK^{lo}ER^{lo}$ cells and $CK^{med}ER^{lo}$ cells) or TME cells (fibroblasts, myofibroblasts, and endothelial cells). Only a small part of the spatial relationships were estimated in more than 50 samples (Figure 16A).

Due to the low number of successful estimations, spatial relationships were also estimated for the alternative cell types. Alternative cell types were more abundant and 48 % of the existing combinations contained more than 20 reference and 5 target cells. Additionally, parameter estimations for 4 % of the combinations did not converge. Estimated combinations contained a wider variety of cell types and 64 % of the combinations were estimated in more than 50 samples (Figure 16B).

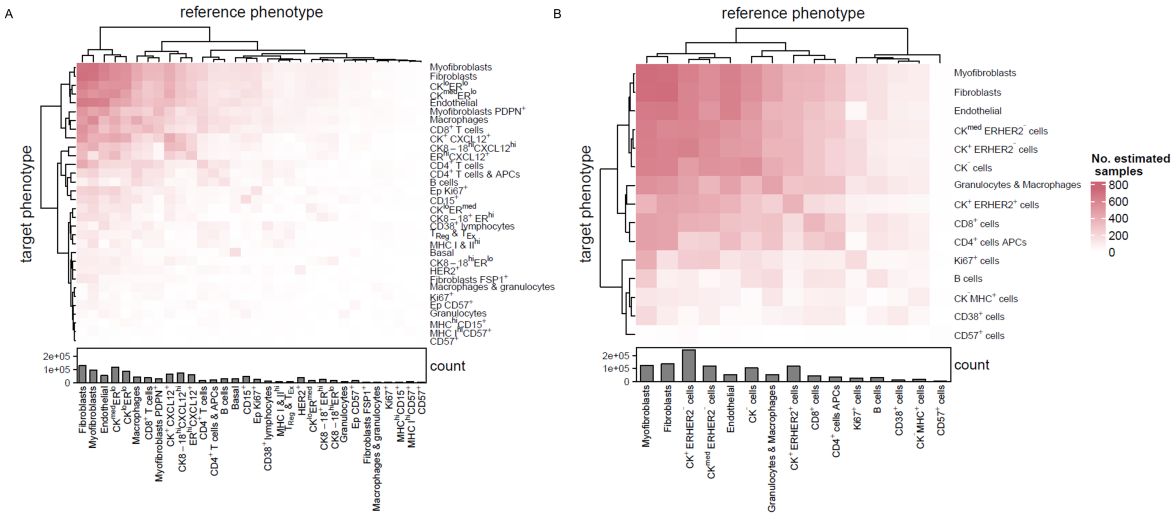

Figure 16: Number of successful parameter estimations for combinations between Danenberg cell types (A) and the alternative cell types (B).

The spatial arrangements and distance distributions of five parameter pairs in the extremes of the parameter space show how the Weibull parameters correspond to unique spatial relationships (Figure 17). Tightly packed cells with small 1-NN distances are characterized by a low scale and high shape (Figure 17B). If there is a large spread in the 1-NN distances, while the overall median remains low, the shape parameter decreases (Figure 17C). The scale also increases when the distance distribution variance increases (Figure 17D). Cells segregated into two compartments are characterized by high shape and scale parameters (Figure 17E). When the target cell type encloses the reference cell type compartment the scale decreases slightly (Figure 17F).

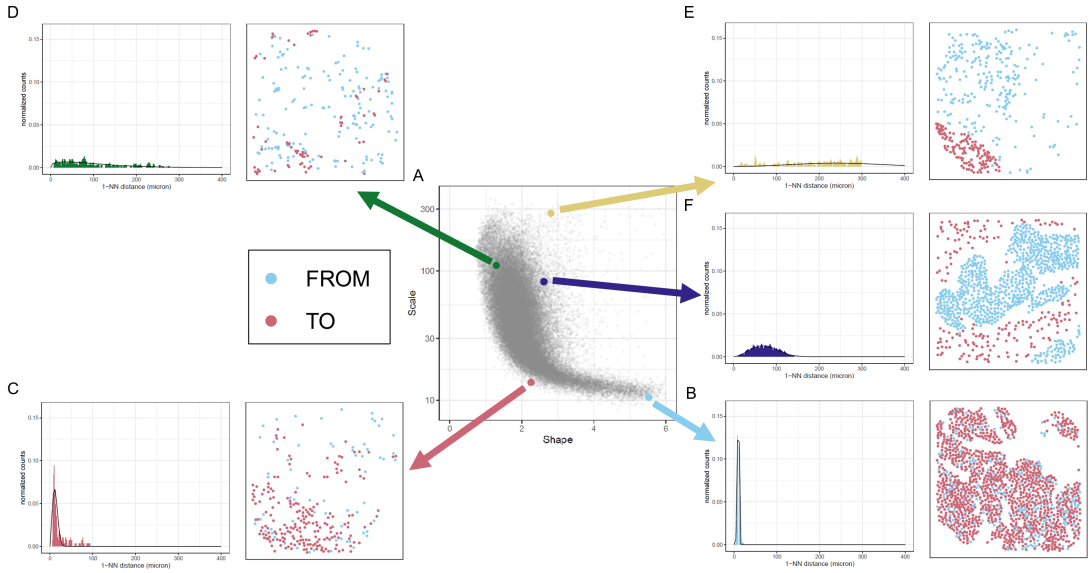

Figure 17: Plotted parameter pairs for every alternative cell type combination (n=324) and sample (n=749) with five example points and their corresponding distance distribution and point pattern.

After the parameter estimations, the goodness-of-fit of the Weibull distribution was tested to remove outliers. Combinations are labeled as outliers if the median of the fitted Weibull distributions differs more than 80 microns from the actual 1-NN distance median. A plot of the outliers shows that outliers are always in the high scale parameter region (Figure 18B). Distance distributions of parameter space extremes revealed that high scale parameters correspond to distributions with a low amplitude and large variance (Figure 17DE). The 1-NN distance distributions are very similar, but the medians differ significantly. Therefore, the median of the distance distribution changes radically due to a small set of outlier points. This effect is larger for distance distributions that contain few bins, explaining why the 1-NN distance distributions of combinations that include rare cell types such as $CD57^{+}$ cells are often labeled as outliers.

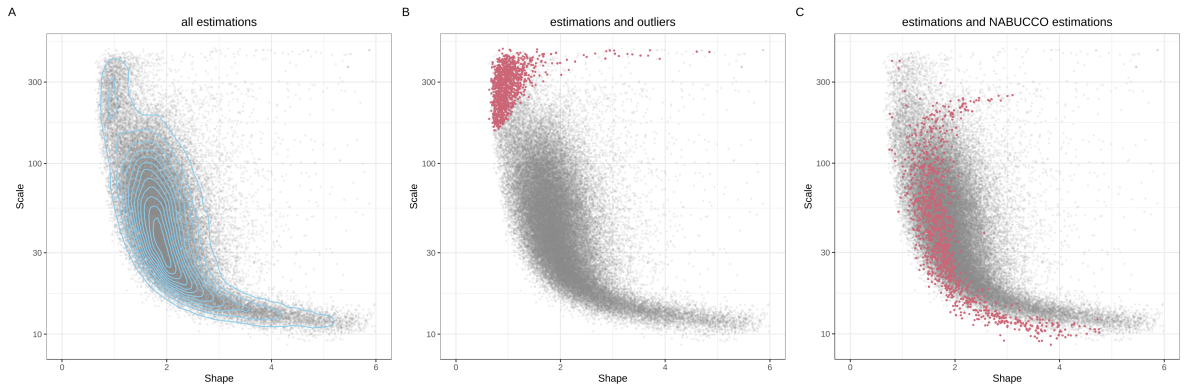

Figure 18: Plotted parameter pairs for every alternative cell type combination (n=324) and sample (n=749) before outlier removal (A). Plotted parameter pairs with outliers (B). Plotted parameter pairs with parameter pairs from the spatial analyses of the NABUCCO cohort (C).

The proportion of tumor-to-tumor, TME-to-TME, tumor-to-TME, and TME-to-tumor combinations throughout regions of the parameter space show that the combinations occur in recurrent distributions. Tumor cell types are either densely packed together (high shape and low scale) or separated in tumor compartments (high shape and high scale) (Figure 19A). The median 1-NN distance between two TME cell types is typically small, but the distance spread may vary (Figure 19B). Furthermore, high proportions of TME-to-tumor and tumor-to-TME combinations in regions with high scales and high

shapes indicate that the cell types reside in separate compartments (Figure 19C). Tumor compartments are often encapsulated by TME cells (Figure 19D).

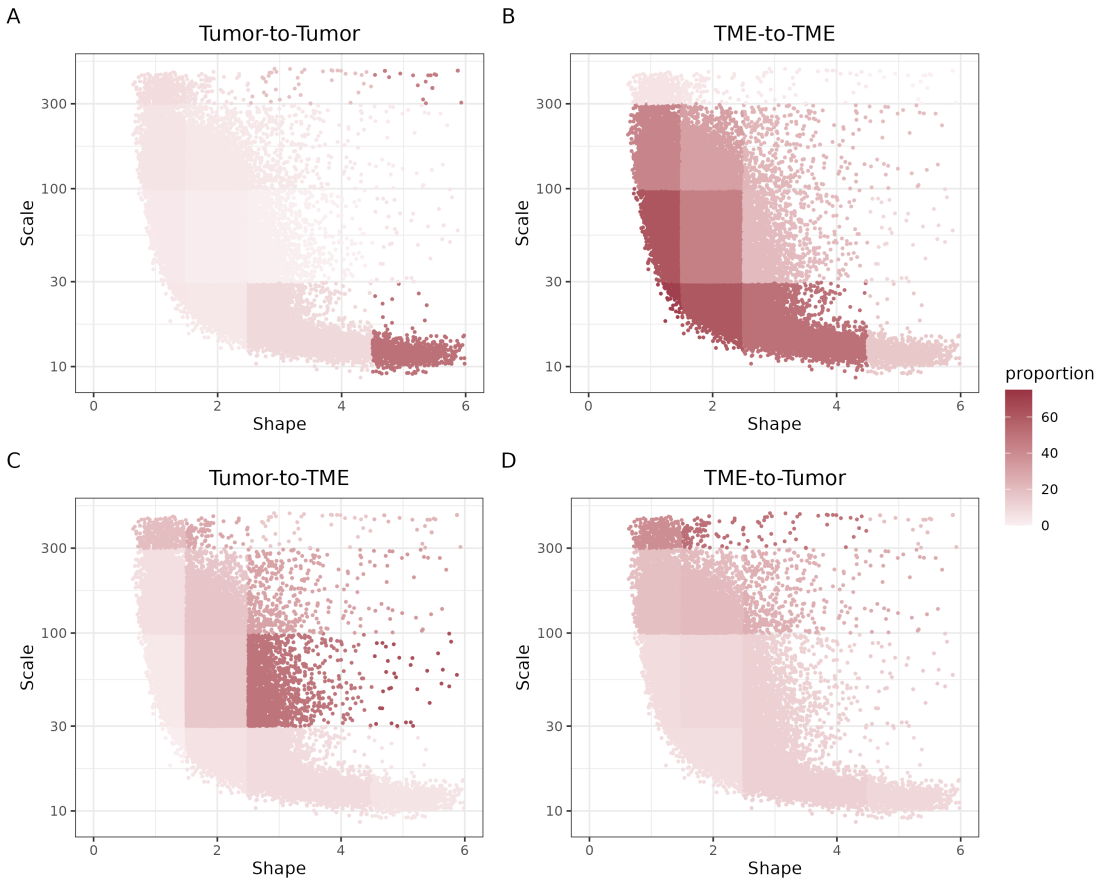

Figure 19: Proportion of tumor-to-tumor (A), TME-to-TME (B), tumor-to-TME (C), and TME-to-tumor (D) combinations within 12 subregions of the parameter space.

Section 4.1 has shown that the abundance of cell types differs significantly across tumor subtypes. Considering the potential effect of cell-type abundance on spatial relationships is important. Cell-type combinations that contain high target cell-type fractions are often characterized by higher shape and lower scale parameters (Figure 20BD). The abundance of the reference cell type does not affect the spatial relationship in this way (Figure 20AC).

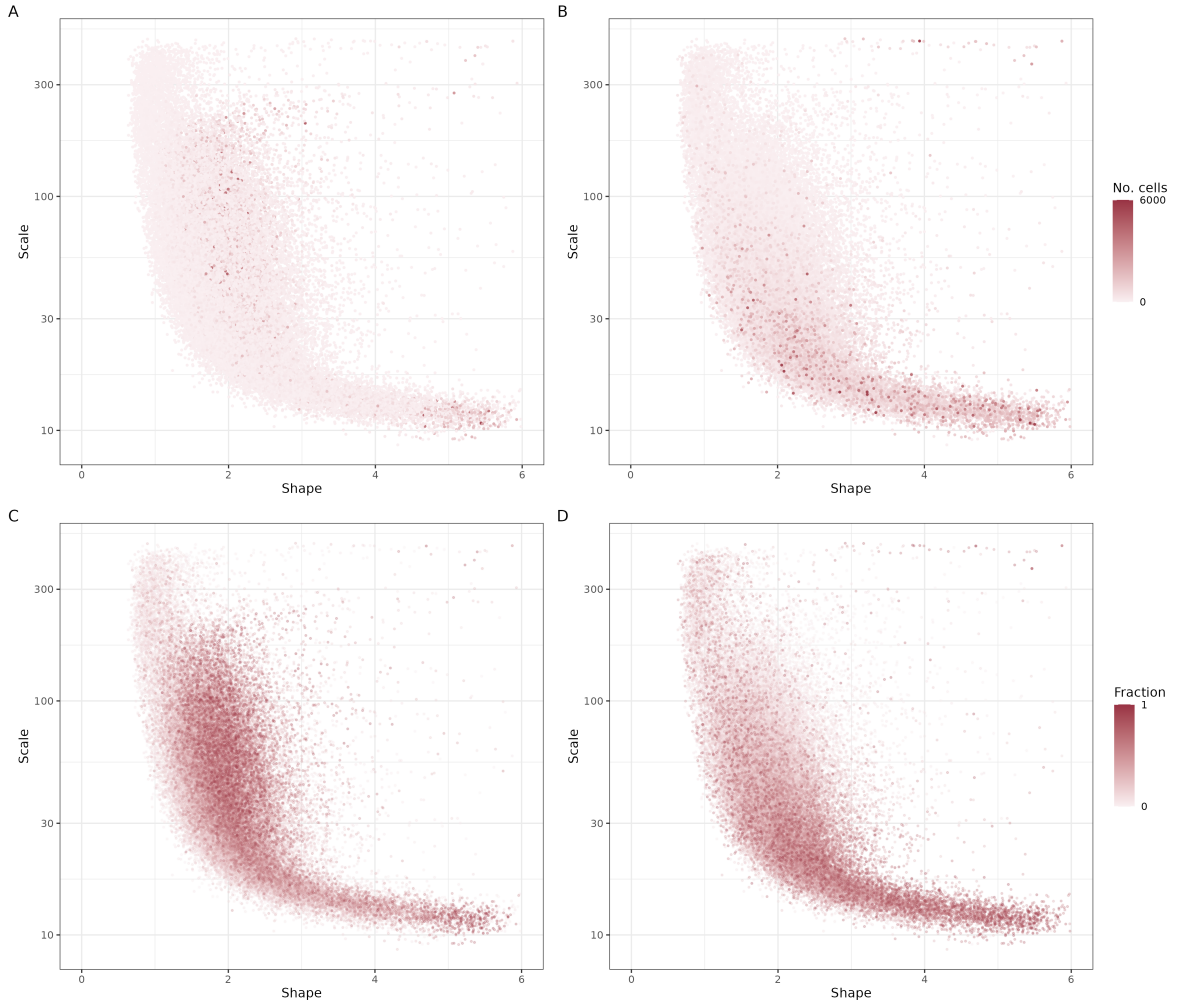

Figure 20: Plotted parameter pairs colored by: reference cell type count (A), reference cell type fraction (C), target cell type count (B), and target cell type fraction (D).

4.2.2. Comparison to spatial analysis of NABUCCO trial

In a previous study by Gil-Jimenez et al., the spatial relationships of 49 cell-type combinations in the NABUCCO trial tissue samples were estimated using the same approach as applied in this research. The representation of Weibull parameter pairs in a two-dimensional space exhibited a distinctive C-shaped distribution (appendix C). The overlay plot of both parameter spaces shows a different distribution (18C). The variations are explained by large differences across the datasets used in both studies. The dataset used in this study consists of 749 images in which 18 cell types were compared. The spatial relationships of four tumor cell types were considered and as these cells are generally tightly packed together, this results in more combinations with high shapes and low scales. The NABUCCO dataset contains tissue images annotated with seven cell types for 24 patients. The cell-type classification contains only one tumor subtype, while the Danenberg classification distinguishes 12 types.

Moreover, tissue slides were imaged with mIF instead of cyTOF, resulting in ten-fold larger images. The images show a larger part of cellular compartments, allowing reference cells to be further apart from target cells. These spatial relationships are characterized by high shape and scale parameters. Therefore, there is also a difference in this region of the parameter space.

Finally, the smaller images and more extensive cell-type stratification result in cell types with low counts. The distance distributions of combinations with lower cell-type counts are more sensitive to the effect of single cells, which explains why there is more variation in the center and top part of the parameter space.

4.2.3. Section Summary

In Section 4.1, we uncovered the occurrences of unique cell types across breast cancer subtypes. Subsequently, the spatial analysis revealed how cell types are organized relative to each other, indicating how cells cooperate and interact. We applied a novel approach to capture unique distance relationships. First, all 1-NN distances between a reference and target cell type are measured. Then a Weibull distribution is fitted to the distance distribution to summarize its behavior with two parameters. The systematic analysis of spatial relationships between all cell-type pairs addressed Research Question 2:

What are the spatial relationships of cell types throughout breast cancer tumors and TMEs?

We find that Weibull parameter pairs represent a broad spectrum of spatial relationships, encompassing arrangements from closely packed clusters to distinct compartments. The regions of the parameter space were linked to tumor and TME cell combinations, confirming the inherent organizational principles of tumors and TMEs.

Furthermore, we examined whether the spatial relationships of cell-type pairs correlated with their abundance. The investigation revealed that cell-type pairs with many target cells often exhibited tightly packed arrangements characterized by high shape and low scale parameters. Interestingly, the reference cell-type count did not display a significant correlation with the parameter estimates.

It is crucial to acknowledge that our dataset consists of relatively small images with many cell types limiting the spatial analysis. Many cell-type combinations have insufficient cell counts to accurately represent the true spatial distribution of cells. To mitigate this, we implemented various lower-bound thresholds and outlier removal steps to address instances of undersized combinations, thus ensuring more reliable results.

Finally, we encountered several combinations with sufficient cell counts for which Weibull distribution estimations failed. These combinations exhibited 1-NN distance distributions with multimodal behavior, making them unsuitable for fitting with a Weibull distribution. Although alternative methods utilizing metrics like the mean or median can provide values for multimodal distance distributions, these estimates tend to oversimplify and potentially mislead. Further research is necessary to ascertain whether multimodal distributions are an artifact of a low cell count.

4.3. Predicting Breast Cancer Subtypes

In this section, we quantitatively assess the features' predictive ability with the performance of a regularized logistic regression model that predicts subtypes from both tumor cell-type fractions, TME cell-type fractions, and Weibull parameters.

Logistic regression models are interpretable prediction models because model coefficients indicate the effect of feature combinations on prediction outcomes. Positive coefficients signify that the subtype probability increases with an increase in the corresponding feature (given that all other features remain constant). Conversely, negative coefficients indicate a smaller probability for higher feature values. Moreover, by using regularization, the model removes uninformative features. Coefficients provide valuable insights into the associations between feature combinations and subtypes.

4.3.1. Predictions based on Danenberg and alternative cell-type fractions

First, the performance of models trained with Danenberg and alternative cell-type fractions are compared to evaluate the effect of the generalization of cell types. In section 4.1, it was shown that subtypes are, in some cases, only associated with joint cell types, while associations with individual cell types are not significant. Basal tumors, for example, are only associated with $CK^{-}MHC^{+}$ cells, and not with individual fractions of MHC-i^{hi} $CD57^{+}$, MHC-I & MHC-II^{hi} cells or $MHC^{hi}CD15^{+}$ cells. Similarly, the alternative cell-type classification removed multiple associations with fractions of Danenberg cell types because we merged the associated cell types with other cell types. Luminal B tumors, for example, are associated with macrophage fractions but not with granulocytes & macrophage fractions.

Despite several differences in univariate feature associations, subtype prediction performances are similar for models trained on the Danenberg and alternative cell-type fractions (Table 2). Hence, the alternative classification does not remove or reveal features significantly affecting the predictions.

Table 2: Mean performance of subtype prediction experiments. Performance is reported as the AUC value of the receiver operating characteristic curve. The highest AUC values are in bold.

Dataset	Basal	HER2	Luminal A	Luminal B	Normal-like
Danenberg tumor fractions	**0.82 (± 0.08)**	0.78 (± 0.08)	**0.70 (± 0.09)**	**0.60 (± 0.09)**	0.52 (± 0.03)
Alternative tumor fractions	0.79 (± 0.03)	**0.79 (± 0.04)**	**0.70 (± 0.04)**	**0.60 (± 0.07)**	0.47 (± 0.03)
Danenberg TME fractions	0.70 (± 0.05)	0.62 (± 0.07)	0.69 (± 0.05)	0.56 (±0.08)	**0.55 (± 0.08)**
Alternative TME fractions	0.69 (± 0.11)	0.61 (± 0.03)	0.69 (±0.02)	0.52 (± 0.05)	**0.55 (± 0.07)**
Shape parameters	0.71 (± 0.10)	0.60 (± 0.08)	0.63 (± 0.03)	0.51 (±)0.04	0.48 (± 0.12)
Scale parameters	0.70 (±0.12)	0.63 (± 0.13)	0.61 (± 0.03)	0.50 (± 0.07)	0.54 (± 0.05)
Shape & scale parameters	0.73 (± 0.12)	0.59 (± 0.06)	0.62 (± 0.01)	0.51 (± 0.08)	0.49 (± 0.10)
Alternative cell fractions & shape parameters & scale parameters	0.77 (±0.12)	0.68 (± 0.08)	0.68 (± 0.03)	0.54 (± 0.05)	0.49 (± 0.10)

4.3.2. Predictions based on epithelial cell-type fractions

In general, prediction models trained on epithelial cell-type fractions performed well. The good performance aligns with earlier findings that tumors contain distinctive cell-type compositions. The model coefficients show that basal tumors exhibit higher numbers of basal cells and reduced numbers of other tumor cell types (alternative cell types $CK^{+}ER^{-}$ and $CK^{+}ER^{+}$ and Danenberg cell types $ER^{hi}CXCL12^{+}$, $CK^{+}CXCL12^{+}$, and $CK8{-}18^{hi}CXCL12^{hi}$) (Figure 21AB). Basal cells also contain higher fractions of proliferating cells in the epithelial tissue. Moreover, predictions with the alternative cell-type fractions show that basal samples contain higher fractions of $CK^{-}MHC^{+}$. The combination of cell types that affect the prediction of basal tumors is in accordance with the associations of individual cell-type fractions presented in section 4.1.

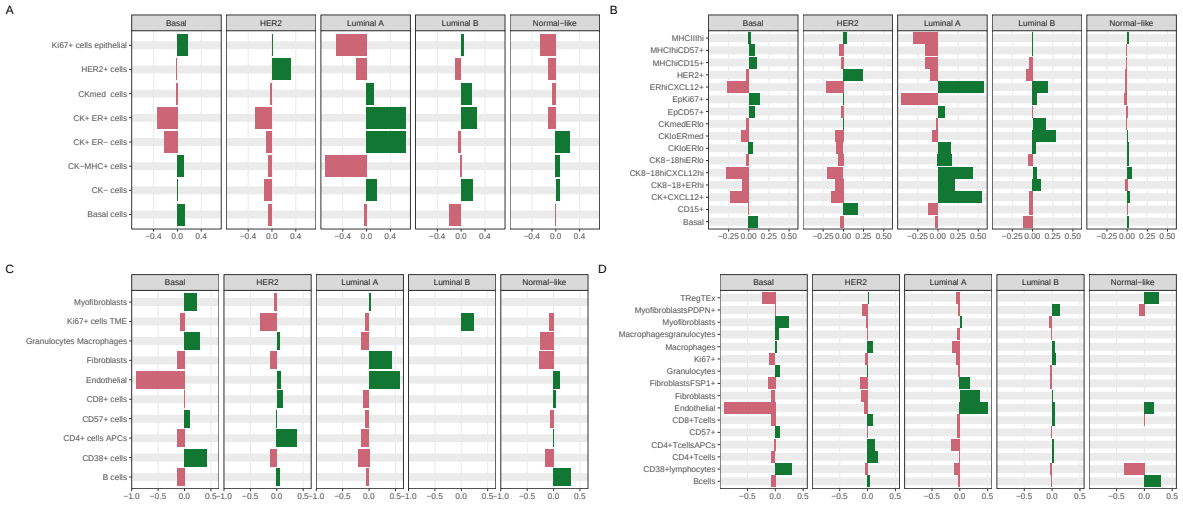

Figure 21: Prediction model coefficients of prediction models trained on the alternative epithelial cell fractions (A), Danenberg epithelial cell fractions (B), alternative TME cell fractions (C), and Danenberg TME cell fractions (D).

HER2-enriched tumors also exhibit a distinguishing tumor cell-type composition. The tumors contain larger fractions of HER2$^+$ cells and lower fractions of other tumor cell types (alternative cell types CK$^+$ER$^+$ cells and Danenber cell types ER^{hi}CXCL12$^+$ cells, CK8−18^{hi}CXCL12^{hi},CK8−18$^+$ER^{hi} and CK$^+$CXCL12$^+$). The combination of cell types that affect HER2-enriched tumor predictions aligns with univariate associations.

The performances of luminal A, luminal B, and normal-like tumors were significantly worse than for basal and HER-2 enriched tumors. Model coefficients indicate that the epithelial cell-type compositions of the subtypes are more alike. The prediction of luminal A tumors is mostly guided by the presence of all tumor cells except for basal, CK$^-$MHC$^+$ and HER2$^+$ cells. Luminal A samples also contain lower fractions of epithelial Ki67$^+$ cells and CK$^-$MHC$^+$ cells.

The prediction performance of luminal B tumors was poor because the fractions do not distinguish between luminal A and luminal B tumors. Luminal B tumors only differ from luminal A samples by higher epithelial Ki67$^+$ fractions and lower fractions of tumor cells without ER protein expression (alternative cell type CK$^+$ER$^-$ and Danenberg cell types CK8−18^{hi}ER^{lo}, CK$^+$CXCL12$^+$, CD15$^+$ cells).

The performance of normal-like tumor predictions was very poor and the AUC value is similar to a random guess (AUC-value = 0.5). In section 4.1, we showed that normal-like tumors are associated with very few epithelial cell types. Therefore, most normal-like samples are misclassified due to the lack of distinctive cell types.

4.3.3. Predictions based On TME cell-type fractions

Predictions based on TME cell-type fractions did not achieve the good performance of predictions based on epithelial cell-type fractions, except for luminal A tumors. The assessment of individual associations between TME cell types and luminal A tumors already found that many distinctive TME cell-type fractions characterize luminal A tumors. The model coefficients confirm a TME composition with an abundance of fibroblasts and endothelial cells and an absence of most other TME cell types (Figure 21CD).

The model coefficients further indicate that basal samples contain higher fractions of granulocytes & macrophages, CD38$^+$ cells and myofibroblasts, and lower fractions of endothelial cells ((Figure 21CD)). In contrast to all other cell-type fractions, the association between myofibroblast and basal tumors was not significant (Figure 14).

The predictions of HER2-enriched tumors had a moderate performance. Fractions of CD4$^+$ cells and TME Ki67$^+$ cells affect the prediction most (Figure 21C), but individual fractions of both cell types are not significantly associated with HER2-enriched tumors.

The model poorly predicted luminal B and normal-like tumors from the TME cell-type fractions. Individual associations already pointed out that few TME cell types are associated with both subtypes. In the prediction model trained on alternative cell-type fractions, only fractions of TME Ki67$^+$ cells affect

the prediction (Figure 21C). The Danenberg cell-type feature coefficients approximate zero, indicating that fractions have minimal effects (Figure 21D).

Although normal-like predictions based on the TME cell-type fractions had the best overall performance, the performance remained poor.

4.3.4. Predictions based on spatial relationships

Comparing all Weibull parameters over samples reveals that the spatial features exhibit considerable sparsity. The sparsity arises because samples do not contain all cell-type combinations or in some cases, have insufficient cell occurrences to estimate Weibull parameters robustly.

A 5-fold evaluation of the prediction performance was conducted to assess the impact of sparse features on the prediction model. The model was trained twenty times, adding five percent additional sparse features to the training dataset with each iteration. The model performances indicate that basal tumor predictions improve significantly for larger dataset sizes (Figure 22). For the remaining predictions, all features are included.

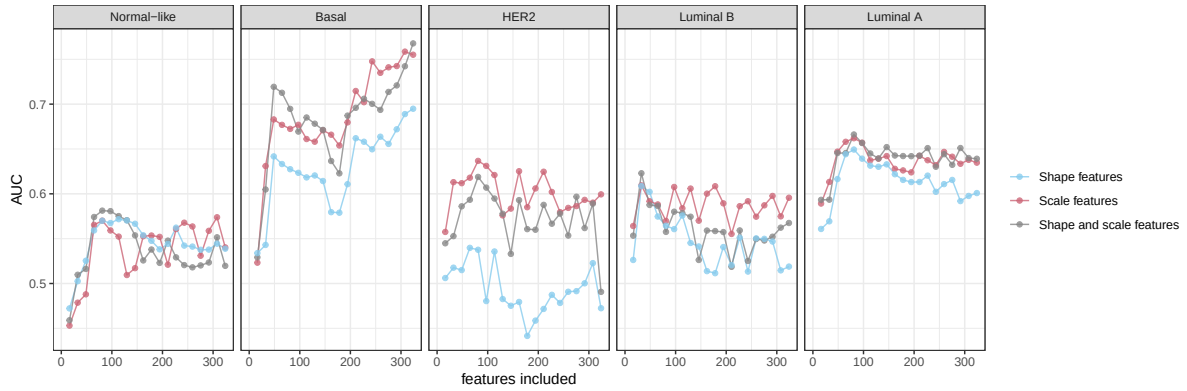

Figure 22: Feature coefficients of prediction models trained on varying spatial feature set sizes. Features are sorted by the number of zeros they contain and sparse features are incrementally added to the model.

The performance of prediction models trained on spatial features is significantly lower than those trained on epithelial cell-type fractions, but similar to the models trained on TME cell-type fractions. Basal samples were still predicted reasonably well, and the coefficients indicate that the prediction is based on various spatial relationships (Figure 23). Basal tumors are characterized by distance distributions with higher shape parameters for the combinations from Myofibroblasts to basal cells and $CD38^+$ to $CD38^+$ cells. Furthermore, the coefficients indicate that basal tumors are characterized by distance distributions with higher scale parameters for the combinations from TME $Ki67^+$ cells to epithelial $Ki67^+$ cells, endothelial to myofibroblasts, and CK^+ER^+ cells to granulocytes & macrophages. Finally, basal tumors are characterized by distance distributions with smaller scale parameters for the combinations from granulocytes & macrophages to CK^- cells, granulocytes & macrophages to $CD38^+$ cells, fibroblasts to basal cells, $CD48^+$ cells to myofibroblasts and $CD38^+$ cells to granulocytes & macrophages.

The prediction performance of HER2-enriched and luminal A tumors was moderate. Most spatial relationships associated with predicting the HER2-enriched subtypes contain $HER2^+$ cells as reference or target cell type. The probability increases for distance distributions with higher shapes or lower scales, both corresponding to a tighter packing of cells. Predictions of luminal A tumors are based on many spatial relationships between tumor cell types. for example higher shapes for the combinations from CK^{med} cells to CK^+ER^- cells, CK^+ER^+ cells to CK^- cells, and CK^- cells to CK^+ ER^- cells increase the probability of being a luminal A tumor. Additionally, the probability increases with higher shape values for myofibroblasts and endothelial cells to fibroblasts. Moreover, the probability increases with higher scales for combinations from TME $Ki67^+$ cells to epithelial $Ki67^+$ cells, $HER2^+$ cells to $HER2^+$ cells, and $CD38^+$ cells to granulocytes & macrophages. Finally, the distance distributions of CK+ ER+ cells to Basal cells and Basal cells to CK^- cells are characterized by lower shapes in luminal A tumors.

Predictions of luminal B and normal-like tumors were poor, and spatial features have poor predictive power for these subtypes.

Figure 23: Feature coefficients of prediction models trained on spatial features.

4.3.5. Predictions based on all features

Finally, we predicted subtypes based on both fractions and spatial features to assess how both feature sets collectively affect predictions and characterize subtypes. The models perform better than the models trained solely on spatial features and TME fractions but do not those trained on tumor cell-type fractions.

Basal tumor predictions are based on the presence of basal cells and granulocytes & macrophages, the absence of endothelial cells, $CK^{+}ER^{+}$ cells, and $CK^{+}ER^{-}$ cells. Moreover, several spatial relationships have a large effect on the model. The model selects similar cell types and spatial relationships as the models trained on the separate feature sets (Figure 24). The spatial relationships of Ki67+ cells epithelial to granulocytes & macrophages and CK−MHC+ cells to fibroblasts are the only features not selected by the individual models. The positive coefficients indicate that both combinations have higher shapes in basal samples.

Figure 24: Feature coefficients of prediction models trained on epithelial cell-type fractions, TME cell-type fractions, and spatial features.

HER2-enriched tumors are predicted based on the presence of $HER2^+$ cells and spatial relationships between $HER2^+$ cells and CK^- cells, $CD57^+$ cells and endothelial cells. The coefficients indicate that $HER2^+$ cells are more tightly packed to $CD57^+$ cells, CK^- cells, endothelial cells, and other $HER2^+$ cells because large shapes parameters and smaller scale parameters increase the probability of being a HER2-enriched tumor.

Luminal A subtype predictions are associated with the presence of fibroblasts, endothelial cells, CK^+ER^+ cells and CK^+ER^- cells. These features were also selected by the prediction models based on cell-type fractions. Similarly, those spatial relationships selected in the models trained on only spatial features are also found significant here.

Finally, it remains hard to predict both luminal B and normal-like tumors, and few features uniquely characterize these tumors.

4.3.6. Section Summary

Logistic regression prediction models are a valuable tool for quantitatively assessing feature sets' predictive power. In our analysis, we found that predictions based on cell-type fractions consistently outperform predictions based on spatial relationships.

It is essential to emphasize that breast cancer subtypes are primarily classified based on the molecular properties of tumor cells and by definition tumor cell-type fractions have strong predictive power. Consequently, subtype predictions relying on the presence of specific tumor cell types are more accurate than those based on spatial relationships alone.

Across all feature sets, the prediction performance for luminal B and normal-like tumors remains notably low. In Section 4.1, we demonstrated that basal, HER2-enriched, and luminal A tumors are associated with a wide variety of distinct epithelial and TME cell types. In contrast, luminal B and normal-like tumors exhibit associations with a limited number of cell-type fractions. Luminal B tumors are similar to luminal A tumors, while normal-like tumors closely resemble healthy breast tissue. Importantly, the subtyping of breast cancer in clinical settings often involves additional morphological features not adequately represented in cell-type fractions or spatial relationships. This discrepancy elucidates the lower prediction performances observed for luminal B and normal-like tumors.

4.4. Spatial Characterization of Breast Cancer Subtypes

We started the characterization of breast cancer subtypes by evaluating the abundance of cell types, revealing significant differences in the composition of the epithelium and TME. The intrinsic breast cancer subtypes were predicted well based on the presence of specific cell types. At the same time, Weibull parameters representing spatial relationships could not distinguish tumor subtypes with the same precision. The spatial relationship analysis does not aim to find the best descriptive features for breast cancer subtypes. Instead, it aims to uncover informative aspects about the tissue organization providing insights into how cells interact and cooperate. Spatial relationships are associated with a tumor subtype if the shape, scale, or both parameters are significantly associated with the subtype.

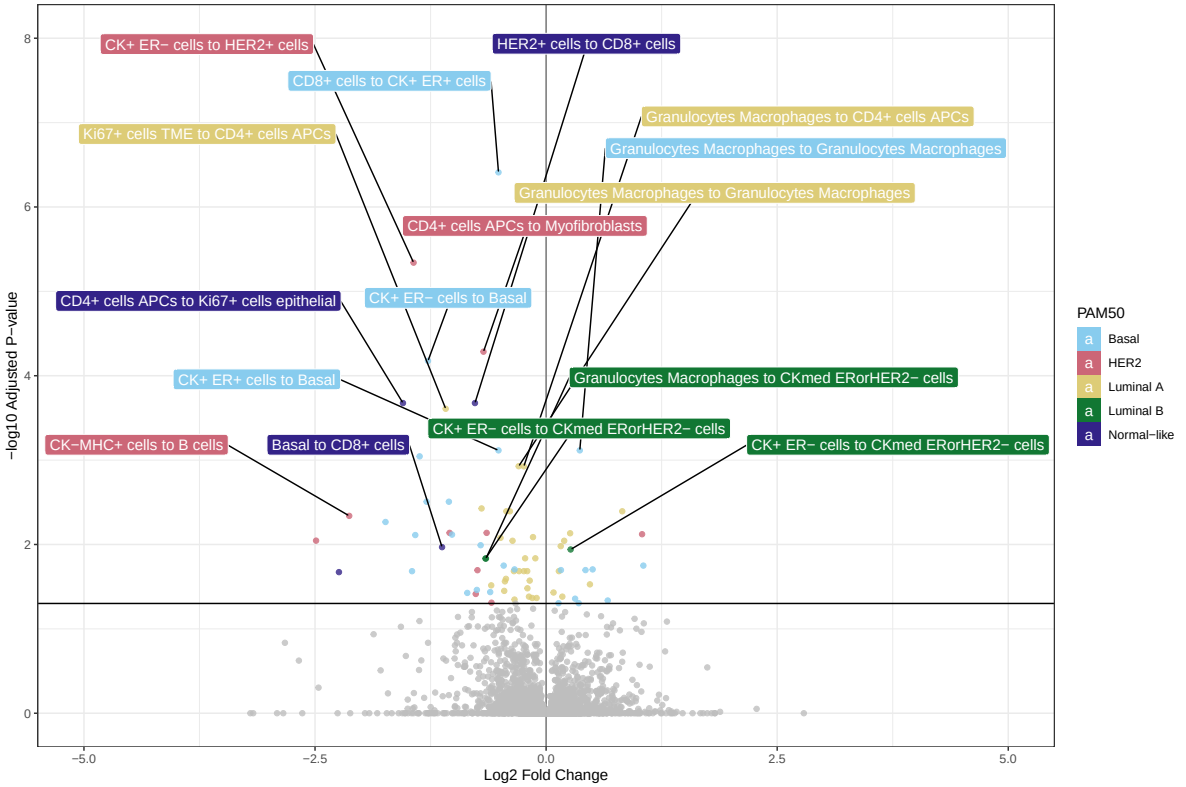

Figure 25: Volcano plot of p-values indicating the association strength between the spatial features and PAM50 subtypes. The labels indicate the three most significant associations per subtype.

As mentioned before, cell-type fractions are correlated with spatial relationships. The correlation is assessed using Pearson correlation coefficients for the Weibull parameters of associated features and reference and target cell-type fractions. The coefficients of the shape parameters and cell-type fractions are plotted against the coefficients of the scale parameters and cell-type fractions to identify the effect for each spatial relationship.

For spatial relationships where cells get closer if cell-type fraction increases, shapes are positively correlated and scales are negatively correlated with fractions. These spatial relationships are located in the bottom right quadrant of the plot (Figure 26). The opposite correlations are found for spatial relationships where distances increase if cell-type fractions increase. We refer to this correlation as an inverse effect and spatial relationships subject to this effect are located in the top left quadrant (Figure 26).

Finally, if the correlations of Weibull parameters with the target and reference cell-type fraction are not significant, spatial relationships do not depend on cell-type abundance (Figure 27).

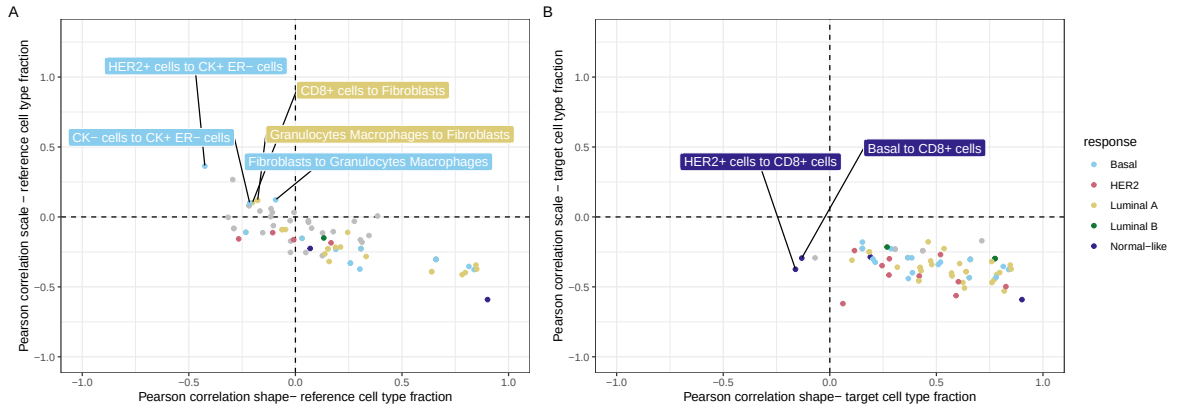

Figure 26: Correlations between Weibull parameters and reference cell-type fractions (A) and correlations between Weibull parameters and target cell-type fractions (B). Colors indicate the associated subtype and points are marked grey if no significant correlation exists. Points in the top left quadrant correspond to spatial relationships for which distances increase when the number of cells increases.

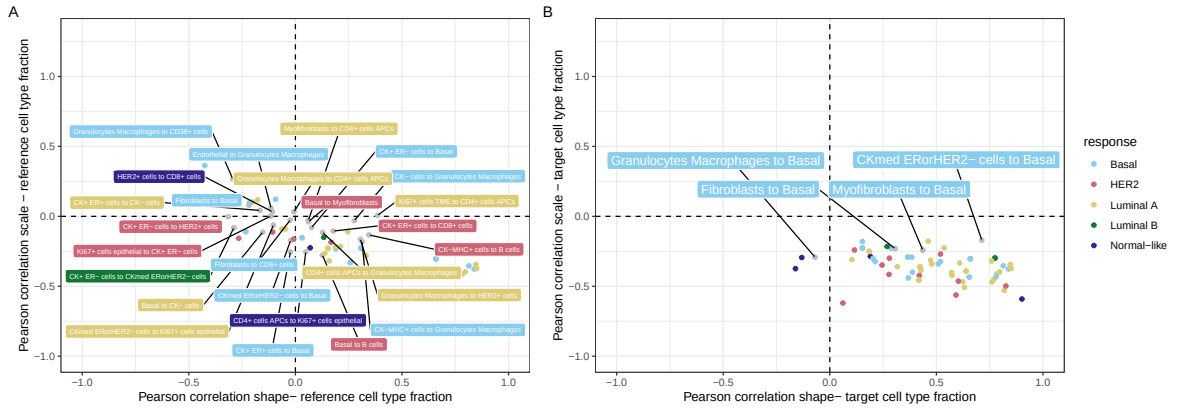

Figure 27: Correlations between Weibull parameters and reference cell-type fractions (A) and correlations between Weibull parameters and target cell-type fractions (B). Colors indicate the associated subtype. Points in the bottom right quadrant correspond to spatial relationships for which distances decrease when the number of cells increases. Grey points are labeled and correspond to spatial relationships not significantly correlated to cell-type fractions.

4.4.1. Spatial characterization of basal tumors

Basal tumors are associated with 21 spatial relationships. The tumors are characterized by six combinations with granulocytes & macrophages as target cell type (Fisher's exact test, p-value = 0.000012), and seven combinations with basal cells as target cell type (Fisher's exact test, p-value = 0.0000254). Most spatial relationships are characterized by higher shapes or lower scales in basal tumors, corresponding to distance distributions with smaller medians and sharper peaks.

All 1-NN distance distributions of combinations with basal cells are characterized by smaller scales and larger shapes in basal tumors. Therefore, fibroblasts, myofibroblasts, granulocytes & macrophages, CK^+ER^+ cells, CK^+ER^- cells and CK^{med} cells are tightly packed to basal cells in basal tumors. An example of the distance distribution is given for CK^+ER^- cells and basal cells (Figure 28). CK^+ER^- cells are less prevalent in basal tumors, but if present, close to basal cells.

Similarly, fibroblasts, epithelial $Ki67^+$ cells, $CD38^+$ lymphocytes, CK^- cells, CK^-MHC^+ cells and endothelial cells are near granulocytes & macrophages in basal tumors. An example is shown for endothelial cells and granulocytes & macrophages (Figure 29). Again the reference cell type is less prevalent in basal tumors, but if endothelial cells are present, they are near granulocytes & macrophages. Finally, granulocytes & macrophages are also densely packed with themselves.

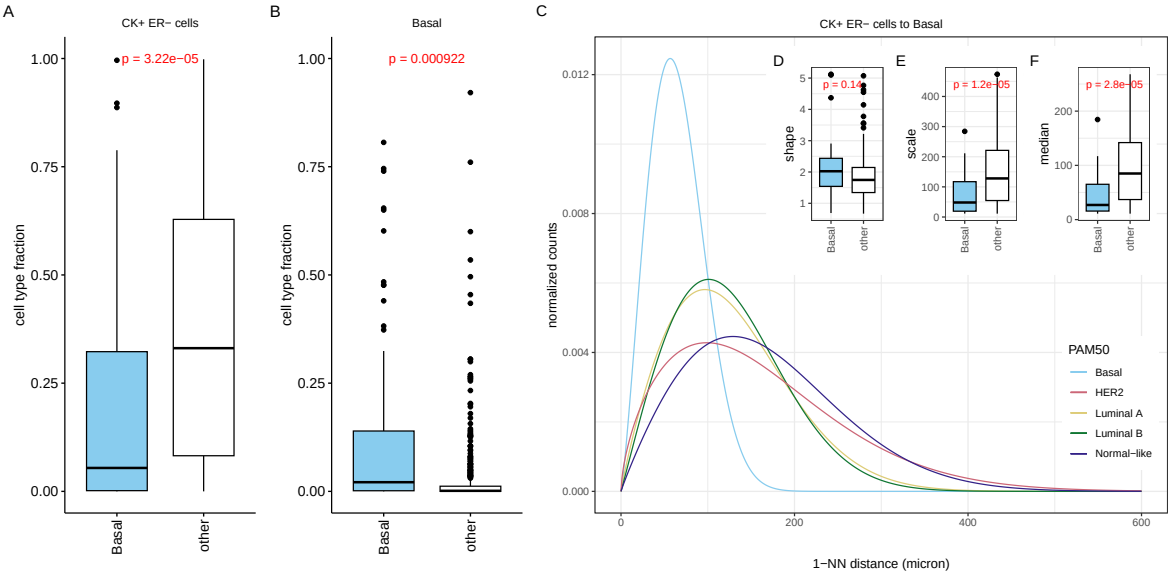

Figure 28: Fractions of $CK^{-}ER^{+}$ cells (A) and basal cells (B) in basal tumors; Weibull curves fitted on the distributions of 1-NN distances from $CK^{-}ER^{+}$ cells to basal cells for all tumor subtypes (C); shape parameters are similar for basal tumors and all other samples (D), while the scale parameter is significantly smaller (E). The median distances between the cell types are smaller (F).

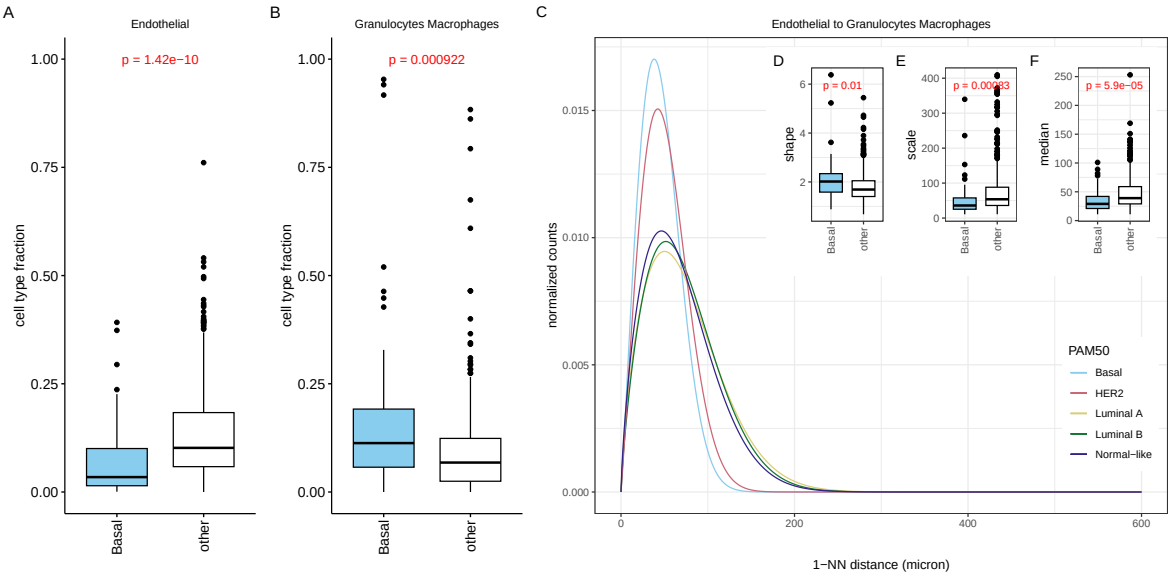

Figure 29: Fractions of endothelial cells (A) and granulocytes & macrophages (B) in basal tumors; Weibull curves fitted on the distributions of 1-NN distances from endothelial cells to granulocytes & macrophages for all tumor subtypes (C); shape parameters are significantly larger for basal tumors (D) and scale parameters are significantly smaller (E). The median distances between the cell types are also smaller (F).

For most combinations with basal cells and granulocytes & macrophages, the Weibull parameters are correlated with cell-type fractions. As both cell types are abundant in basal tumors, this likely explains their tight packing to other cell types. The Weibull parameters of spatial relationships from fibroblasts and CK^{med} cells to basal cells, however, are not correlated with reference and target cell fractions (Figure 27B). Therefore, CK^{med} cells and fibroblasts are always tightly packed with basal cells regardless of the number of cells.

Larger shapes and smaller scales in basal tumors do not characterize three cell-type combinations. CK^{-} cells and $CK^{+}ER^{-}$ cells (Figure 30), and CK^{med} cells and $CK^{+}Er^{-}$ cells (Figure 31) are further apart

in basal tumors. Both combinations include CK^+ER^- cells that are less abundant in basal tumors which likely explains why the distances are larger. Finally, the spatial relationship of $CD8^+$ cells to CK^+ER^+ cells is characterized by smaller shapes in basal tumors. The median distance between both cell types is similar, but the 1-NN distance distribution has a lower peak and more variance (Figure 31).

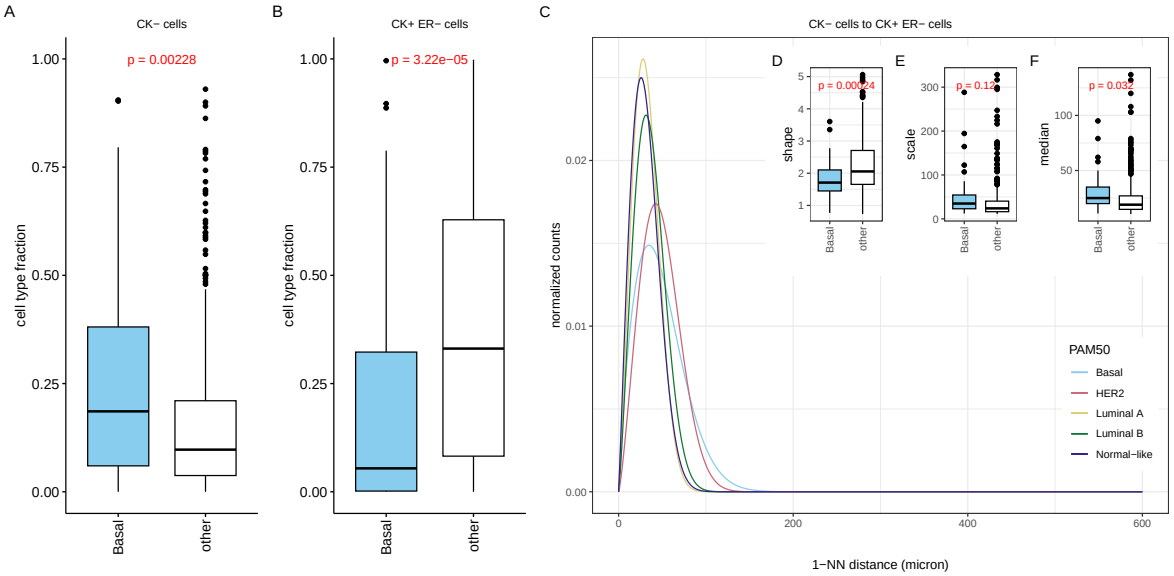

Figure 30: Fractions of CK^- cells (A) and CK^+ER^- cells (B) in basal tumors; Weibull curves fitted on the distributions of 1-NN distances from CL^- cells to CK^+ER^- cells for all tumor subtypes (C); shape parameters are smaller for basal tumors (D) while scale parameters are similar (E). The median distances between the cell types are larger (F).

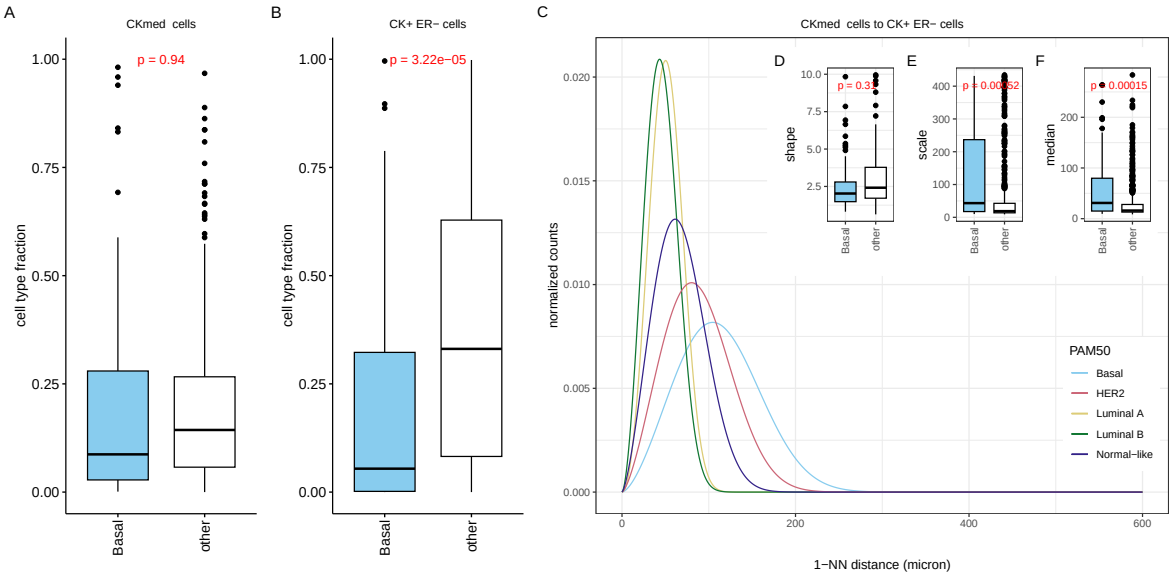

Figure 31: Fractions of $CD8^+$ cells (A) and CK^+ER^+ cells (B) in basal tumors; Weibull curves fitted on the distributions of 1-NN distances from $CD8^+$ cells to CK^+ER^+ cells for all tumor subtypes (C); shape parameters are smaller (D) while scale parameters similar for basal tumors (E). The median distances between the cell types are also not different (F).

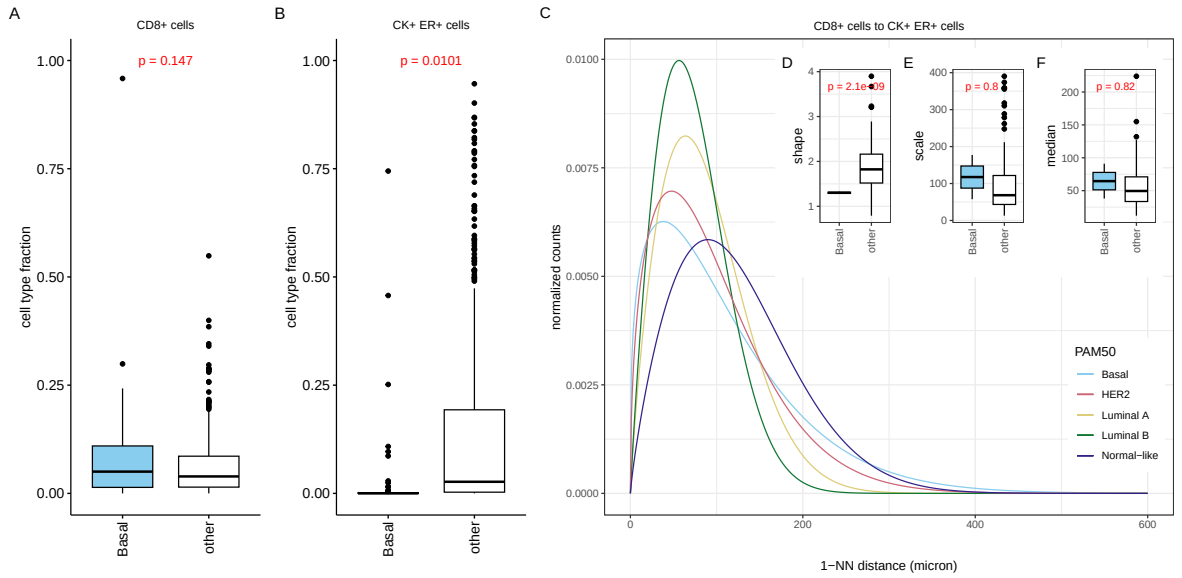

Figure 32: Fractions of HER2$^+$ cells (A) and CK$^+$ER$^-$ cells (B) in basal tumors; Weibull curves fitted on the distributions of 1-NN distances from HER2$^+$ cells to CK$^+$ER$^-$ cells for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller for basal tumors (E). The median distances between the cell types are not significantly different (F).

For the spatial relationships of fibroblasts to granulocytes & macrophages, HER2$^+$ cells to CK$^+$ER$^-$ cells and CK$^-$ cells to CK$^+$ER$^-$ cells, the shape parameter is negatively correlated with reference cell-type fractions and the scale parameter is positively correlated with reference cell-type fractions suggesting an inverse effect (Figure 26B). The effect is observed for HER2$^+$ cells to CK$^+$ER$^-$ cells (Figure 33). HER2$^+$ cell fractions are not different and CK$^+$ER$^-$ cell fractions are lower in basal tumors. Yet, the distance distribution has a smaller variance (Figure 33). Medians of the distance distributions are also not significantly different and this highlights the importance of evaluating the entire distance distribution instead of only comparing medians.

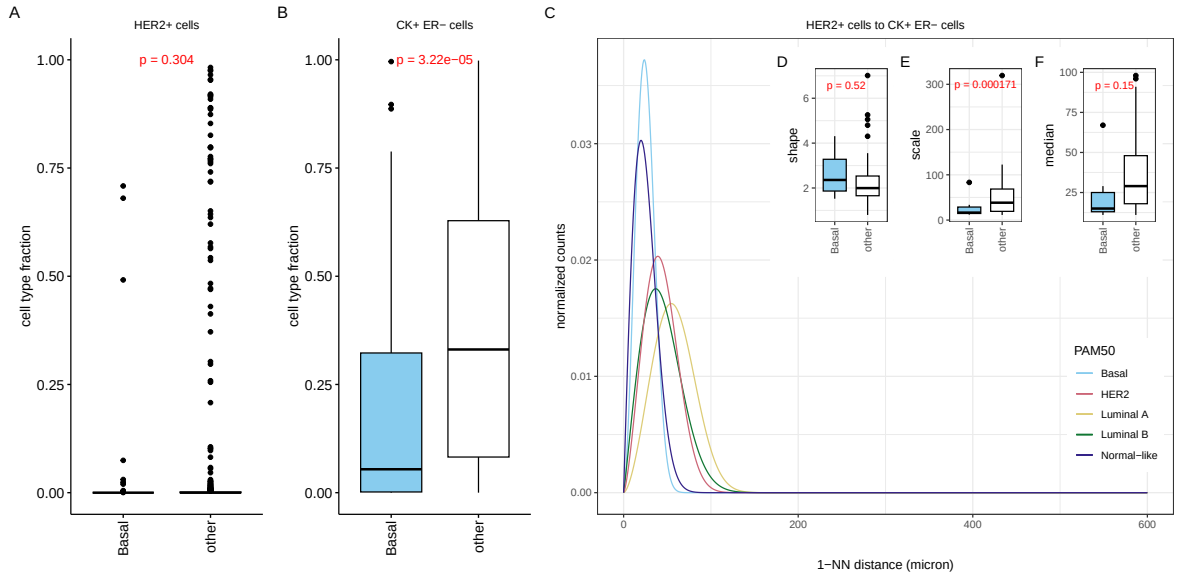

Figure 33: Fractions of HER2$^+$ cells (A) and CK$^+$ER$^-$ cells (B) in basal tumors; Weibull curves fitted on the distributions of 1-NN distances from HER2$^+$ cells to CK$^+$ER$^-$ cells for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller for basal tumors (E). The median distances between the cell types are not significantly different (F).

4.4.2. Spatial characterization of HER2-enriched tumors

HER2-enriched tumors are associated with ten spatial relationships; only the scale parameter is significantly different for all relationships. All distance distributions of HER2-enriched tumors are characterized by smaller scales, except for the distribution of CK^+ER^- cells to CK^+ER^+ cells, which has a higher scale corresponding to cells which are further apart (Figure 34).

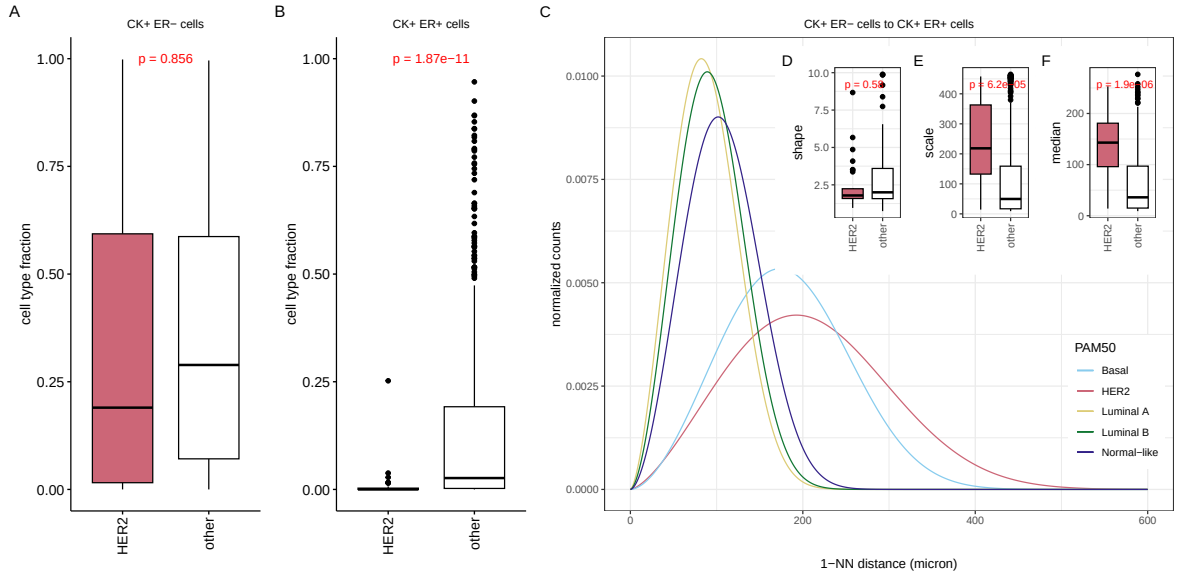

Figure 34: Fractions of CK^+ER^- cells (A) and CK^+ER^+ cells (B) in HER-enriched tumors; Weibull curves fitted on the distributions of 1-NN distances from CK^+ER^- cells to CK^+ER^+ cells for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly larger in HER2-enriched tumors (E). The median distances between the cell types are also larger (F).

The remaining associated spatial relationships correspond to cells in closer proximity, while the cell type fractions are not significantly different. Epithelial $Ki67^+$ cells and CK^+ER^- cells, for example, are tightly packed while both cell types do not occur more frequently in HER2-enriched samples (Figure 35). The combinations of basal cells to B cells, basal cells to myofibroblasts, $CD4^+$ T cells & APCs to myofibroblasts and CK^-MHC^+ cells to B cells have similar distributions in HER2-enriched samples and all involved cell types are not significantly more abundant in HER2-enriched tumors.

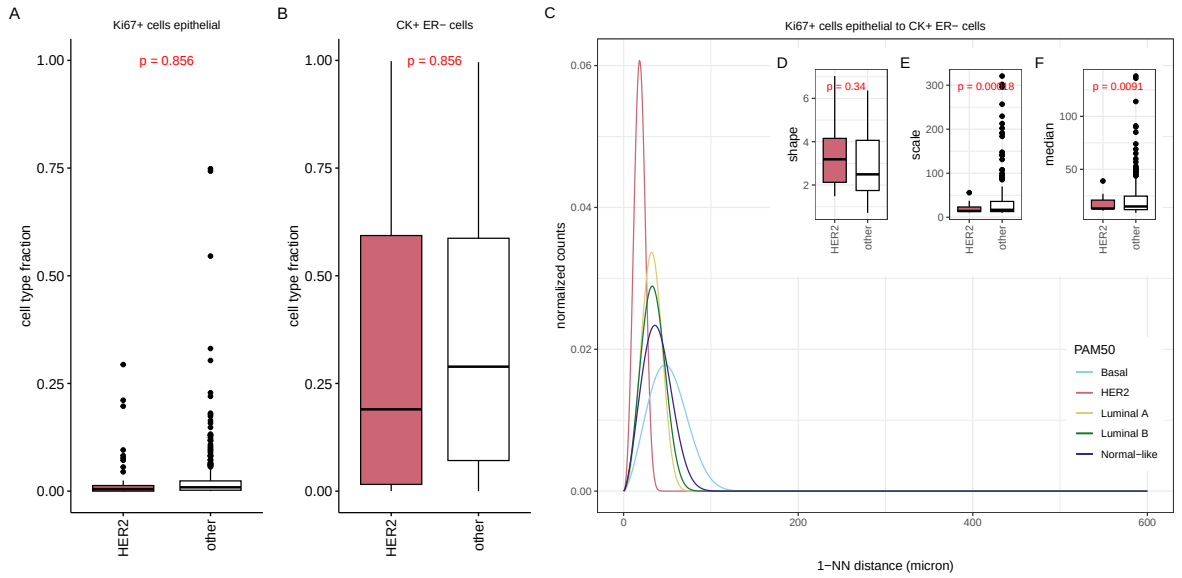

Figure 35: Fractions of epithelial Ki67$^+$ cells (A) and CK$^+$ER$^-$ cells (B) in HER-enriched tumors; Weibull curves fitted on the distributions of 1-NN distances from epithelial Ki67$^+$ cells to CK$^+$ER$^-$ cells for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller in HER2-enriched tumors (E). The median distances between the cell types are also smaller (F).

HER2-enriched tumors contain low numbers of CK$^+$ER$^+$ cells. Still, the cells that are present are close to myofibroblasts (Figure 36) and CD8$^+$ T cells (Figure 37). Furthermore, the spatial relationship from B cells to endothelial cells is characterized by smaller scales in HER2-enriched tumors, but the distribution medians are not significantly different (Figure 38).

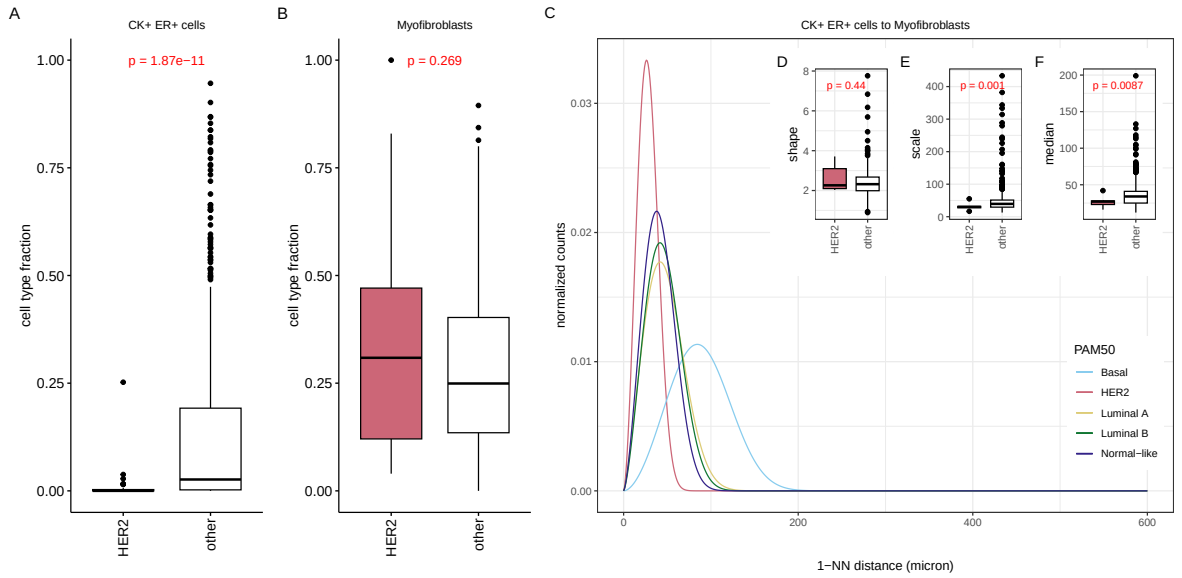

Figure 36: Fractions of CK$^+$ER$^+$ cells (A) and myofibroblasts (B) in HER-enriched tumors; Weibull curves fitted on the distributions of 1-NN distances from CK$^+$ER$^+$ cells to myofibroblasts for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller in HER2-enriched tumors (E). The median distances between the cell types are also smaller (F).

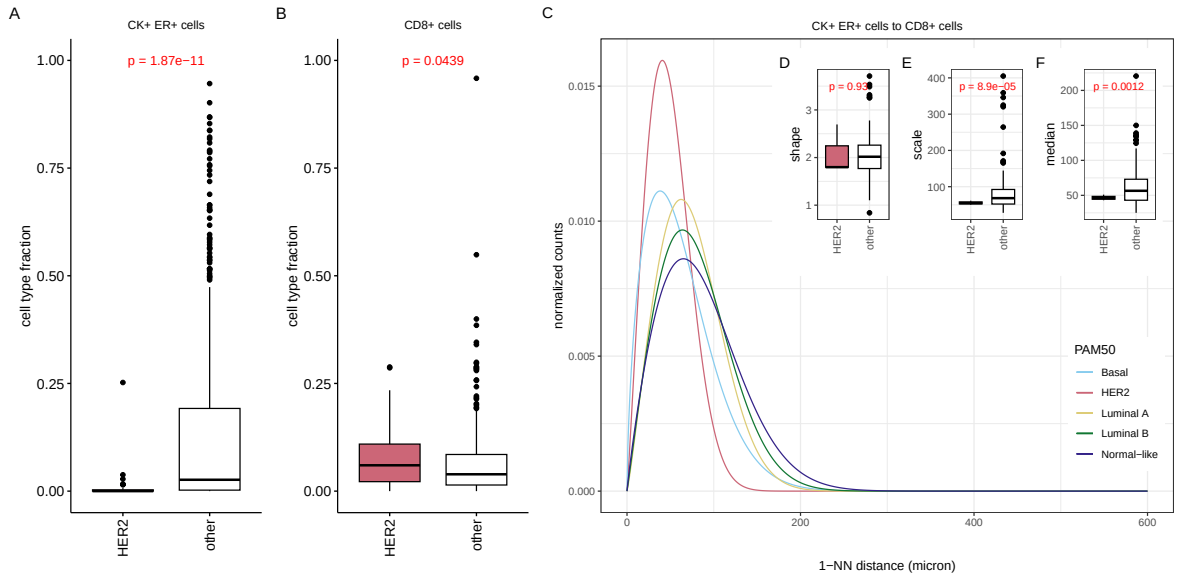

Figure 37: Fractions of CK^+ER^+ cells (A) and $CD8^+$ cells (B) in HER-enriched tumors; Weibull curves fitted on the distributions of 1-NN distances from CK^+ER^+ cells to $CD8^+$ cells for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller in HER2-enriched tumors (E). The median distances between the cell types are also smaller (F).

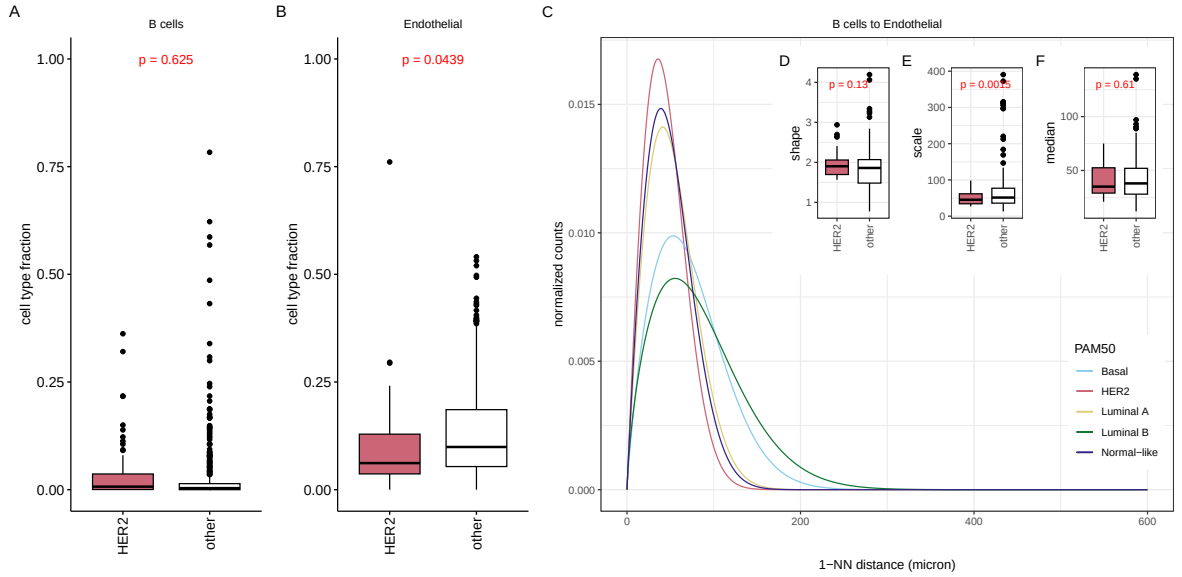

Figure 38: Fractions of B cells (A) and endothelial cells (B) in HER-enriched tumors; Weibull curves fitted on the distributions of 1-NN distances from B cells to endothelial cells for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller in HER2-enriched tumors (E). The median distances are also similar (F).

Finally, only one spatial relationship of $HER2^+$ cells characterizes HER2-enriched tumors. CK^+ER^- cells are densely packed with $HER2^+$ cells (Figure 39). This likely explains the smaller distances as $HER2^+$ cells are more abundant in HER2-enriched tumors.

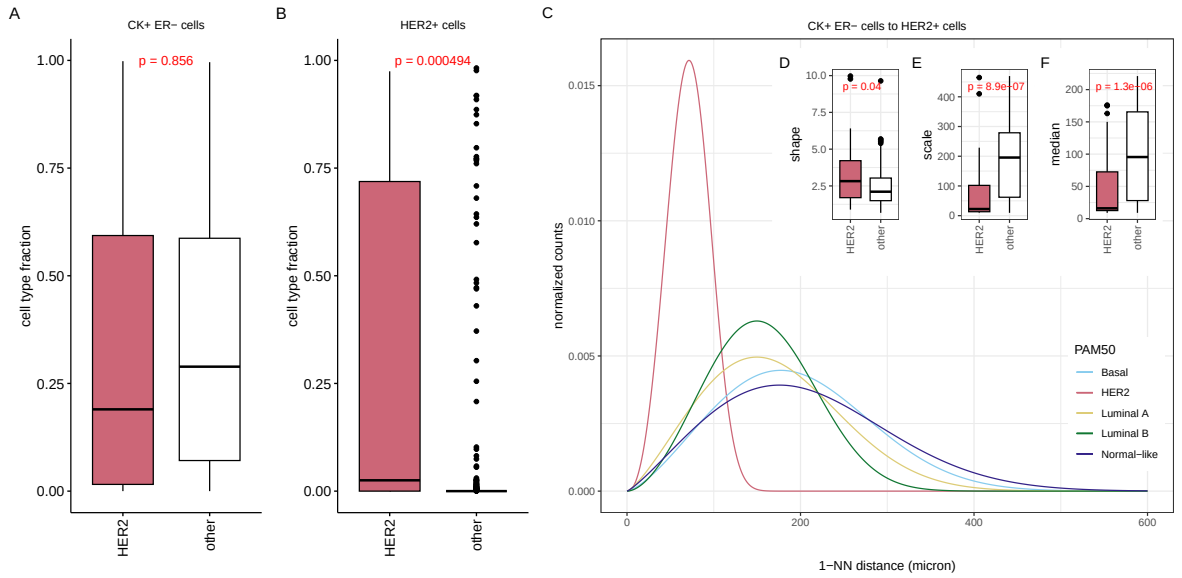

Figure 39: Fractions of $CK^{+}ER^{-}$ cells (A) and $HER2^{+}$ cells (B) in HER-enriched tumors; Weibull curves fitted on the distributions of 1-NN distances from $CK^{+}ER^{-}$ cells to $HER2^{+}$ cells for all tumor subtypes (C); shape parameters are significantly larger (D) while scale parameters are significantly smaller in HER2-enriched tumors (E). The median distances between the cell types are also smaller (F).

4.4.3. Spatial characterization of luminal A tumors

Luminal A tumors are associated with 28 spatial relationships. Tumors are characterized by five relationships to $CD4^{+}$ T cells & APCs (Fisher's exact test, p-value = 0.00067), five to fibroblasts (Fisher's exact test, p-value = 0.00013), and six to granulocytes & macrophages (Fisher's exact test, p-value = 0.000012).

All distance distributions with fibroblasts as the target type have smaller medians and variances and are characterized by larger shapes or smaller scales. An example is shown for endothelial cells to fibroblasts (Figure 40). Like endothelial cells, $CD8^{+}$ cells, CK^{-} cells, and granulocytes & macrophages are close to fibroblasts. Additionally, fibroblasts are also densely packed with themselves.

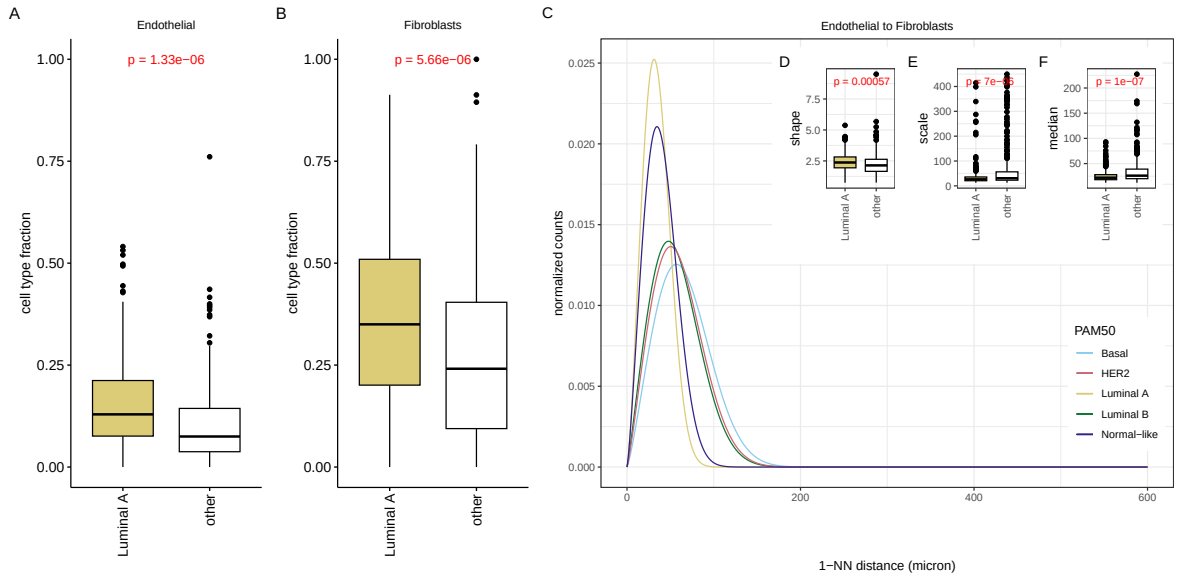

Figure 40: Fractions of endothelial cells (A) and fibroblasts (B) in luminal A tumors; Weibull curves fitted on the distributions of 1-NN distances from endothelial cells to fibroblasts for all tumor subtypes (C); shape parameters are significantly larger (D) while scale parameters are significantly smaller in luminal A tumors (E). The median distances between the cell types are also smaller (F).

Granulocytes & macrophages and CD4$^+$ T cells & APCs occur in low fractions throughout luminal A samples, and all associated spatial relationships are characterized by larger scales or smaller shapes corresponding to distance distributions with larger medians and variances. An example is shown for myofibroblasts to both granulocytes & macrophages (Figure 41) and CD4$^+$ T cells & APCs (Figure 42). In addition, epithelial Ki67$^+$ cells, CD48$^+$ cells, CD4$^+$ cells & APCs and CD8$^+$ T cells are further away from granulocytes & macrophages. Similarly, TME Ki67$^+$ cells, granulocytes & macrophages and CD8$^+$ T cells are further away from CD4$^+$ cells. The self-self relationships of granulocytes & macrophages and CD4$^+$ cells & APCs are also characterized by distance distributions with larger medians and variances.

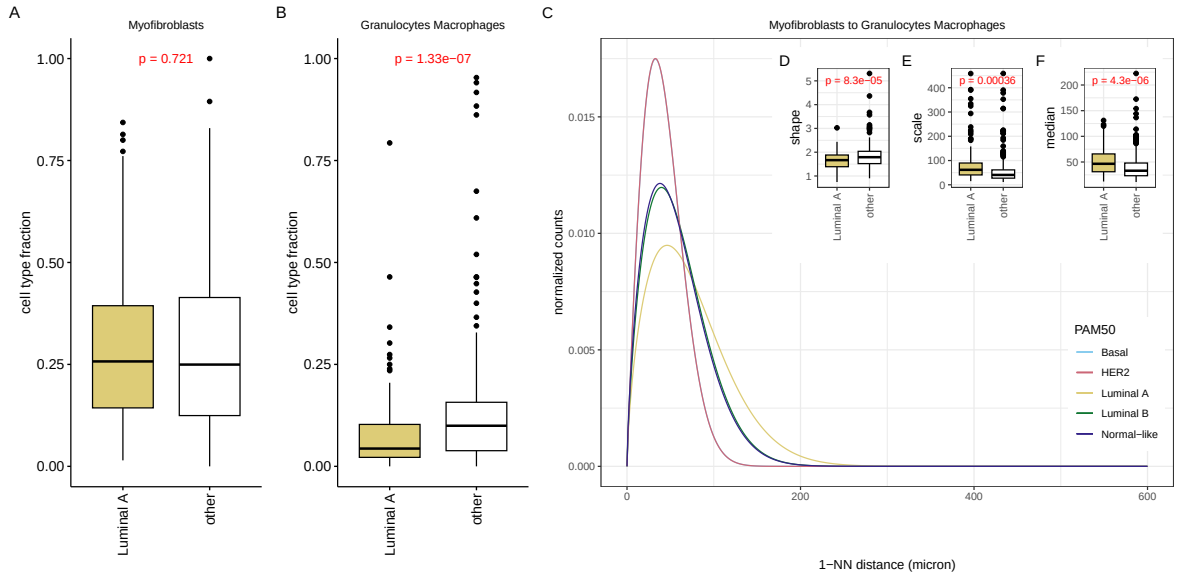

Figure 41: Fractions of myofibroblasts (A) and granulocytes & macrophages (B) in luminal A tumors; Weibull curves fitted on the distributions of 1-NN distances from myofibroblasts to granulocytes & macrophages for all tumor subtypes (C); shape parameters are significantly smaller (D) while scale parameters are significantly larger in luminal A tumors (E). The median distances between the cell types are also larger (F).

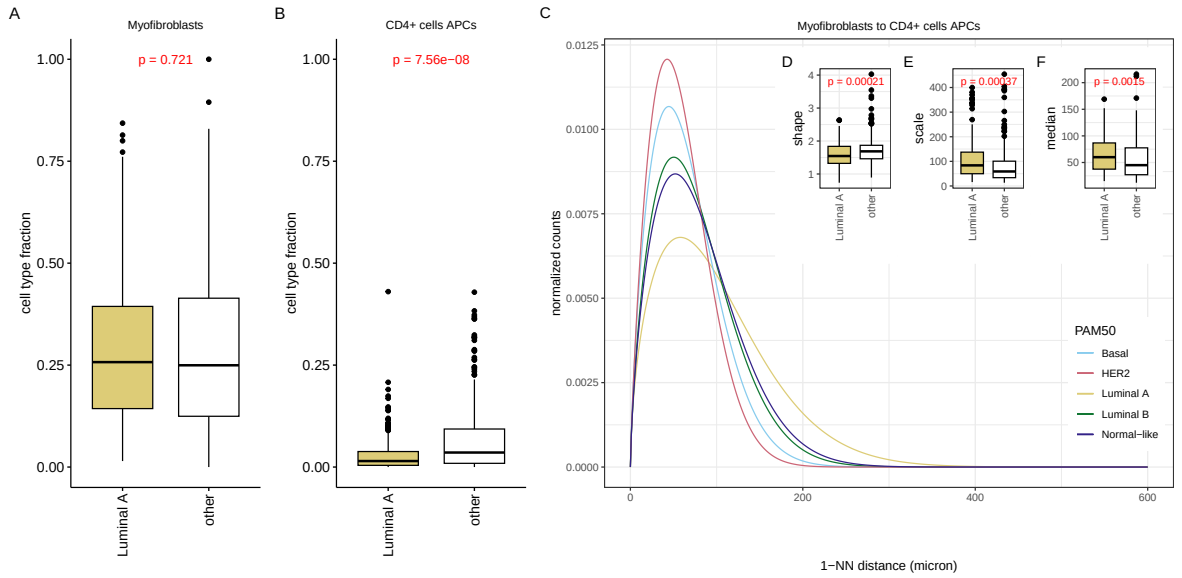

Figure 42: Fractions of myofibroblasts (A) and CD$^+$ cells & APCs (B) in luminal A tumors; Weibull curves fitted on the distributions of 1-NN distances from endothelial cells to fibroblasts for all tumor subtypes (C); shape parameters are significantly smaller (D) while scale parameters are significantly larger in luminal A tumors (E). The median distances between the cell types are also larger (F).

Luminal A tumors are associated with the spatial relationships of CD8$^+$ cells to fibroblasts and granulocytes & macrophages to fibroblasts. For both combinations, reference cell-type fractions are positively correlated with the scale parameters and negatively correlated with the shape parameters, suggesting an inverse effect (Figure 26). Moreover, medians and variances of the 1-NN distance distributions are smaller in luminal A tumors although they contain lower CD8$^+$ T cell and granulocytes & macrophages fractions (Figure 43 and Figure 44). Yet, the shape parameters of the spatial relationships are positively correlated with the fraction of fibroblasts, and the scale parameters are negatively correlated with the fraction. CD8$^+$ cells and granulocytes & macrophages are, therefore, tightly packed to fibroblasts regardless of their occurrence due to a high number of fibroblasts.

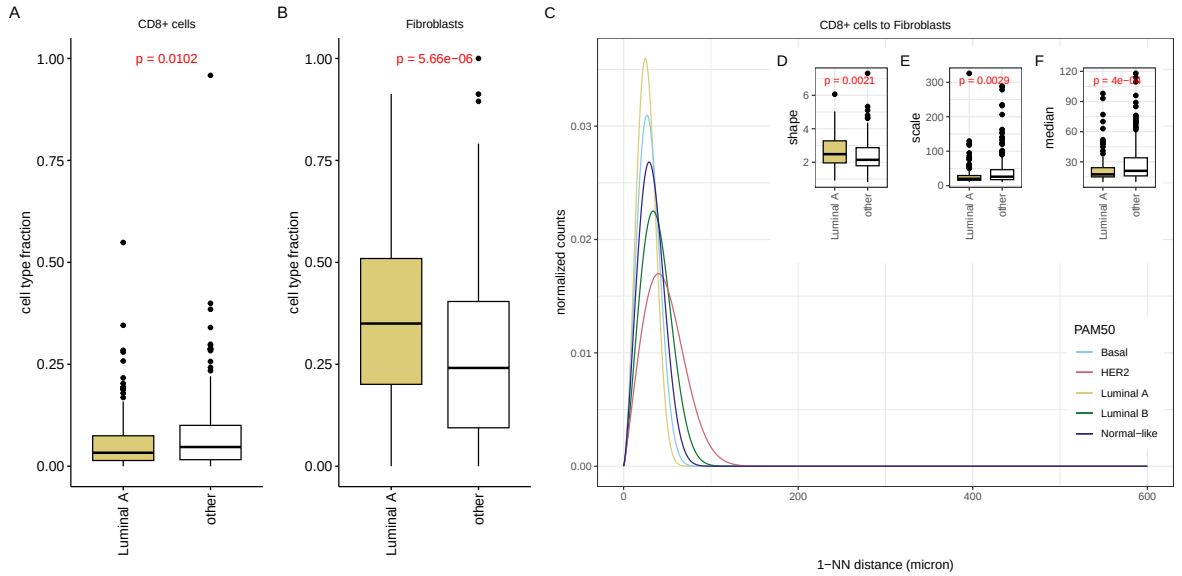

Figure 43: Fractions of CD8$^+$ cells (A) and fibroblasts (B) in luminal A tumors; Weibull curves fitted on the distributions of 1-NN distances from CD8$^+$ cells to fibroblasts for all tumor subtypes (C); shape parameters are significantly larger (D) while scale parameters are significantly smaller in luminal A tumors (E). The median distances between the cell types are also smaller (F).

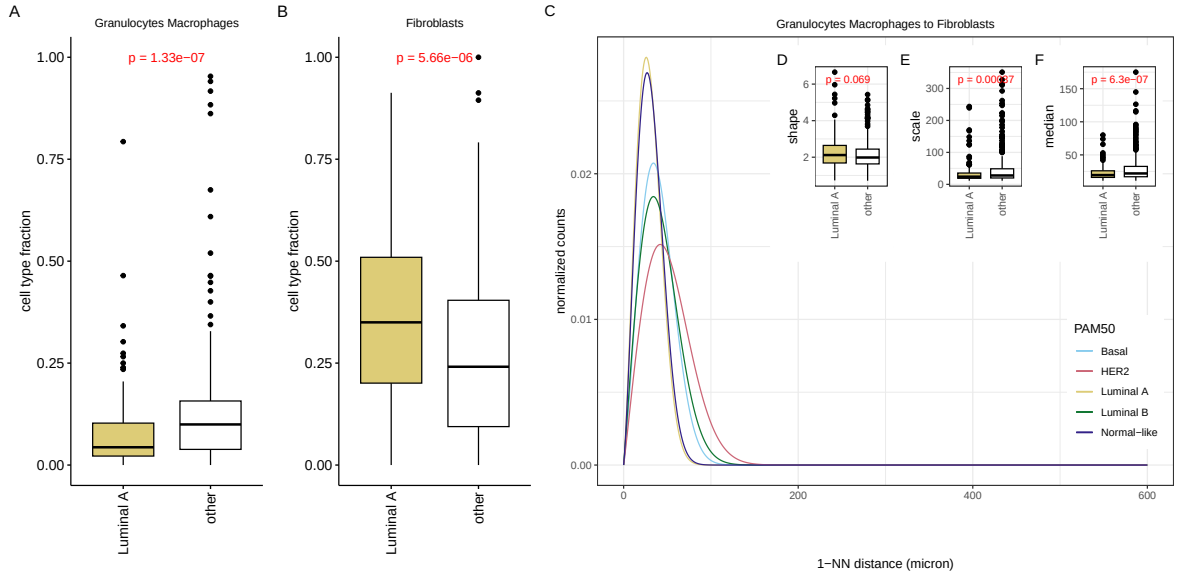

Figure 44: Fractions of granulocytes & macrophages (A) and fibroblasts (B) in luminal A tumors; Weibull curves fitted on the distributions of 1-NN distances from granulocytes & macrophages to fibroblasts for all tumor subtypes (C); shape parameters are similar (D) while scale parameters are significantly smaller in luminal A tumors (E). The median distances between the cell types are also smaller (F).

4.4.4. Spatial characterization of luminal B tumors

Luminal B tumors are associated with two spatial relationships: $CK^{+}ER^{-}$ cells to $CK^{med}ER^{-}$ cells and granulocytes & macrophages to $CK^{med}ER^{-}$ cells. No cell-type fractions are associated with luminal B samples. The 1-NN distance distributions of both combinations are characterized by smaller medians and variances in luminal B tumors (Figure 45 and 46).

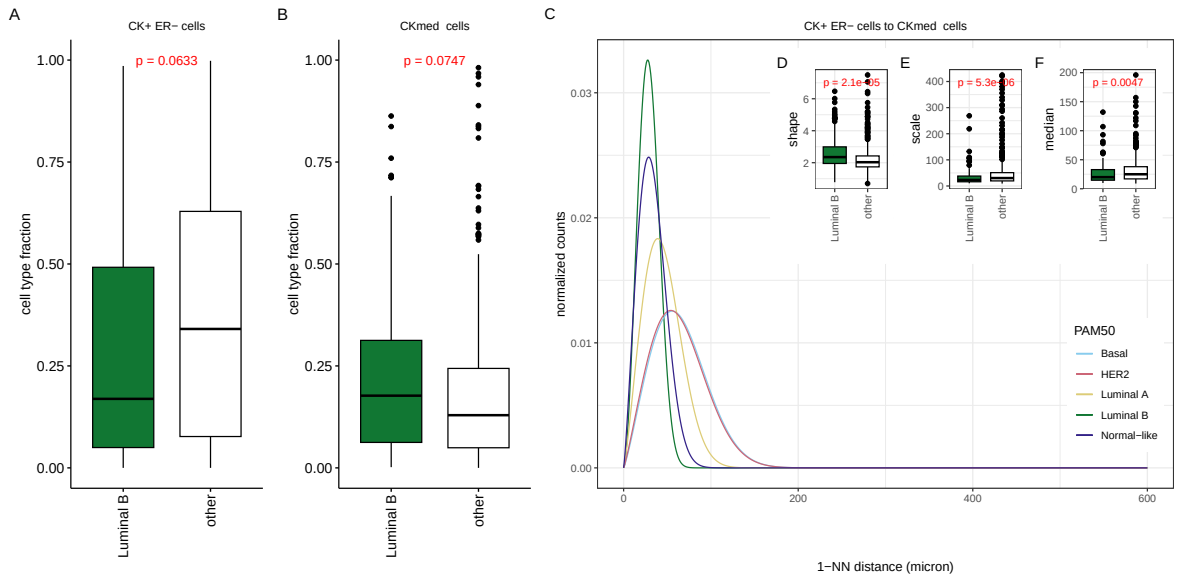

Figure 45: Fractions of $CK^{+}ER^{-}$ cells (A) and CK^{med} cells (B) in luminal B tumors; Weibull curves fitted on the distributions of 1-NN distances from $CK^{+}ER^{-}$ cells to CK^{med} cells for all tumor subtypes (C); shape parameters are significantly larger (D) while scale parameters are significantly smaller in luminal B tumors (E). The median distances between the cell types are also smaller (F).

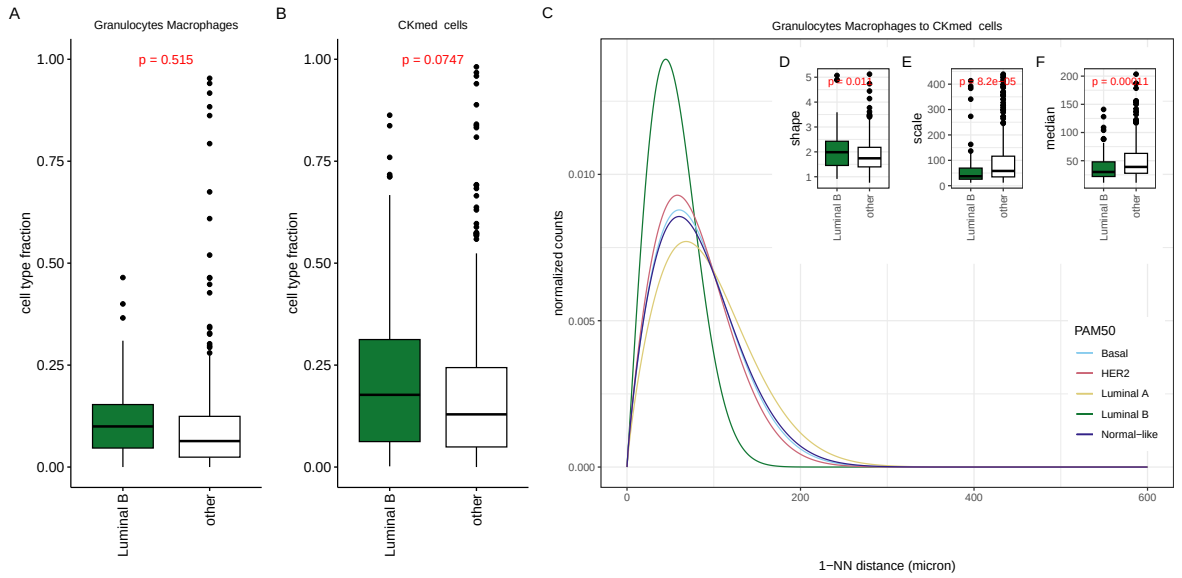

Figure 46: Fractions of granulocytes & macrophages (A) and CK^{med} cells (B) in luminal B tumors; Weibull curves fitted on the distributions of 1-NN distances from granulocytes& macrophages to CK^{med} cells for all tumor subtypes (C); shape parameters are significantly larger (D) while scale parameters are significantly smaller in luminal B tumors (E). The median distances between the cell types are also smaller (F).

4.4.5. Spatial characterization of normal-like tumors

Normal-like tumors are associated with four spatial relationships: basal cells to $CD8^+$ T cells (Figure 47), $CD4^+$ T cells & APCs to epithelial $Ki67^+$ cells (Figure 48), $HER2^+$ cells to $CD8^+$ T cells (Figure 49) and the self-self relationship of $CD8^+$ T cells (Figure 50). All relationships concern cell types without a significant abundance in normal-like tumors, even though the spatial relationships indicate that the cells are densely packed.

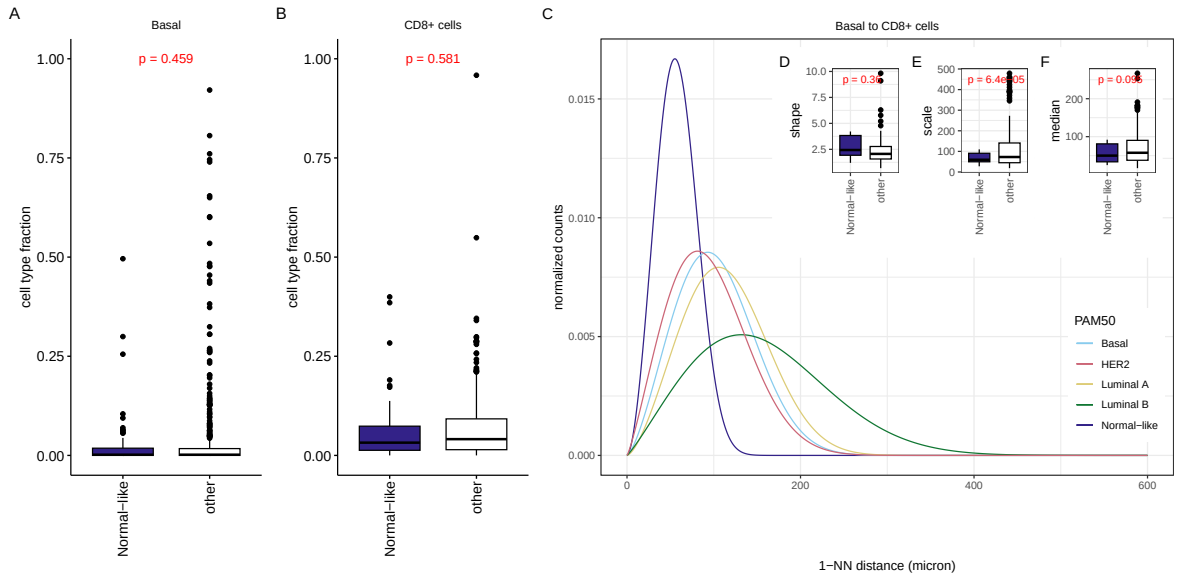

Figure 47: Fractions of basal cells (A) and $CD8^+$ cells (B) in normal-like tumors; Weibull curves fitted on the distributions of 1-NN distances from basal cells to $CD8^+$ cells for all tumor subtypes (C); shape parameters are not significantly different for normal-like tumors (D) while scale parameters are significantly smaller (E). The median distances between the cell types are also not smaller (F).

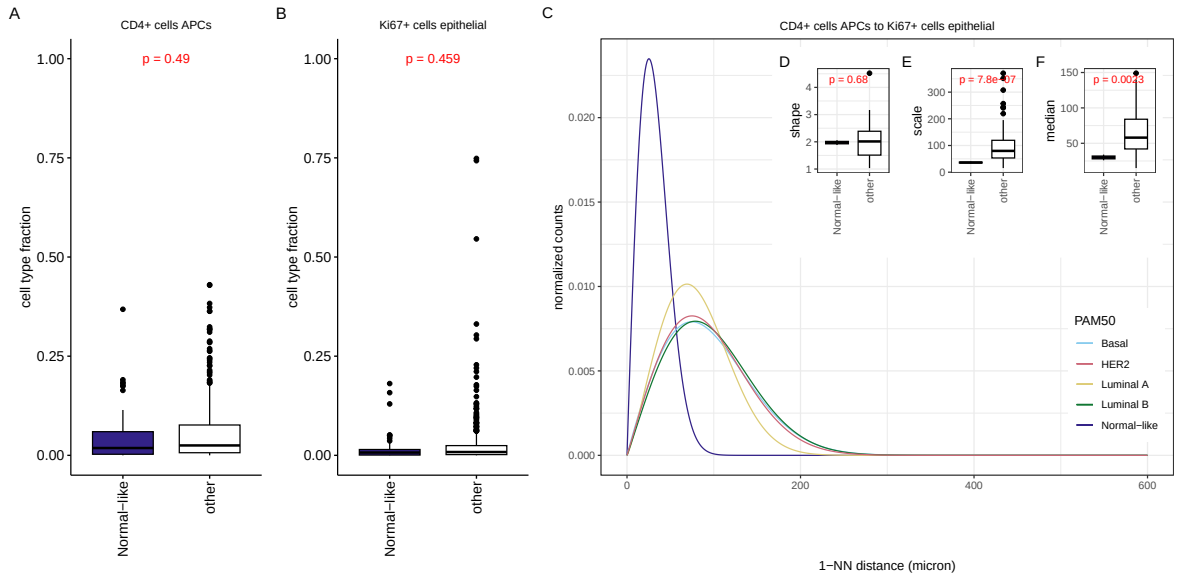

Figure 48: Fractions of CD4$^+$ cells & APCs (A) and epithelial Ki67$^+$ cells (B) in normal-like tumors; Weibull curves fitted on the distributions of 1-NN distances from CD4$^+$ cells & APCs to epithelial Ki67$^+$ cells for all tumor subtypes (C); shape parameters are not significantly different for normal-like tumors (D) while scale parameters are significantly smaller (E). The median distances are also smaller (F).

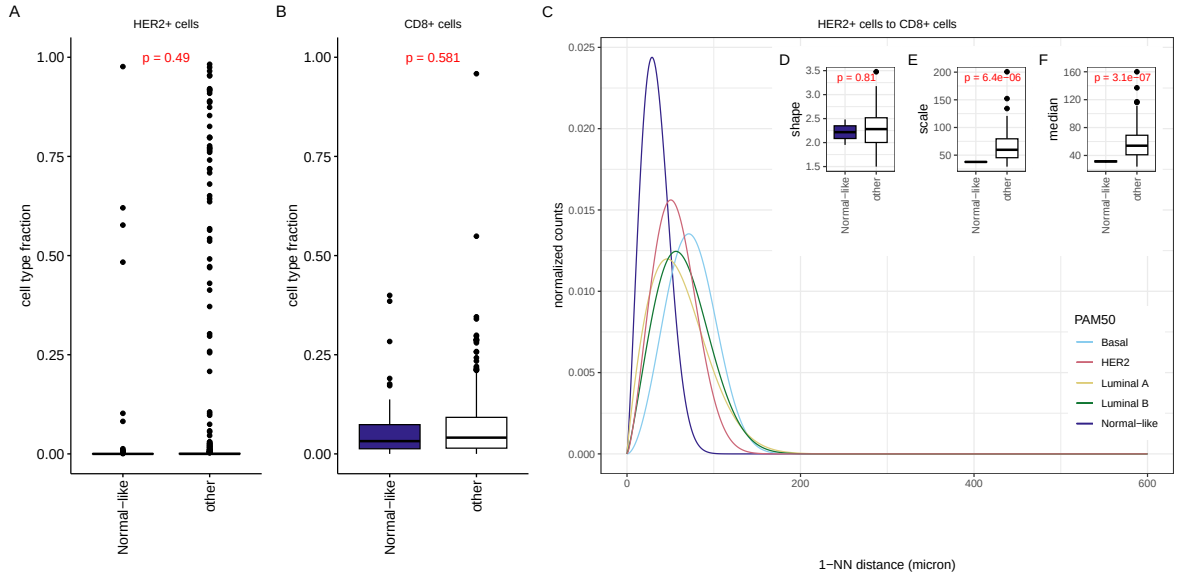

Figure 49: Fractions of HER2$^+$ cells (A) and CD8$^+$ cells (B) in normal-like tumors; Weibull curves fitted on the distributions of 1-NN distances from HER2$^+$ cells to CD8$^+$ cells for all tumor subtypes (C); shape parameters are not significantly different for normal-like tumors (D) while scale parameters are significantly smaller (E). The median distances between the cell types are also smaller (F).

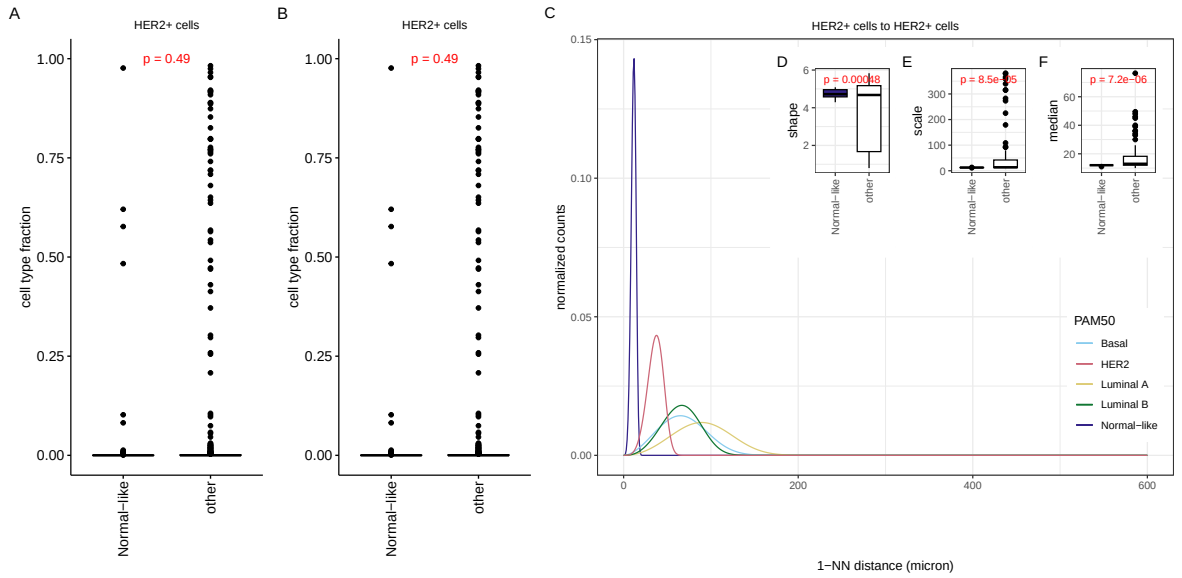

Figure 50: Fractions of HER2$^+$ cells (AB) in normal-like tumors; Weibull curves fitted on the distributions of 1-NN distances from HER2$^+$ cells to HER2$^+$ cells for all tumor subtypes (C); shape parameters are significantly large (D) while scale parameters are significantly smaller for normal-like tumors (E). The median distances between the cell types are also smaller (F)

The shape parameters of the spatial relationships HER2$^+$ cells to CD8$^+$ cells and basal cells to CD8$^+$ cells are negatively correlated with the basal cell fractions indicating an inverse effect (Figure 26). However, the shape parameters and CD8$^+$ cell fractions are both not significantly different in normal-like samples.

4.4.6. Section Summary

Using Weibull parameters to describe spatial relationships provided a means to compare subtypes and identify spatial relationships uniquely associated with each subtype. In this section, we addressed Research Question 3:

What cell types and spatial relationships are associated with breast cancer subtypes?

As previously demonstrated, basal, HER2-enriched, and luminal A samples exhibit a diverse array of distinct cell types. These abundant cell types, especially for basal and luminal A samples, also occurred in many unique spatial relationships. Most spatial relationships revealed a denser packing of abundant cell types. The observation is consistent with the findings in section 4.2 showing a correlation between cell-type abundance and spatial relationships.

The spatial relationships linked to basal tumors comprehensively depict the specific neighboring cells surrounding abundant basal cells and granulocytes & macrophages. Moreover, the spatial relationships also revealed that the limited presence of other tumor cell types such as CK$^-$, CK$^+$ER$^-$ and CK^{med} cells in basal tumors leads to relatively larger distances between them when present. The relationship between CK^{med} and CK$^+$ER$^-$ cells is also significantly associated with luminal A tumors but in this subtype, the cells are tighter packed. The spatial relationship of HER2$^+$ cells and CK$^+$ER$^-$ cells indicates that cells are densely packed while both cell types occur in lower numbers in basal tumors.

HER2-enriched tumors were rarely associated with spatial relationships involving HER2$^+$ cells suggesting that HER2$^+$ cells are not arranged differently in HER2-enriched tumors. Notably, HER2-enriched tumors were often characterized by densely packed cell types for which fractions were not significantly different or were even relatively lower. Only CK$^+$ER$^-$ cells and CK$^+$ER$^+$ cells are further apart in HER2-enriched tumors which is likely explained by low fractions of the latter cell type.

The spatial relationships characterizing luminal A tumors show the dense packing of many cell types with abundant fibroblasts and endothelial cells. Other TME cells such as CD4$^+$ cells & APCs and granulocytes & macrophages were more distanced in luminal A tumors consistent with their low abundance. The self-self relationship of granulocytes & macrophages stands out as it is characterized by large dis-

tances in luminal A tumors and a dense packing in basal tumors. The same is observed for epithelial $Ki67^+$ cells and macrophages & granulocytes, and $CD38^+$ cells and macrophages & granulocytes.

Similar to cell-type fractions, few spatial relationships were associated with luminal B and normal-like tumors. In luminal B tumors, CK^+ER^- cells and granulocytes & macrophages are densely packed with CK^{med} cells. Normal-like tumors are associated with the dense packing of basal cells and $CD8^+$ cells, $CD4^+$ cells & APCs and epithelial $Ki67^+$ cells, $HER2^+$ cells and $CD8^+$ cells, and $HER2^+$ cells to themselves. Fractions of all involved cell types were not significantly different in luminal B and normal-like tumors.

Characterizing spatial relationships between cell-type pairs shows that cells are organized in natural ways following their abundance or absence in subtypes. Moreover, we find many relationships independent of cell-type fractions or show an inverse effect where cell types are rare but densely packed together. Both demonstrate the additional value of considering spatial relationships alongside cell-type fractions.

4.5. Predicting Patient Survival

We assess the prognostic significance of cell-type fractions and spatial features by examining associations with hazard rates. The hazard rate represents the rate of mortality over time. We express the strength of these associations in terms of hazard ratios, which represent the change in hazard rate proportionate to the increase or decrease of a given variable.

The unique impact of significantly associated variables is visualized by categorizing patients into two groups based on the median. Essentially, the continuous variable is transformed into a discrete one, taking on a value of zero if it is lower than or equal to the median and one if it surpasses the median. This discretization process allows for a clearer understanding of the association between low and high values. We evaluate the difference between the resulting survival curves using the log-rank test.

Associations are initially evaluated across the entire patient cohort and subsequently computed separately for patients with ER-positive and ER-negative breast cancer, recognizing the importance of hormone receptor status in breast cancer prognosis.

4.5.1. Predicting patient survival

In our initial analysis, we evaluated associations with cell-type fractions, revealing that higher fractions of epithelial Ki67$^+$ cells, CD4$^+$ cells & APCs and CD57$^+$ cells are associated with a worse prognosis. In contrast, higher fractions of endothelial cells and fibroblasts are associated with a better prognosis (Figure 51A). The hazard ratios for Ki67$^+$ cells, CD4$^+$ cells and CD57$^+$ cells are exceptionally high. These associations are based on a few samples that contain high fractions, and the absolute hazard ratio should be interpreted cautiously. Yet, in the raw data, we do observe that samples with higher fractions have a lower survival probability.

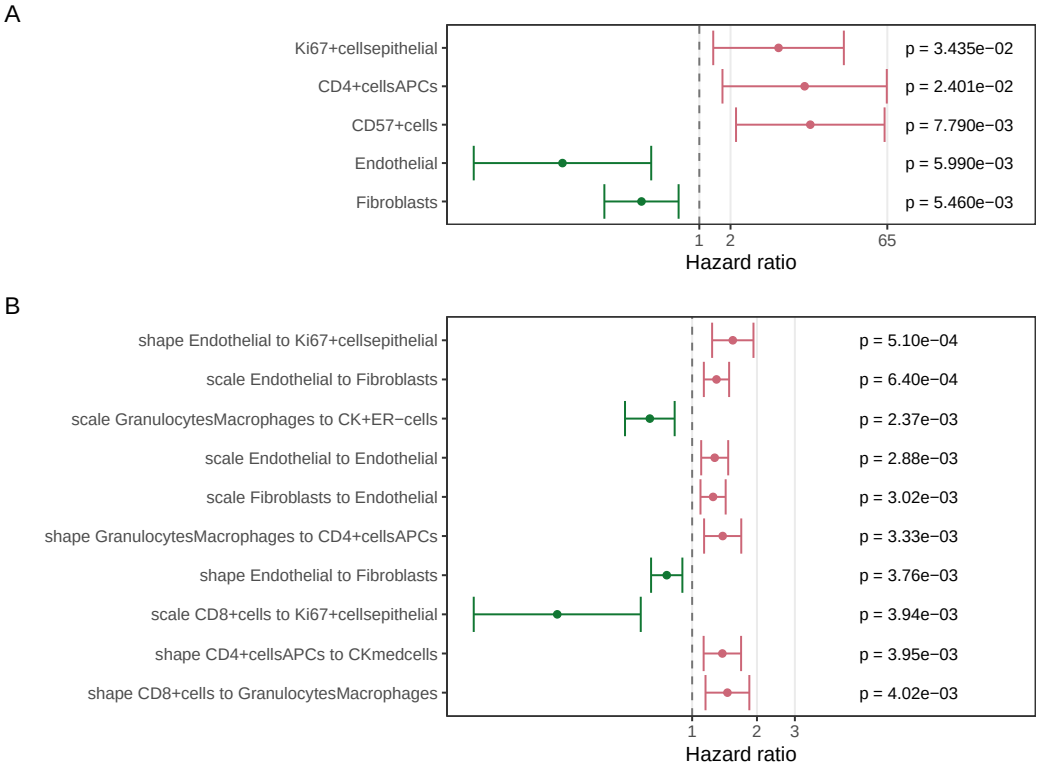

Figure 51: Hazard ratios of univariate Cox proportional hazard regression model with cell-type fractions (A) and spatial relationships (B). A positive hazard ratio indicates a worse prognosis and a negative hazard ratio indicates a protective effect. The p-value comes from testing the null hypothesis that this hazard ratio is 1, or that there is no difference in the relative risk of the event comparing individuals with varying levels of the variable. The x-axis is log-scaled.

For Ki67$^+$ cells, CD4$^+$ cells & APCs, and CD57$^+$ cells, the discretized variables do not have a significant association with the survival rates (Figure 52). The median splits patients into two groups with little difference in survival probability.

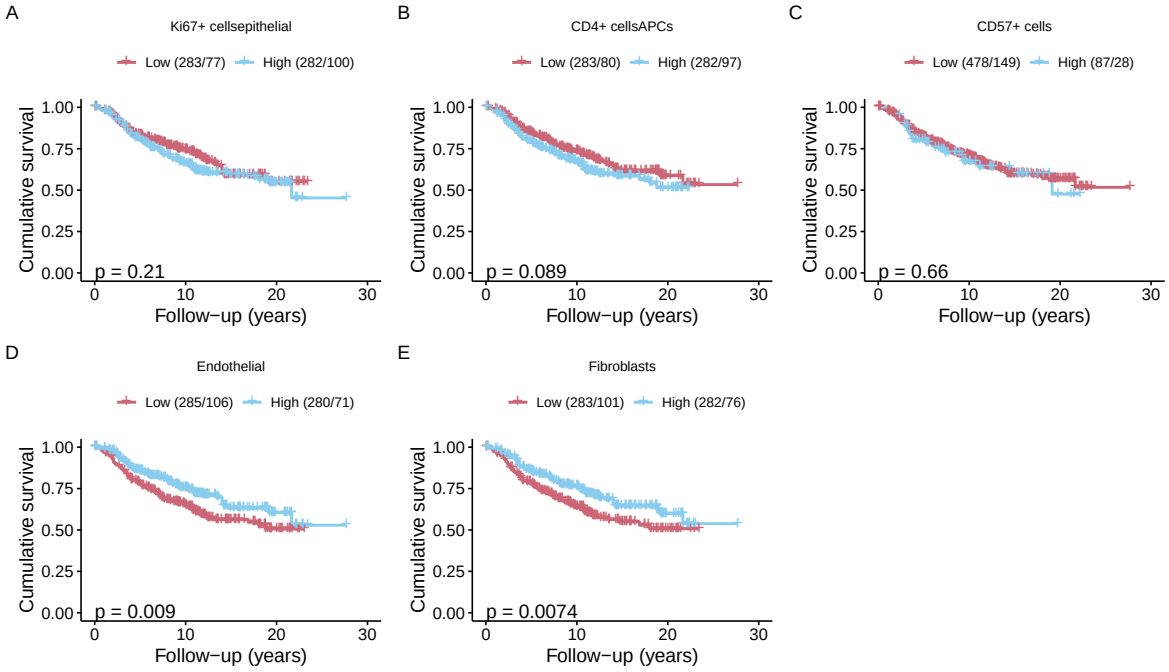

Figure 52: Predicted survival proportion over time for two risk groups (low: smaller or equal to the median, high: higher than the median) stratified by cell-type fractions significantly associated with the hazard rates of all samples. Predictions are adjusted for HER2 status. The legends indicate the number of patients and the number of events. All depicted p-values are from log-rank tests.

Additionally, we find that 31 spatial relationships are associated with survival outcomes. Figure 51B shows the ten most significant associations.

Higher shapes for endothelial cells to epithelial Ki67$^+$ cells (Figure 53A), granulocytes & macrophages to CD4$^+$ cells & APCs (Figure 53F), CD4$^+$ cells & APCs to CK^{med} cells (Figure 53I) and CD8$^+$ cells to granulocytes & macrophages (Figure 53J) are associated with a worse prognosis. The discrete variable is not significantly associated with the hazard rate for the final two combinations. Low scales for the combinations granulocytes & macrophages to CK$^+$ER$^-$ (Figure 53C) and CD8$^+$ cells to epithelial Ki67$^+$ cells (Figure 53H) are associated with a poor prognosis. Both high shapes and low scales correspond to densely packed cells.

In contrast, for some spatial relationships, the survival probability increases if cells are in closer proximity. Endothelial cells, for example, are associated with improved survival if they are close to fibroblasts (Figure 53BEG) and themselves (Figure 53D). Therefore, improved survival is associated with an abundance of both cell types and with both cell types being near each other.

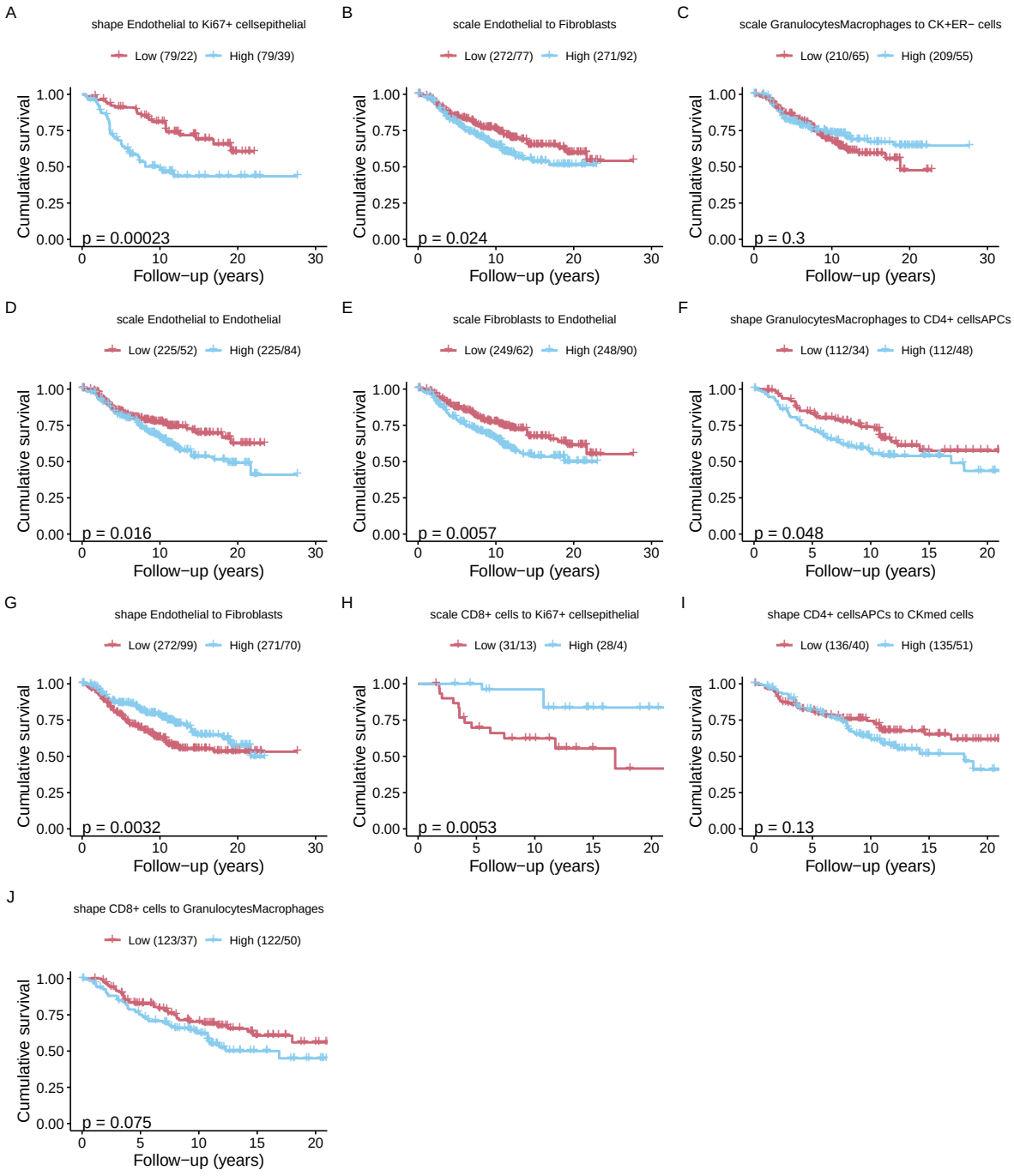

Figure 53: Predicted survival proportion over time for two risk groups (low: $\leq$ median, high: $>$ median) stratified by spatial relationships significantly associated with the hazard rates of all samples. Predictions are adjusted for HER2 status. The legends indicate the number of patients and the number of events. All depicted p-values are from log-rank tests.

4.5.2. Predicting patient survival per ER status

In general ER-negative patients (HER2-enriched and basal tumors) have a worse prognosis than ER-positive patients (luminal A, luminal B, and normal-like tumors). Therefore, associations between cell-type fractions and spatial features were also assessed separately for ER-negative and ER-positive tumors. No features were significantly associated with the survival probability of ER-negative samples. In contrast, in ER-positive samples, higher fractions of fibroblasts are associated with a higher survival probability, while higher fractions of $CD4^+$ T cells & APCs are associated with worse prognosis (Figure 54A). The discretized variables are not significantly associated with survival rates (Figure 55).

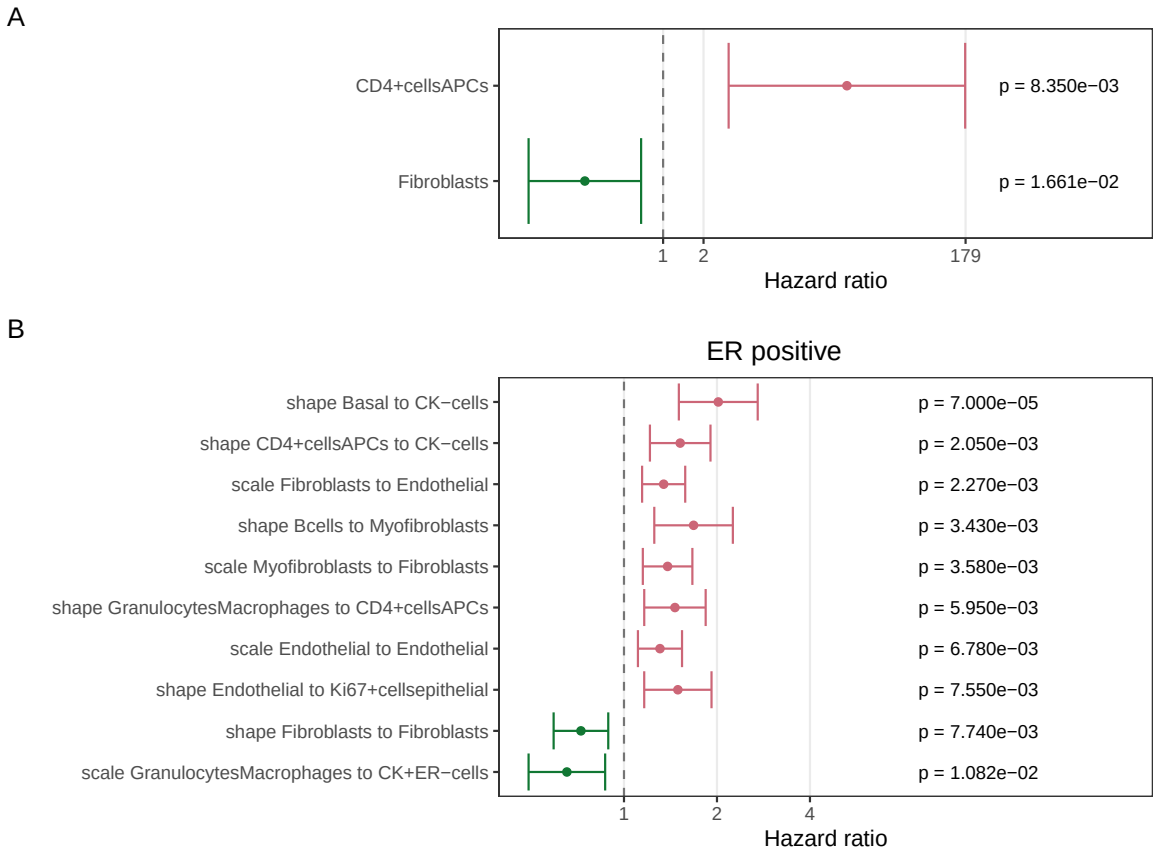

Figure 54: Hazard ratios of univariate Cox proportional hazard regression model for all ER-positive patients with cell-type fractions (A) and spatial relationships (B). A positive hazard ratio indicates a worse prognosis and a negative hazard ratio indicates a protective effect. The p-value comes from testing the null hypothesis that this hazard ratio is 1, or that there is no difference in the relative risk of the event comparing individuals with varying levels of the variable. The x-axis is log-scaled.

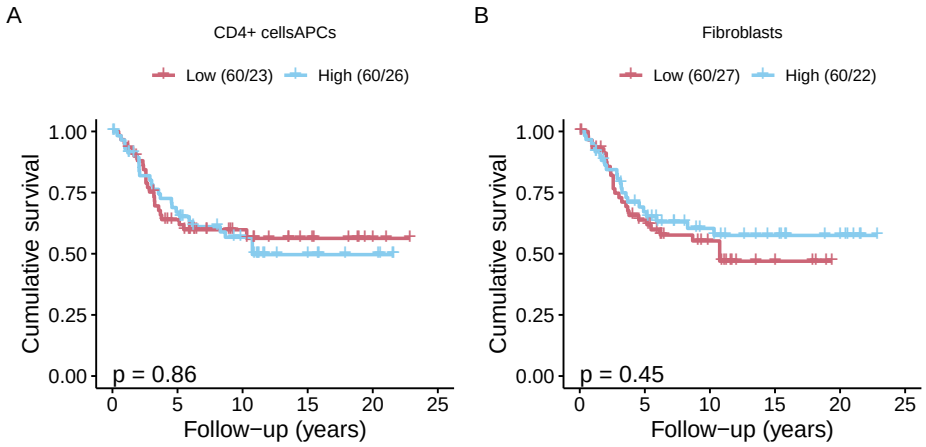

Figure 55: Predicted survival proportion over time for two risk groups (low: $\leq$ median, high: $>$ median) stratified by cell-type fractions significantly associated with the hazard rates of all ER-positive samples. Predictions are adjusted for HER2 status. The legends indicate the number of patients and the number of events. All depicted p-values are from log-rank tests.

Furthermore, there are 26 spatial features significantly associated with the survival of ER-positive samples. Figure 54B shows the ten most significant features. Similar to the associations with all samples, higher shapes and lower scales for spatial relationships with endothelial cells and fibroblasts

are associated with improved survival. In contrast, higher shape values for basal cells to CK^- cells, $CD4^+$ cells & APCs to CK^- cells, B cells to myofibroblasts, granulocytes & macrophages to $CD4^+$ cells & APCs, and endothelial cells to epithelial $Ki67^+$ cells have a poor prognostic impact. Higher scale values for granulocytes & macrophages and CK^+ER^- cells are also associated with poor prognoses.

The survival probability increases significantly if the scale parameter is higher than the median for the spatial relationships from granulocytes & macrophages to CK^+ER^- cells (Figure 56). For the remaining spatial relationships, the association is only significant with the continuous values and not with the discretized variable.

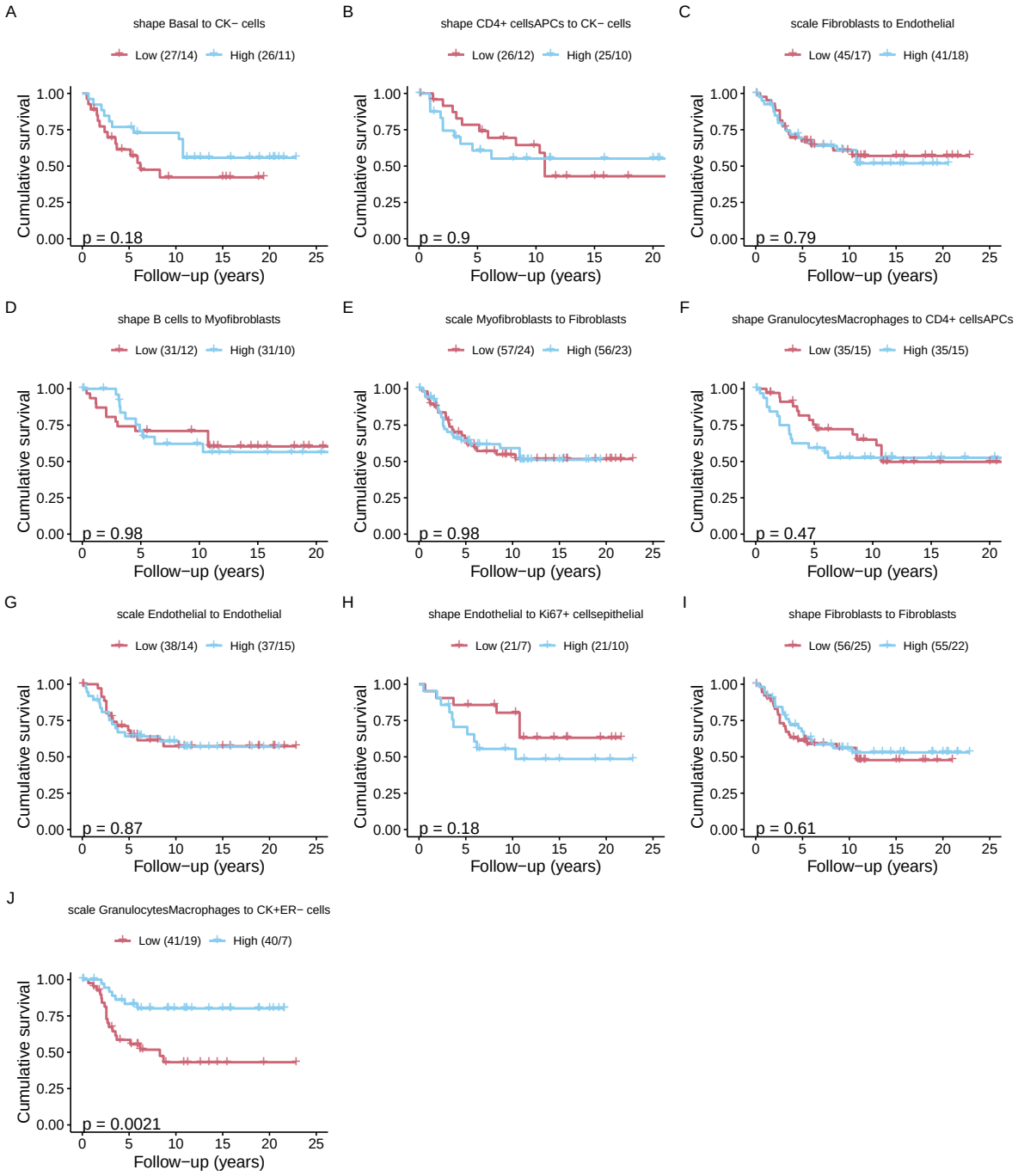

Figure 56: Predicted survival proportion over time for two risk groups (low: $\leq$ median, high: $>$ median) stratified by spatial relationships significantly associated with the hazard rates of all ER-positive samples. Predictions are adjusted for HER2 status. The legends indicate the number of patients and the number of events. All depicted p-values are from log-rank tests.

4.5.3. Section Summary

We evaluated the prognostic impact of cell-type fractions and spatial relationships by examining their univariate associations with hazard rates. We also visualized the effect of low and high values by discretizing the variables based on their median. This analysis addressed Research Question 4:

> What cell types and spatial relationships are associated with patient survival?

We found that fibroblasts and endothelial cells' fractions and spatial relationships were significantly associated with survival outcomes. When both cell types were abundant and in close proximity, patients had a higher likelihood of survival. At the same time, larger distances between endothelial cells and epithelial $Ki67^{+}$ cells, granulocytes & macrophages and $CD4^{+}$ cells & APCs, $CD4^{+}$ cells & APCs and CK^{med} cells, and $CD8^{+}$ cells and granulocytes & macrophages were associated with a better prognosis.

We also investigated whether spatial relationships were associated with the survival of ER-positive and ER-negative patients separately. Consistent with the findings of Danenberg et al. (2022), we did not observe significant prognostic impacts of spatial relationships in the ER-negative group. However, for ER-positive tumors, numerous spatial relationships were associated with survival rates. Once again, densely packed endothelial cells and fibroblasts were linked to improved survival. Similarly, dense packing of granulocytes & macrophages and $CK^{+}ER^{-}$ cells was associated with a better prognosis. The remaining spatial relationships showed the opposite, where higher survival probabilities were associated with larger distances between cells.

The significant spatial relationships were discretized based on the median to compare the survival probabilities of two patient groups. Yet, the discretized variable was not significantly associated with survival rates for most relationships, especially within ER-positive tumors. Splitting patients on more extreme values results in survival curves with larger differences.

4.6. Neighborhoods and Spatial Relationships

Neighborhoods are cellular regions with recurrent properties that potentially influence tumor phenotype and treatment response. Danenberg et al. (2022) identified ten recurrent neighborhoods in the METABRIC cohort (Figure 57). In this last section, we characterized the spatial organization of these neighborhoods, using the same method that was used to characterize the spatial relationships in the tissue slides. In contrast to these analyses, the neighborhood region was isolated from the remaining tissue, and separate analyses were performed on both sections. Significantly associated spatial relationships were found for active stroma, vascular stroma, $CD8^+$ & macrophage, TLS-like, and APC-enriched neighborhoods.

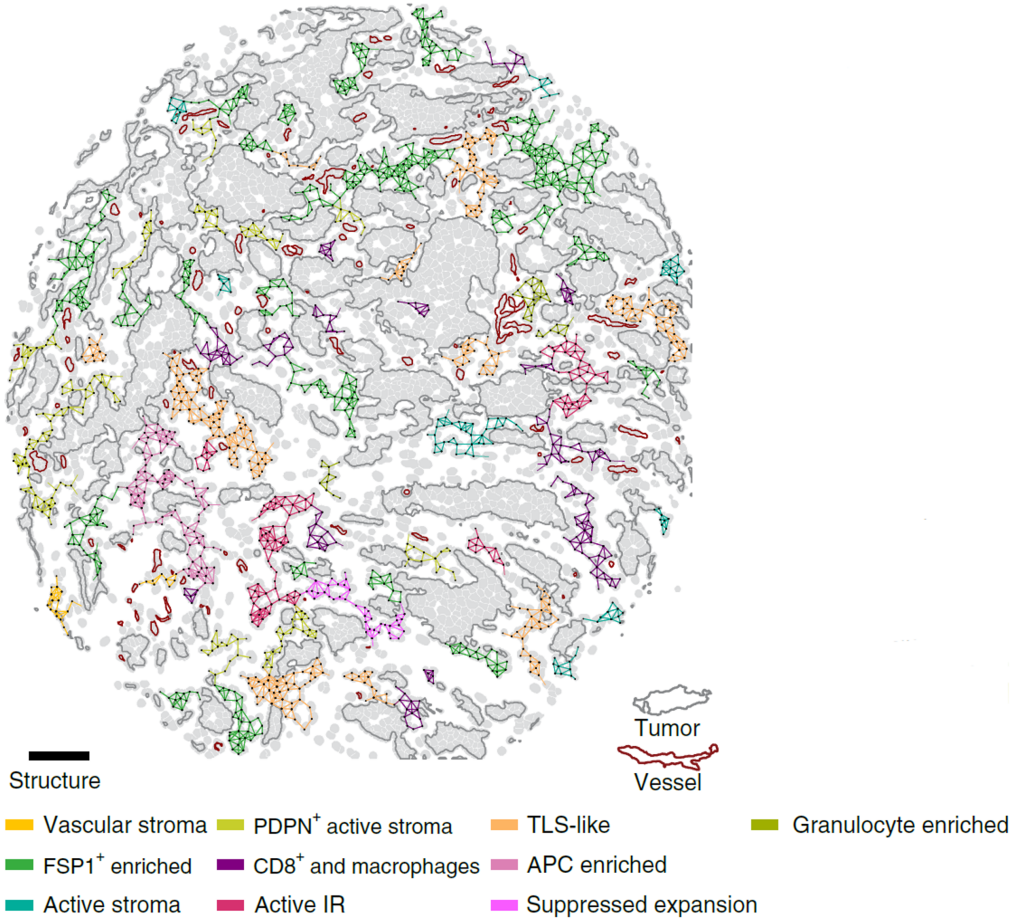

Figure 57: Example of neighborhoods in a tissue slide. Cells part of dense regions are marked with a black dot. Cells outside the regions are shown without centroid. Reprinted from "Breast tumor microenvironment structures are associated with genomic features and clinical outcome" by Danenberg et al. *Nature genetics, 54(5),* 660–669.

4.6.1. Spatial relationships in vascular stroma

Regions classified as vascular stroma are characterized by higher fractions of CD8$^+$ cells, CD4$^+$ cells & APCs, B cells, and granulocytes & macrophages, while it contains lower fractions of endothelial cells and fibroblasts. The associations are in contrast with the conventional characterizations of vascularization stating that vascular stroma contains an abundance of endothelial cells (Madu et al., 2020). No other neighborhood is associated with high endothelial cell fractions, indicating that endothelial cells are abundant throughout the remaining tissue but are not contained in dense regions.

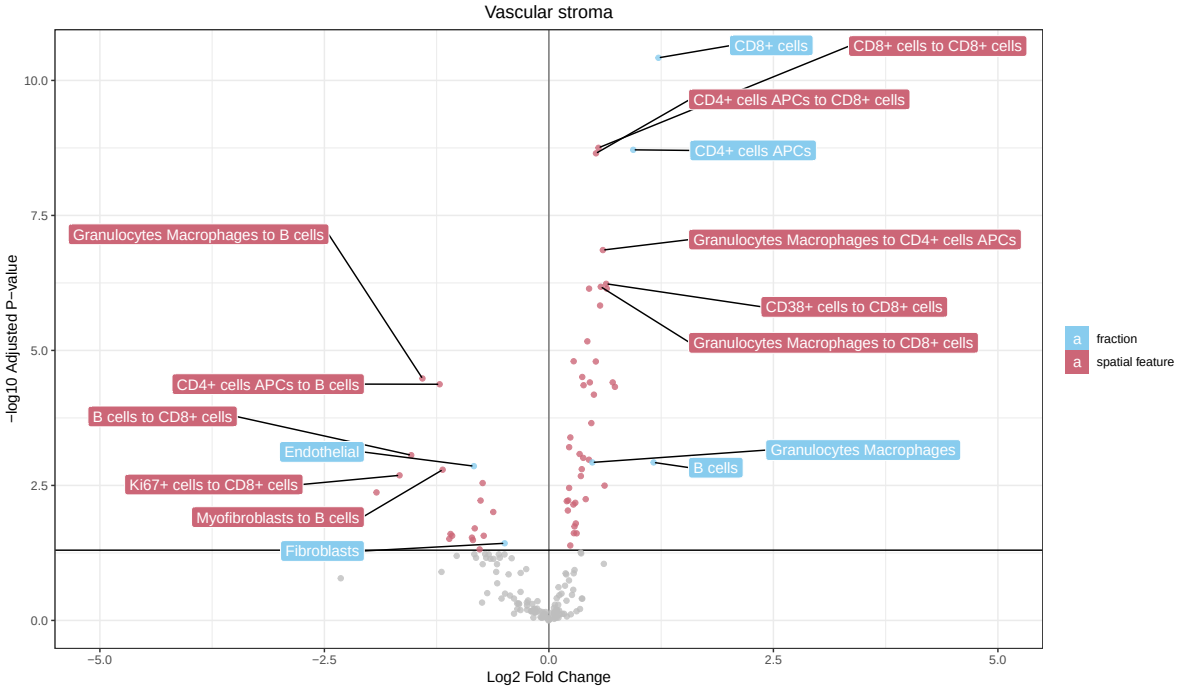

Figure 58: Associations between cell-type fractions and vascular stroma (blue) and between spatial relationships and vascular stroma (red). P-values are adjusted for multiple testing; Fold change is computed by dividing the mean of features in neighborhoods by the mean of features in the remaining tissue; the five most significantly associated spatial relationships are labeled.

The associated spatial relationships are characterized by higher shapes or smaller scales in vascular stroma indicating a tighter packing of cells (Figure 58).

Even though endothelial cells are not more prevalent, the cells are significantly differently arranged in vascular stroma. Endothelial cells are more densely packed with CD8$^+$ cells (Figure 59) and CD4$^+$ cells & APCs.

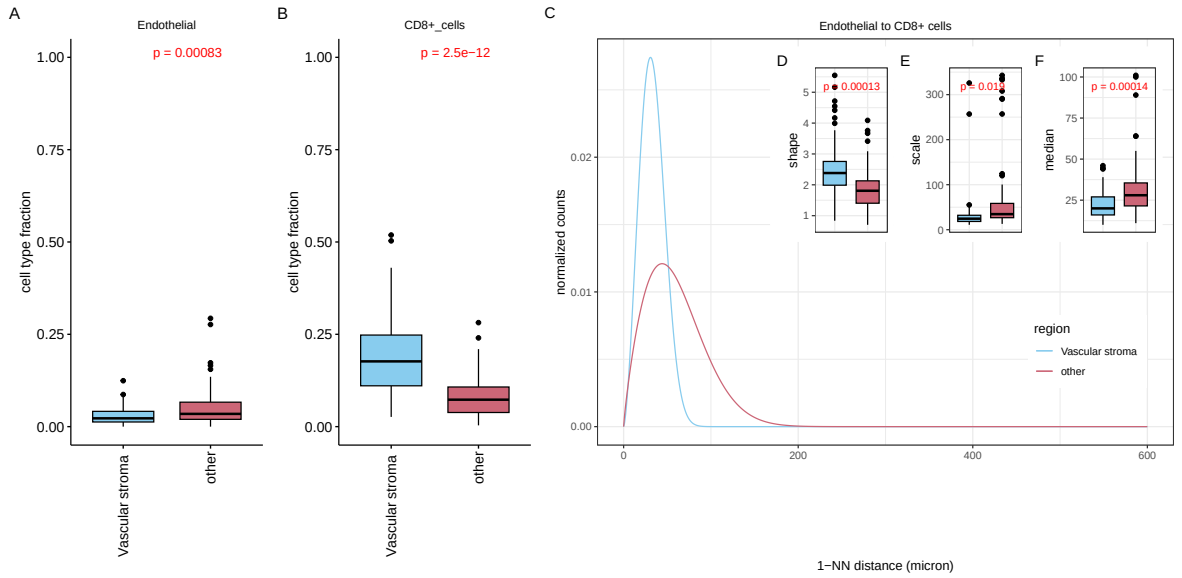

Figure 59: Fractions of endothelial cells (A) and CD8$^+$ cells in vascular stroma (B); Weibull curve fitted on the distribution of 1-NN distances from endothelial cells to CD8$^+$ cells in the vascular stroma (C); shape parameters are significantly larger in vascular stroma (D), while scale values are smaller (E); median 1-NN distances are smaller (F).

Additionally, CD8$^+$ cells occur in dense packings with several cell types. The cells are aggregated with endothelial cells (Figure 60), CD38$^+$ cells, and granulocytes & macrophages, and other CD8$^+$ cells.

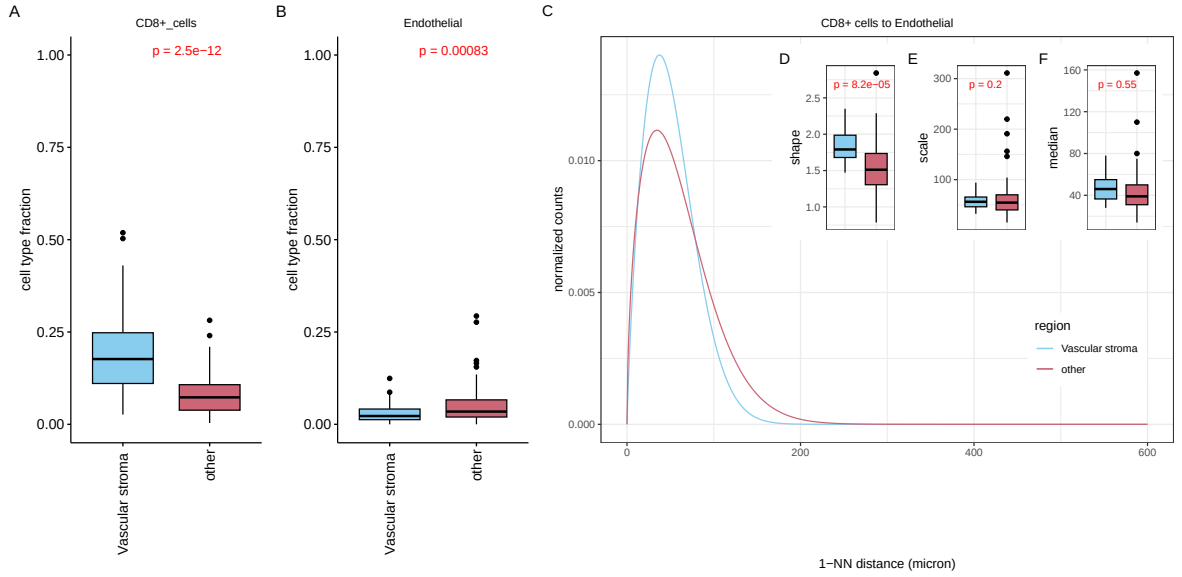

Figure 60: Fractions of CD8$^+$ T cells and endothelial cells in vascular stroma (AB); Weibull curve fitted on the distribution of 1-NN distances from CD8$^+$ cells to endothelial cells in the vascular stroma (C); shape parameters are significantly larger in vascular stroma (D), while scale values are the same (E); median 1-NN distances are smaller (F).

Furthermore, we find that myofibroblasts are densely packed with granulocytes & macrophages, CD8$^+$ cells, CD4$^+$ cells & APCs, and B cells. An example is given for the spatial relationship between myofibroblasts and granulocytes & macrophages (Figure 61).

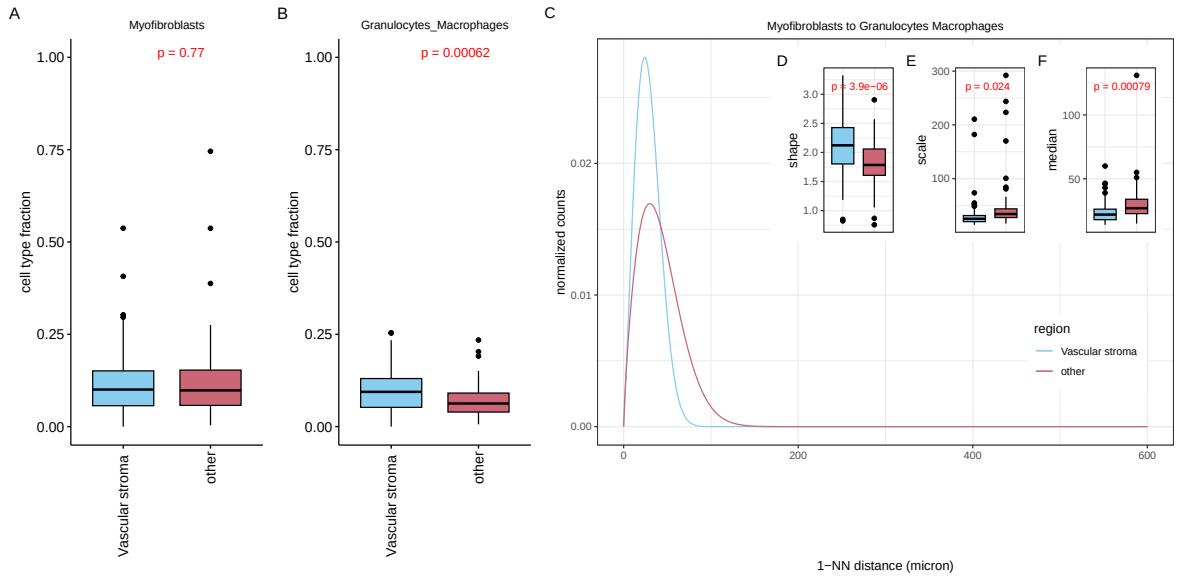

Figure 61: Fractions of myofibroblasts (A) and granulocytes & macrophages in vascular stroma (B); Weibull curve fitted on the distribution of 1-NN distances from myofibroblasts to granulocytes & macrophages in the vascular stroma (C); shape parameters are significantly larger in vascular stroma (D), while scale values are smaller (E); median 1-NN distances are smaller (F).

4.6.2. Spatial relationships in active stroma

Neighborhoods classified as active stroma are involved in stromal activation which is a process in which fibroblasts differentiate into myofibroblasts (Zainab, Sultana, et al., 2019). Active stroma contains higher fractions of fibroblasts and the spatial relationships expose that they are more densely packed together (Figure 62). Moreover, myofibroblasts are also closer to fibroblasts (Figure 63).

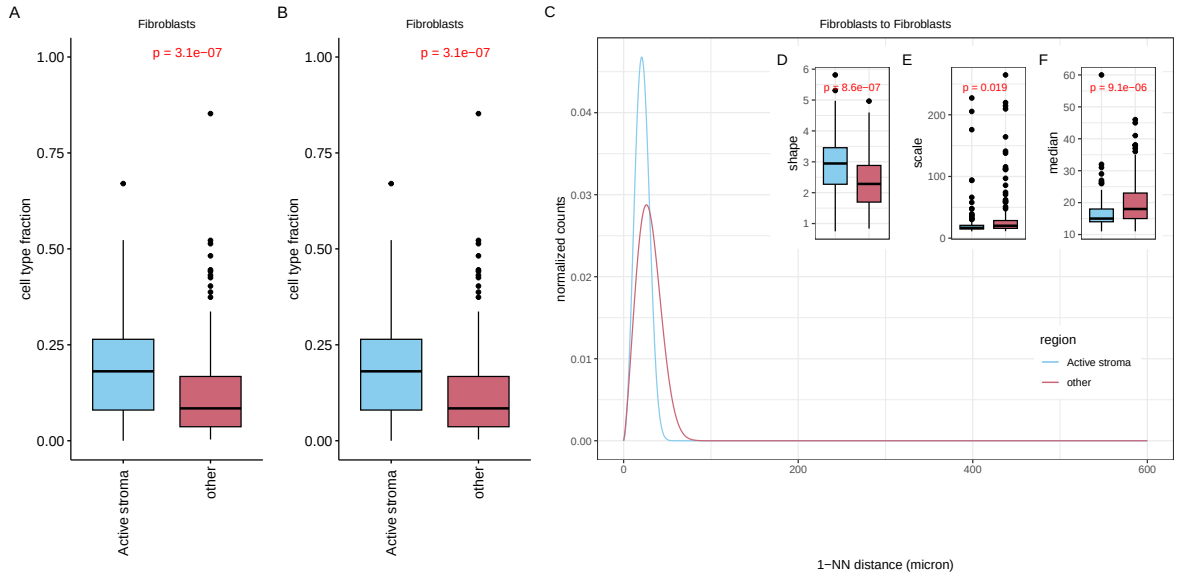

Figure 62: Fractions of fibroblasts in active stroma (AB); Weibull curve fitted on the distribution of 1-NN distances from fibroblasts to fibroblasts in active stroma (C); shape parameters are significantly larger in active stroma (D), while scale values are smaller (E); median 1-NN distances are also smaller in active stroma (F).

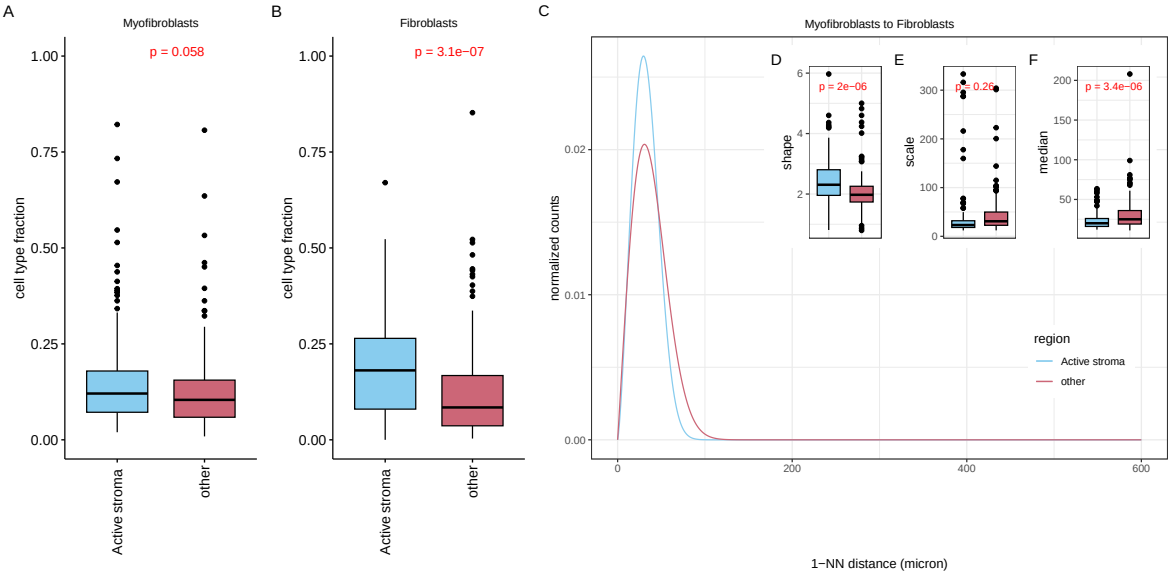

Figure 63: Fractions of myofibroblasts and fibroblasts in active stroma (AB); Weibull curve fitted on the distribution of 1-NN distances from myofibroblasts to fibroblasts in active stroma (C); shape parameters are significantly larger in active stroma (D), while scale parameters are not significantly different (E). Median 1-NN distances are also smaller in active stroma (F).

Associations with cell-type fractions reveal that active stroma contains higher fractions of granulocytes & macrophages, $CD4^+$ cells & APCs, $CD8^+$ cells, fibroblasts, $CD38^+$ cells and B cells and lower fractions of $Ki67^+$ cells and endothelial cells. Furthermore, many spatial relationships between these cell types are associated with active stroma (Figure 64).

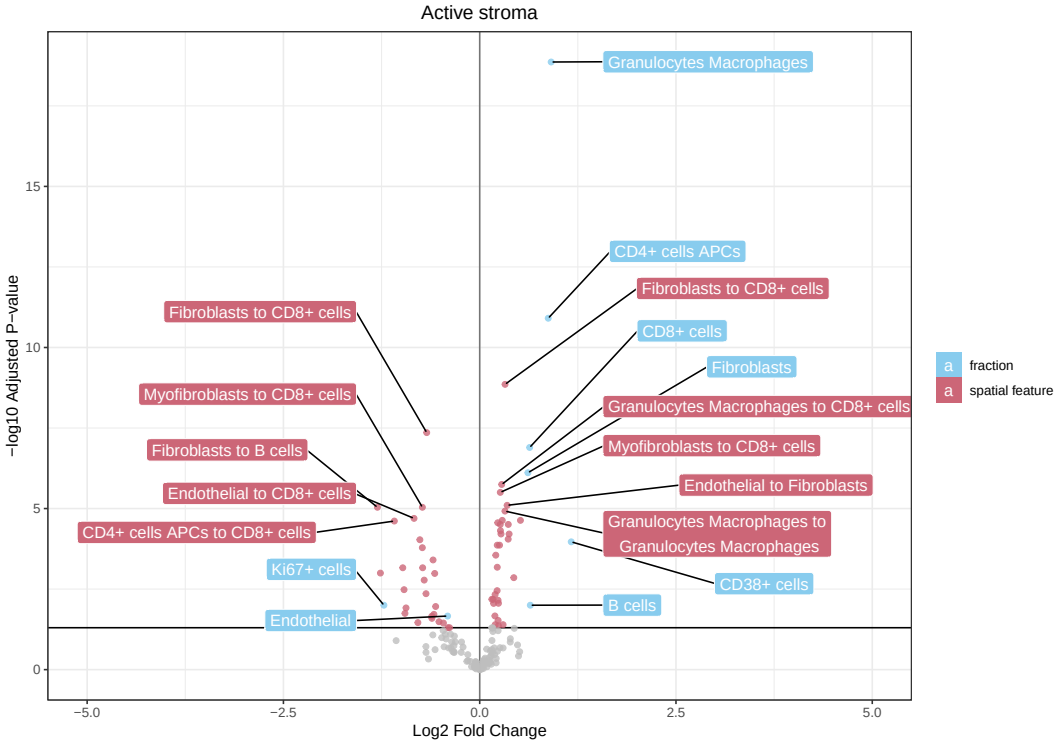

Figure 64: Associations between active stroma and cell-type fractions (blue) and between active stroma and spatial relationships (red). P-values are adjusted for multiple testing; Fold change is computed by dividing the mean of neighborhood regions by the mean of all other regions; The five most significantly associated spatial relationships are labeled.

All associated spatial relationships show similar behavior and the distance distributions are always fitted with higher shape parameters and smaller scale parameters in active stroma. An example is given for CD8$^+$ cells to B cells (Figure 65). Granulocytes & macrophages, CD4$^+$ cells & APCs, CD8$^+$ cells, fibroblasts, CD38$^+$ cells and B cells are therefore, more abundant and more tightly packed.

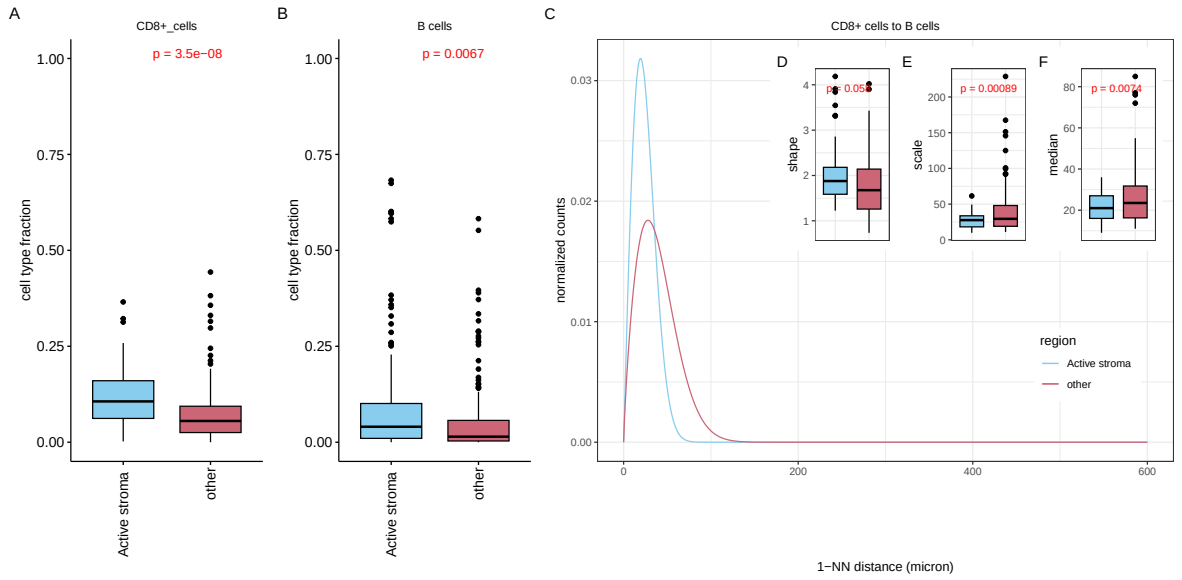

Figure 65: Fractions of CD8$^+$ cells and B cells in active stroma and other tissue (AB). Weibull curve fitted on the distribution of 1-NN distances from CD8$^+$ cells and B cells in active stroma and other tissue (C). shape parameters are similar in active stroma and other tissue (D), but scale parameters are smaller (E). The median 1-NN distances are also smaller in active stroma (F).

4.6.3. Spatial relationships in CD8$^+$ and macrophage neighborhoods

CD8$^+$ cells and macrophages are not more prevalent in CD8$^+$ and macrophage neighborhoods compared to the remaining tissue.

Associations reveal that the neighborhoods contain higher fractions of fibroblasts and that granulocytes & macrophages are closer to these fibroblasts (Figure 67).

Finally, distributions of the 1-NN distances between endothelial cells and fibroblasts have the same median for both CD8$^+$ and macrophage neighborhoods and the remaining tissue, but the variance of the distances is smaller (Figure 68).

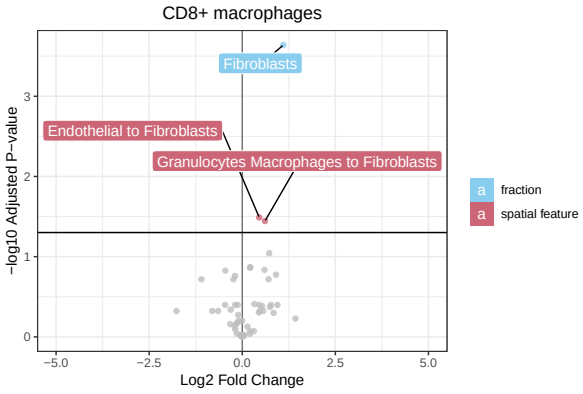

Figure 66: Associations between cell type fractions and CD8$^+$ and macrophage neighborhoods (blue) and between spatial relationships and neighborhoods (red). P-values are adjusted for multiple testing; Fold change is computed by dividing the mean of neighborhood regions by the mean of all other regions.

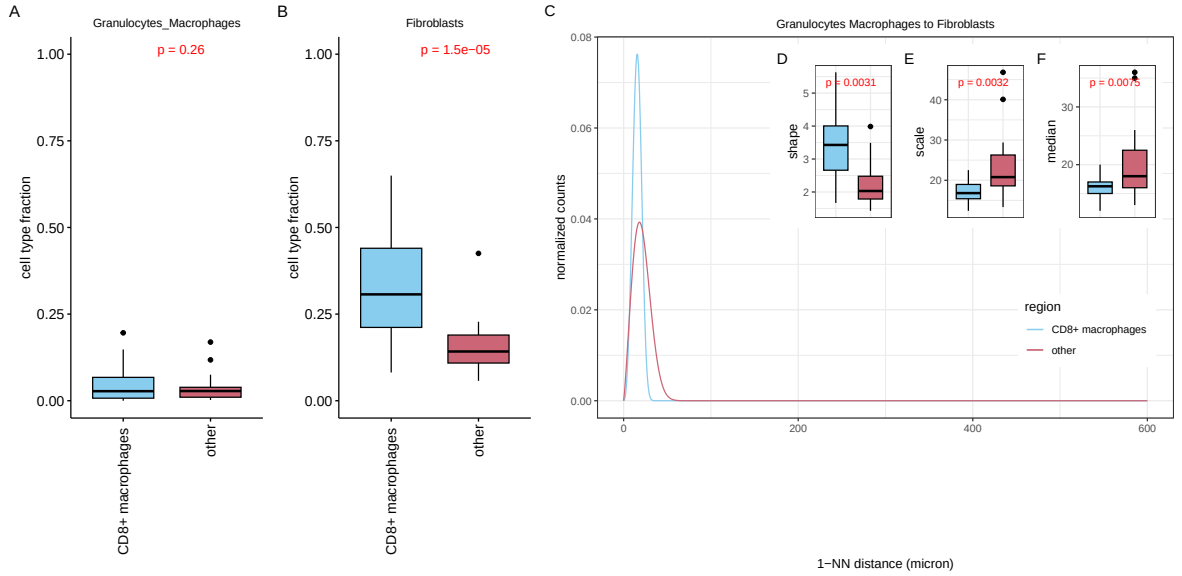

Figure 67: Fractions of granulocytes & macrophages and fibroblasts in CD8$^+$ and macrophage neighborhoods (AB); Weibull curve fitted on the distribution of 1-NN distances from granulocytes & macrophages to fibroblasts in CD8$^+$ and macrophage neighborhoods (C); shape parameters are significantly larger in the neighborhoods (D) and scale parameters are significantly smaller (E); median 1-NN distances are also significantly smaller (F).

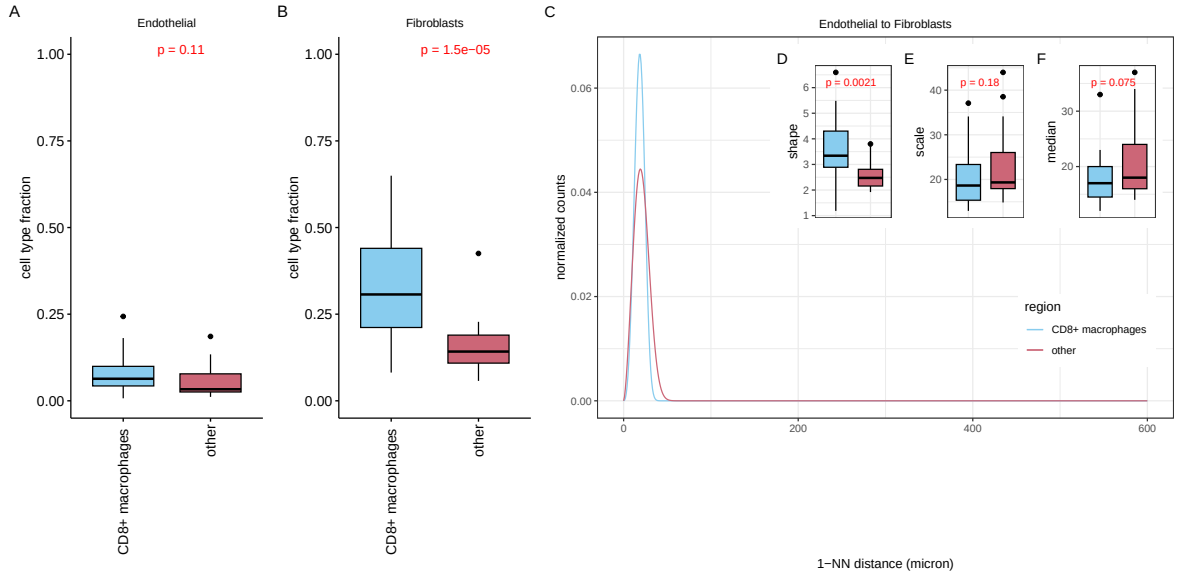

Figure 68: Fractions of endothelial cells and fibroblasts in CD8$^+$ and macrophage neighborhoods (AB); Weibull curve fitted on the distribution of 1-NN distances from endothelial cells to fibroblasts in CD8$^+$ and macrophage neighborhoods (C); shape parameters are significantly larger in the neighborhoods(D) while scale parameters are not significantly different (E); median 1-NN distances are also not different (F).

4.6.4. Spatial relationships in TLS-like neighborhoods

Regions classified as TLS-like neighborhoods are characterized by complex heterocellular compositions reminiscent of TLSs. As mentioned before, TLSs are aggregates of immune cells with high proportions of B cells. Yet, no significant associations with B cell fractions or spatial relationships concerning B cells are found for TLS-like neighborhoods (Figure 69).

Myofibroblasts and fibroblasts are abundant in TLS-like neighborhoods, while the neighborhoods contain lower fractions of CD38$^+$ and Ki67$^+$ cells. Myofibroblasts are also more tightly packed with CD8$^+$ cells (Figure 70).

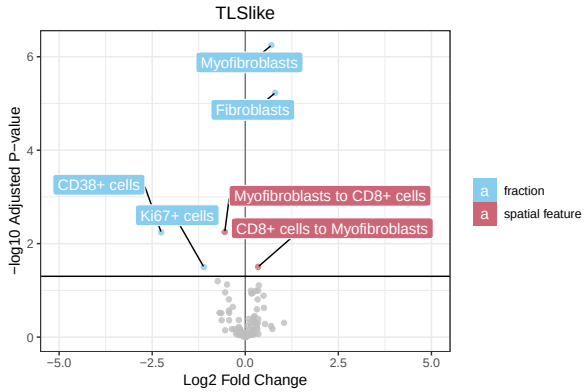

Figure 69: Associations between cell type fractions and TLS-like neighborhoods (blue) and between spatial relationships and neighborhoods (red). P-values are adjusted for multiple testing; Fold change is computed by dividing the mean of neighborhood regions by the mean of all other regions; The five most significantly associated spatial relationships are labeled.

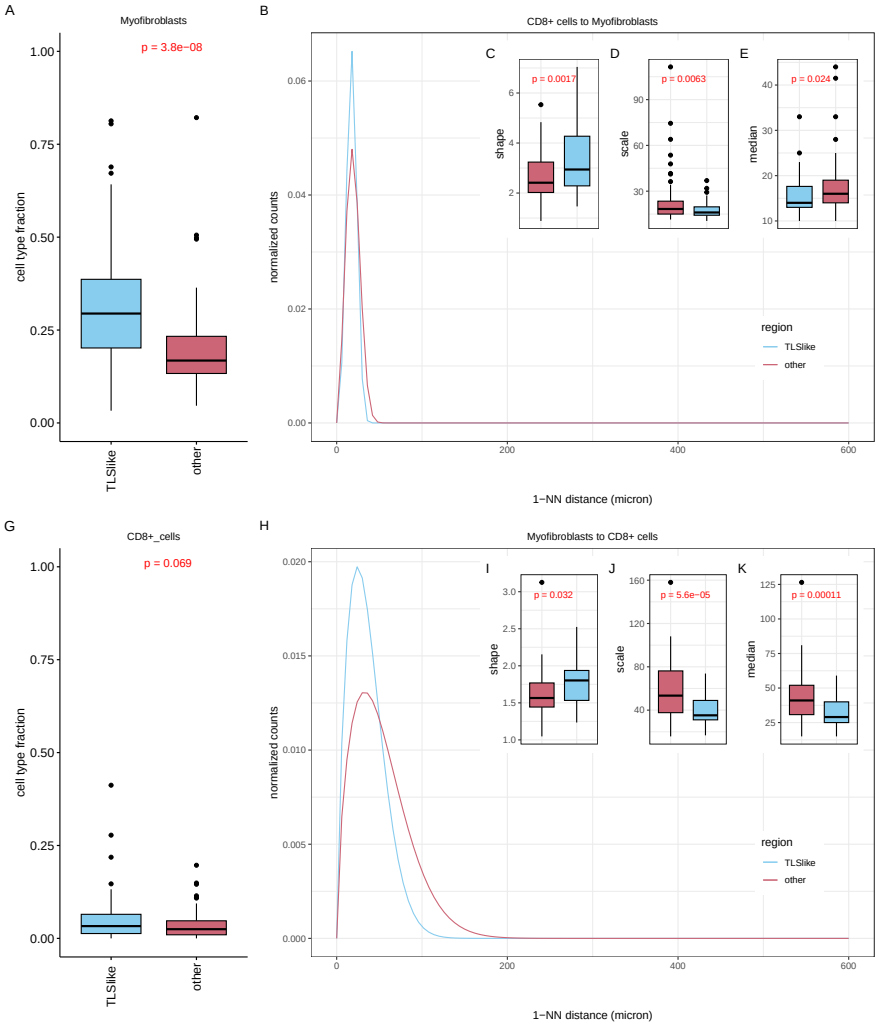

Figure 70: Fractions of Myofibroblasts (A) and $CD8^+$ cells (G) in TLS-like neighborhoods; Weibull curve fitted on the distribution of 1-NN distances from $CD8^+$ cells to Myofibroblasts in TLS-like neighborhoods (B); shape parameters are significantly larger (C) and scale parameters are significantly smaller (D). A tighter packing of the cell types in TLS-like neighborhoods is confirmed by the median 1-NN distances (E). The distribution of 1-NN distances from myofibroblasts to $CD8^+$ cells is also fitted with a Weibull distribution with larger shapes (I), smaller scales (J), and smaller medians (K).

4.6.5. Spatial relationships in APC-enriched neighborhoods

In contrast to what its name suggests, APC-enriched neighborhoods contain lower fractions of $CD4^+$ cells & APCs. $CD4^+$ cells & APCs comprise both $CD4^+$ cells and APCs, and low numbers of $CD4^+$ T cells possibly disguise the abundance of APCs.

The spatial relationship from fibroblasts to endothelial cells is the only spatial feature that characterizes the neighborhood (Figure 71). The distribution of 1-NN distances from fibroblasts to endothelial cells has a larger median and variance than the distributions in the remaining tissue (Figure 72).

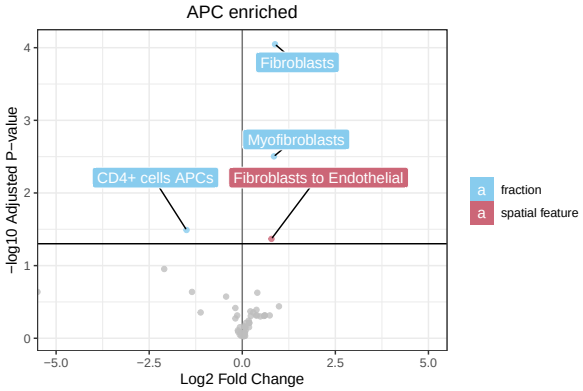

Figure 71: Associations between cell type fractions and APC-enriched neighborhoods (blue) and between spatial relationships and the neighborhoods (red). P-values are adjusted for multiple testing; Fold change is computed by dividing the mean of neighborhood regions by the mean of all other regions.

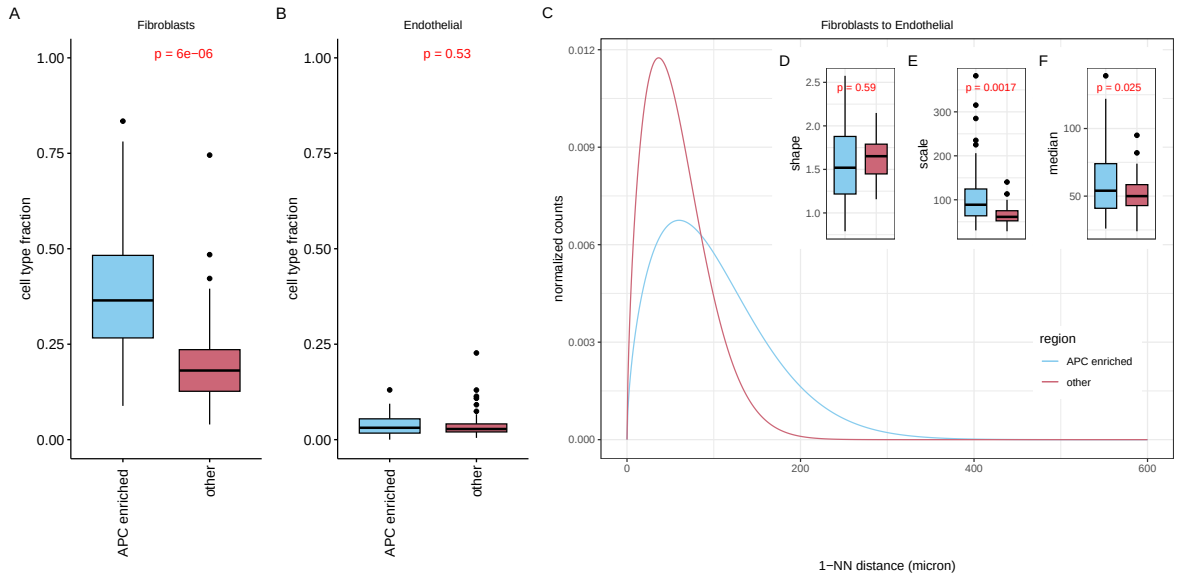

Figure 72: Fractions of fibroblasts and endothelial cells in APC-enriched neighborhoods (AB); Weibull curve fitted on the distribution of 1-NN distances from fibroblasts to endothelial cells in APC-enriched neighborhoods (C); shape parameters are not significantly different (D), but scale parameters are significantly larger (E); median 1-NN distances between the two cell types are also significantly larger (E).

4.6.6. Section Summary

Evaluating distance relationships within neighborhoods provides an alternative understanding of the composition of these regions. This section addressed Research Question 5:

> What cell types and spatial relationships are associated with cellular neighborhoods?

Several spatial relationships confirmed the properties described by Danenberg et al. (2022), particularly in vascular and active stroma. For instance, vascular stroma exhibited a dense packing of $CD8^+$ cells, while active stroma displayed a dense packing of fibroblasts. Vascular and active stroma were associated with numerous spatial relationships, all indicating a denser arrangement of cells in these neighborhoods. Notably, these relationships included cell types not noted by Danenberg et al.

Several neighborhood properties were not evident from our spatial analysis. Endothelial cells were not abundant in vascular stroma, but the spatial relationships revealed that the cells occurred in significantly different arrangements. Moreover, $CD8^+$ and macrophage neighborhoods did not contain higher fractions of $CD8^+$ cells and macrophages, while APC-enriched neighborhoods contained few $CD4^+$ cells & APCs.

Danenberg et al. characterized neighborhoods by comparing the compositions across these regions. In contrast, our analysis involved the study of distinctions between neighborhoods and the remaining tissue, encompassing cells not assigned to dense regions (refer to Figure 57).

The number of spatial relationships associated with $CD8^+$ and macrophage, APC-enriched, and TLS-like neighborhoods was limited. Furthermore, no significant spatial relationships were identified for the remaining five neighborhoods. This can be attributed to the relatively small sizes of these neighborhood regions, which hindered the detection of spatial relationships between cell-type pairs. These limitations are similar to those described in Section 4.2. The spatial relationship analysis demands an adequate number of reference and target cells for robust results.

5 Discussion

5.1. Conclusions

The characterization of cancer cells based on molecular features has led to the distinction of intrinsic tumor subtypes with differences in incidence, survival, and treatment response. Yet, breast cancer is a heterogeneous disease and underlying mechanisms driving individual tumor behavior remain unknown. Traditional analyses have typically disregarded the role of cells in the TME, despite the evident influence of factors beyond tumor cells alone. Additionally, little attention has been given to the spatial relationships among cells, even though cell arrangement is intricately connected to distinct functional aspects.

In this research, we addressed these gaps by extending the characterization of breast cancer tumors with a comprehensive description of the epithelium and TME and a systematic analysis of the spatial relationships between all cell-type pairs. Our study encompasses 749 tissue biopsies included in the METABRIC dataset, providing a sufficiently large and diverse sample pool to capture the inherent heterogeneity of breast cancer. As we conclude this final section, we reflect on the principal biological insights of our research, acknowledging the value of these findings in advancing the comprehension of this complex disease.

Our study leveraged the protein expression profiles of single cells, acquired through imaging mass cytometry, to distinguish cell types. Through our association analysis with breast cancer subtypes, we unveiled that these cell types are present across different tumor subtypes. Importantly, the subtypes encompass specialized tumor cell types like basal, $HER2^+$, and ER^+ cells and exhibit significant fractions of specialized epithelial and TME cell types.

Basal tumors were found to contain elevated fractions of proliferating epithelial cells and MHC-presenting cells. The TME in basal and HER2-enriched tumors exhibited a diverse array of cell types but notably lacked fibroblasts and endothelial cells. In contrast, the TME of luminal A tumors was predominantly composed of fibroblasts and endothelial cells. The higher TME cell type diversity in basal and HER2-enriched tumors and lower diversity in luminal A tumors have been shown before (Ali et al., 2020, Danenberg et al., 2022). In our research, we confirmed this observation and provided the precise compositions.

Furthermore, we demonstrated that cell-type fractions possess robust predictive capabilities for distinguishing breast cancer subtypes, particularly in the case of basal and HER2-enriched tumors. While it has been recognized that these subtypes comprise specialized tumor cell types, our study uncovered that the TME also hosts distinctive cell types. In contrast, prediction performances for luminal A, luminal B, and normal-like tumors were less robust due to the similarity in cell-type compositions among these subtypes.

Notably, distinguishing luminal B and normal-like tumors proved particularly challenging, with prediction performance failing to reach a moderate level. Normal-like tumors closely resemble healthy breast tissue, complicating their discrimination. Moreover, subtype classification is also based on morphological features not represented by cell-type fractions, further contributing to the complexity of subtype differentiation. Similar challenges have been addressed in earlier studies (Schneider et al., 2022) and features such as cell morphology and tissue structure are required to improve subtype classification.

Our exploration of spatial relationships has illuminated four noteworthy patterns within breast cancer subtypes. Firstly, we observed several spatial relationships that reflect cell-type abundance. Prevalent cell types such as basal cells and granulocytes & macrophages in basal tumors and fibroblasts and endothelial cells in luminal A tumors are densely packed in these tumor subtypes. Similarly, we find that cell types that occur rarely in subtypes are often separated by larger distances. In basal and HER2-enriched tumors, for example, tumor cells such as CK^+ER^-, CK^+ER^+ and CK^{med} are far apart.

Secondly, we observed spatial relationships between cell types for which fractions are not significantly different. These findings emphasize the value of spatial relationships as an additional layer of information into tumor characteristics. The relationships are very diverse and most are associated with HER2-enriched, luminal B, and normal-like tumors.

Thirdly, we found several spatial relationships that exhibit an inverse effect, where low occurrences of cells were associated with dense packings. For example, in basal tumors, $HER2^+$ cells and CK^+ER^- cells occur rarely, but the cells that occur are tightly packed. Similar behavior is found for CK^+ER^+ cells and myofibroblasts, CK^+ER^+ cells and $CD8^+$ cells, and B cells and endothelial cells in HER2-enriched tumors.

Finally, we identified four spatial relationships associated with multiple tumor subtypes. CK^{med} and CK^+ER^- cells are far apart in basal tumors but close in luminal A tumors. Both tumor cell types are characteristic of luminal A tumors and occur rarely in basal tumors. Additionally, the self-self relationship of granulocytes & macrophages, spatial relationships between epithelial $Ki67^+$ cells and granulocytes & macrophages, and $CD38^+$ cells and granulocytes & macrophages show a denser packing in basal cells and a separation in luminal A samples. The important role of macrophages in tumor growth has been addressed before (Schapiro et al., 2017) and here we have shown that their role differs across subtypes with different growth rates.

Furthermore, we found that denser arrangements of fibroblasts and endothelial cells are linked to improved survival across all samples. These same spatial relationships were also observed in luminal A tumors, a subtype associated with a generally positive prognosis. Remarkably, this association persisted when examining ER-positive samples, indicating that the spatial configurations of fibroblasts and endothelial cells have a prognostic impact within the broader category of luminal tumors.

However, it's notable that we did not identify significant associations between survival and spatial relationships in ER-negative samples. This finding aligns with the study by Danenberg et al. (2022), which likewise did not uncover associations between neighborhood presence and survival in ER-negative samples. Together, these findings suggest that other analytical approaches may be needed to uncover spatial arrangements associated with the survival of ER-negative breast cancer patients.

The final objective of this research was to provide a spatial characterization of conserved neighborhoods. Neighborhoods have received considerable attention due to their link with local TME behavior and their effect on tumor phenotypes. However, an assessment of distance relationships has been performed just now.

We observed that vascular and active stroma neighborhoods exhibit numerous recurring spatial relationships, all indicating a tightly packed configuration of a limited set of TME cell types. The spatial relationships provided insights into the spatial arrangement of cells involved in stromal activation and vascularization.

In the case of $CD8^+$ and macrophage, APC-enriched, and TLS-like neighborhoods, we identified several recurrent spatial features, although none of these associations could be linked to known characteristics of the neighborhoods. For instance, TLS-like neighborhoods are recognized for their abundance of B cells, yet we did not find any significant relationship with B cells within these neighborhoods. As mentioned by Danenberg et al., TLS-like neighborhoods are characterized by complex heterocellular compositions and they potentially encompass cellular regions lacking clear TLS characteristics.

Moreover, spatial analyses require sufficient cells to yield meaningful results. It is possible that the segmentation of tissue slides led to regions that were too small for robust spatial analyses.

Finally, it is important to evaluate the effectiveness of the spatial relationship analysis that was used in this study as it is a novel approach that has the potential to be used for the spatial profiling of other cohorts.

Spatial relationships were summarized by fitting Weibull distributions to the 1-NN distance distributions using an NLME model. The shape and scale parameters uniquely describe the Weibull distribution. Associations between spatial relationships and subtypes were identified by comparing the Weibull parameters among samples. Notably, several distance distributions linked to subtypes exhibited varia-

tions solely in terms of variance or amplitude, while the medians remained equal. This underscores the superiority of quantifying spatial relationships via Weibull parameters over other existing methods that rely on means (Ma et al., 2022) or medians (Parra et al., 2021) for comparing distance relationships.

While this method has seen limited application across a few datasets in the past, its extension to images acquired through mass-spectrometry-based imaging is entirely novel. Our explorations of various Weibull parameter combinations demonstrated the method's efficacy in capturing a wide array of spatial relationships, ranging from densely packed cell configurations to fully segregated compartments.

5.2. Limitations

Ongoing developments in multiplex imaging make it possible to consider cells with increasing detail, enabling the identification of precise cell states. Our cell classification efforts underscore that obtaining a highly detailed perspective of cell types can sometimes compromise the robustness and reliability of spatial relationship assessments. For some cell types, we were only able to define spatial relationships for merged groups such as macrophages & granulocytes and $CD4^+$ T cells & APCs.

Additionally, the image sizes also limited the spatial analysis. The images covered tissue areas roughly ten times smaller than the slides used in previous applications. The size reduction is an inherent outcome of employing mass-spectrometry-based imaging. A major threat of low cell counts to a robust spatial characterization is the observation that single cells can exert a large effect on the distance distribution. Our analysis of the parameter space showed increased parameter estimate variation, which may be attributed to the influence of single cells.

Furthermore, it's worth noting that the Weibull approximation assumes an unimodal distribution of the 1-NN distances. Failed parameter estimations were often the consequence of multimodal distributions. These distributions result from cells organized in small clusters. Even though small clusters often contain few cells and might disappear for larger cell counts, our method misses these spatial relationships.

5.3. Recommendations and Future Work

The conducted research has provided valuable biological insights, but it is important to acknowledge its limitations. We offer several recommendations to facilitate the adoption of our methods in future work and improve the generalizability of the results.

The spatial analysis has specific requirements for its applicability. Firstly, it relies on datasets with an adequate cell count. It may be necessary to adapt cell-type classifications to ensure the presence of robust cell-type populations, even if this entails merging multiple cell types.

Another crucial consideration is that tumor tissue biopsies offer a limited representation of the entire tumor tissue. Variations between different samples of the same tumor can be significant. Therefore, datasets need to be sufficiently large to ensure that disparities in cell-type composition and spatial relationships are not merely the result of local irregularities. This is especially important when utilizing single-cell information obtained through mass-spectrometry-based methods, given that the resulting tissue images are notably smaller than those acquired through microscopy-based methods.

This research primarily focused on univariate associations concerning subtypes and survival. Univariate comparisons, while valuable, provide a simplified view of the complex biological reality. Enhancing our understanding of associations with subtypes and survival could be achieved by extending our analysis to include multivariate approaches.

The logistic regression prediction models have demonstrated that spatial features are not well-suited for accurately predicting breast cancer subtypes. Subtyping primarily relies on intrinsic tumor properties, and our research did not aim to identify the best predictive features. Nevertheless, future studies aimed at addressing this challenge should consider incorporating morphological properties of cells into their analyses.

Additionally, we have uncovered limited spatial aspects associated with the survival of ER-negative patients. Given that ER-negative tumors are more aggressive than ER-positive tumors, further investigations into these aspects are essential to gain a deeper understanding of their behavior and potentially identify new prognostic factors.

Finally, it is important to note that characterizations of neighborhoods often oversimplify or do not consider spatial arrangements. Existing methods typically rely on factors such as cell counts or direct neighbors. Our study aimed for a more comprehensive description of these neighborhoods, although

with mixed success due to the small tissue regions. Moving forward, it is advisable for future methods to incorporate distance properties as well. By doing so, they obtain more interpretable and nuanced characteristics of the neighborhoods, allowing for a more accurate understanding of their structural intricacies.

References

Abdelaal, T., van Unen, V., Höllt, T., Koning, F., Reinders, M. J., & Mahfouz, A. (2019). Predicting cell populations in single cell mass cytometry data. *Cytometry Part A*, *95*(7), 769–781.

Ali, H. R., Jackson, H. W., Zanotelli, V. R., Danenberg, E., Fischer, J. R., Bardwell, H., Provenzano, E., Rueda, O. M., Chin, S.-F., et al. (2020). Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nature Cancer*, *1*(2), 163–175.

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., Siesling, S., et al. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, *66*, 15–23.

Barak, V., Goike, H., Panaretakis, K. W., & Einarsson, R. (2004). Clinical utility of cytokeratins as tumor markers. *Clinical biochemistry*, *37*(7), 529–540.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Cabrita, R., Lauss, M., Sanna, A., Donia, M., Skaarup Larsen, M., Mitra, S., Johansson, I., Phung, B., Harbst, K., Vallon-Christersson, J., et al. (2020). Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature*, *577*(7791), 561–565.

Cheung, A. M.-Y., Wang, D., Liu, K., Hope, T., Murray, M., Ginty, F., Nofech-Mozes, S., Martel, A. L., & Yaffe, M. J. (2021). Quantitative single-cell analysis of immunofluorescence protein multiplex images illustrates biomarker spatial heterogeneity within breast cancer subtypes. *Breast Cancer Research*, *23*(1), 1–17.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346–352.

Danenberg, E., Bardwell, H., Zanotelli, V. R., Provenzano, E., Chin, S.-F., Rueda, O. M., Green, A., Rakha, E., Aparicio, S., Ellis, I. O., et al. (2022). Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature genetics*, *54*(5), 660–669.

Elhanani, O., Ben-Uri, R., & Keren, L. (2023). Spatial profiling technologies illuminate the tumor microenvironment. *Cancer cell*.

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, *11*(4), 417–422.

Jackson, H. W., Fischer, J. R., Zanotelli, V. R., Ali, H. R., Mechera, R., Soysal, S. D., Moch, H., Muenst, S., Varga, Z., Weber, W. P., et al. (2020). The single-cell pathology landscape of breast cancer. *Nature*, *578*(7796), 615–620.

Kensler, K. H., Sankar, V. N., Wang, J., Zhang, X., Rubadue, C. A., Baker, G. M., Parker, J. S., Hoadley, K. A., Stancu, A. L., Pyle, M. E., et al. (2019). Pam50 molecular intrinsic subtypes in the nurses' health study cohorts. *Cancer Epidemiology, Biomarkers & Prevention*, *28*(4), 798–806.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Latapy, M., & Pons, P. (2004). Computing communities in large networks using random walks. *arXiv preprint cond-mat/0412368*.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. (2015). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, *162*(1), 184–197.

Ma, X., Guo, Z., Wei, X., Zhao, G., Han, D., Zhang, T., Chen, X., Cao, F., Dong, J., Zhao, L., et al. (2022). Spatial distribution and predictive significance of dendritic cells and macrophages in esophageal cancer treated with combined chemoradiotherapy and pd-1 blockade. *Frontiers in Immunology*, *12*, 786429.

Madu, C. O., Wang, S., Madu, C. O., & Lu, Y. (2020). Angiogenesis in breast cancer progression, diagnosis, and treatment. *Journal of Cancer*, *11*(15), 4474.

Mittal, S., Brown, N. J., & Holen, I. (2018). The breast tumor microenvironment: Role in cancer development, progression and response to therapy. *Expert review of molecular diagnostics*, *18*(3), 227–243.

Nolan, E., Lindeman, G. J., & Visvader, J. E. (2023). Deciphering breast cancer: From biology to the clinic. *Cell*.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, *27*(8), 1160.

Parra, E. R., Ferrufino-Schmidt, M. C., Tamegnon, A., Zhang, J., Solis, L., Jiang, M., Ibarguen, H., Haymaker, C., Lee, J. J., Bernatchez, C., et al. (2021). Immuno-profiling and cellular spatial analysis using five immune oncology multiplex immunofluorescence panels for paraffin tumor tissue. *Scientific reports*, *11*(1), 1–15.

Patwa, A., Yamashita, R., Long, J., Risom, T., Angelo, M., Keren, L., & Rubin, D. L. (2021). Multiplexed imaging analysis of the tumor-immune microenvironment reveals predictors of outcome in triple-negative breast cancer. *Communications Biology*, *4*(1), 852.

RIVM. (2019). https://www.vzinfo.nl/borstkanker/zorguitgaven

Rueda, O. M., Sammut, S.-J., Seoane, J. A., Chin, S.-F., Caswell-Jin, J. L., Callari, M., Batra, R., Pereira, B., Bruna, A., Ali, H. R., et al. (2019). Dynamics of breast-cancer relapse reveal late-recurring er-positive genomic subgroups. *Nature*, *567*(7748), 399–404.

Savas, P., Salgado, R., Denkert, C., Sotiriou, C., Darcy, P. K., Smyth, M. J., & Loi, S. (2016). Clinical relevance of host immunity in breast cancer: From tils to the clinic. *Nature reviews Clinical oncology*, *13*(4), 228–241.

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R., Schulz, D., Giesen, C., Catena, R., Varga, Z., & Bodenmiller, B. (2017). Histocat: Analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, *14*(9), 873–876.

Schneider, L., Laiouar-Pedari, S., Kuntz, S., Krieghoff-Henning, E., Hekler, A., Kather, J. N., Gaiser, T., Froehling, S., & Brinker, T. J. (2022). Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *European journal of cancer*, *160*, 80–91.

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., et al. (2020). Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, *182*(5), 1341–1359.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209–249.

Toth, C., Helic, D., & Geiger, B. C. (2022). Synwalk: Community detection via random walk modelling. *Data Mining and Knowledge Discovery*, *36*(2), 739–780.

Tsujikawa, T., Kumar, S., Borkar, R. N., Azimi, V., Thibault, G., Chang, Y. H., Balter, A., Kawashima, R., Choe, G., Sauer, D., et al. (2017). Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis. *Cell reports*, *19*(1), 203–217.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, *87*(7), 636–645.

Vos, J. L., Elbers, J. B., Krijgsman, O., Traets, J. J., Qiao, X., van der Leun, A. M., Lubeck, Y., Seignette, I. M., Smit, L. A., Willems, S. M., et al. (2021). Neoadjuvant immunotherapy with nivolumab and ipilimumab induces major pathological responses in patients with head and neck squamous cell carcinoma. *Nature communications*, *12*(1), 7348.

Zainab, H., Sultana, A., et al. (2019). Stromal desmoplasia as a possible prognostic indicator in different grades of oral squamous cell carcinoma. *Journal of oral and maxillofacial pathology: JOMFP*, *23*(3), 338.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

A

Protein Panel

Table 1: Proteins used for the epithelial and non-epithelial cell-type classification.

	Protein	Description
Epithelal	Histone H3	One of the five main histones involved in the structure of chromatin
	CK5	Marker for basal breast cancer
	HLA-DR	MHC class II cell surface receptor
	CK8-18	Cytokeratin produced in breast tumor cells
	CD15	Marker for human myeloid cells
	HER2 (3B5)	Receptor protein for epidermal growth factor
	Podoplanin	Transmembrane receptor glycoprotein that is upregulated on transformed cells, cancer-associated fibroblasts, and inflammatory macrophages
	HER2 (D8F12)	Receptor protein for epidermal growth factor
	B2M	A component of MHC class I molecules
	ER	Receptor protein for estrogen
	CD57	Marker for natural killer cells
	Ki-67	Marker for proliferating cells
	CXCL12	Strong chemotactic for lymphocytes
	HLA-ABC	MHC class II cell surface receptor
	pan-CK	Marker for tumor cells
	c-Casapse3	Protein playing a key role in both the death receptor pathway
TME	Histone H3	One of the five main histones involved in the structure of chromatin
	SMA	Marker for myoepithelial cells
	CD38	Marker of cell activation
	HLA-DR	MHC class II cell surface receptor
	CD15	Marker for human myeloid cells
	FSP1	Glutathione-independent ferroptosis suppressor
	CD163	Marker of cells from the monocyte/macrophage lineage
	ICOS	Immune checkpoint protein
	OX40	Secondary co-stimulatory immune checkpoint molecule
	CD68	Protein highly expressed by cells in the monocyte lineage
	CD3	Protein involved in activating both the cytotoxic T cell (CD8+ naive T cells) and T helper cells (CD4+ naive T cells).
	Podoplanin	Transmembrane receptor glycoprotein that is upregulated on transformed cells, cancer-associated fibroblasts, and inflammatory macrophages
	CD11c	Integrin alpha X chain protein highly abundant in human dendritic cells
	PD-1	Protein on the surface of T and B cells that has a role in regulating the immune system's response
	GITR	Marker for CD25+CD4+ regulatory T cells
	CD16	Marker for natural killer cells, neutrophils, monocytes, macrophages, and certain T cells.
	CD45RA	Marker for CD8+ memory T cells indicating terminal differentiation
	B2M	A component of MHC class I molecules
	CD45RO	Marker for CD8+ memory T cells
	FOXP3	Protein involved in immune system responses
	CD20	Marker for B cells
	CD8	Plays a major role in T cell signaling and aiding with cytotoxic T cell-antigen interactions binding to MHC molecules
	CD57	Marker for natural killer cells
	Ki-67	Marker for proliferating cells
	PDGFRB	Protein essential for vascular development
	Caveolin-1	Main component of the caveolae plasma membranes found in most cell types
	CD4	Co-receptor for the T-cell receptor (TCR) found on the surface of immune cells such as T helper cells, monocytes, macrophages, and dendritic cells
	CD31-vWF	Marker for endothelial cells
	HLA-ABC	MHC class II cell surface receptor
	c-Caspase3	Protein with a key role in both the death receptor pathway

B

Cell-type Descriptions

Table 2: Description of cell types in the epithelium and TME.

	Cell type	Description
Epitelial	Normal epithelial cell	Healthy breast tissue cells responsible for milk production
	Basal cells	Tumor cells without hormone receptors
	ER^+ cells	Tumor cells with estrogen receptors
	$HER2^+$ cells	Tumor cells with human epidermal growth factor 2 receptor
	$CD15^+$ tumor cells	Invasive breast tumor cells
	MHC-presenting cells	Cells with major histocompatibility complex that mediate interactions with leukocytes
TME	Endothelial cells	Cells that form the layer around blood vessels and regulate exchanges with the bloodstream and surrounding tissue
	Fibroblasts	Cells synthesize the extracellular matrix and collagen giving tissue its support. Moreover, fibroblasts play a critical role in wound healing
	Myofibibroblasts	Cells that are in a state between fibroblasts and smooth muscle cells
	Granulocytes	First responders of the innate immune system that clear invading microbes and necrotic cells
	Natural Killer (NK) cells	Cells of the innate immune system destroying harmful cells in the early stages
	B cells	Cells that are part of the adaptive immune system and produce antibodies that bind to harmful cells
	$CD4^+$ T cells	Cells that are part of the adaptive immune system and destroy marked cells
	Antigen-presenting cells (APCs)	Cells that display antigens critical for lymphocyte functioning
	$CD8^+$ T cells	Cells that are part of the adaptive immune system and produce signals to recruit other immune cells
	Regulatory T (T_{reg}) cells	Cells that are part of the adaptive immune system and protect healthy cells by suppressing the immune response
	Macrophages	Cells that engulf and digest cells that are marked as hostile
	$Ki67^+$ cells	Proliferating cells

C

Weibull Parameter Estimations NABUCCO trial

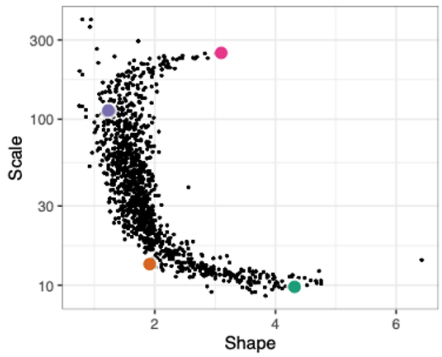

Figure 1: Shape versus scale parameters fitted on the 1-NN distribution for the cell type combinations (n=7) of all samples of the NABUCCO cohort (n=24).

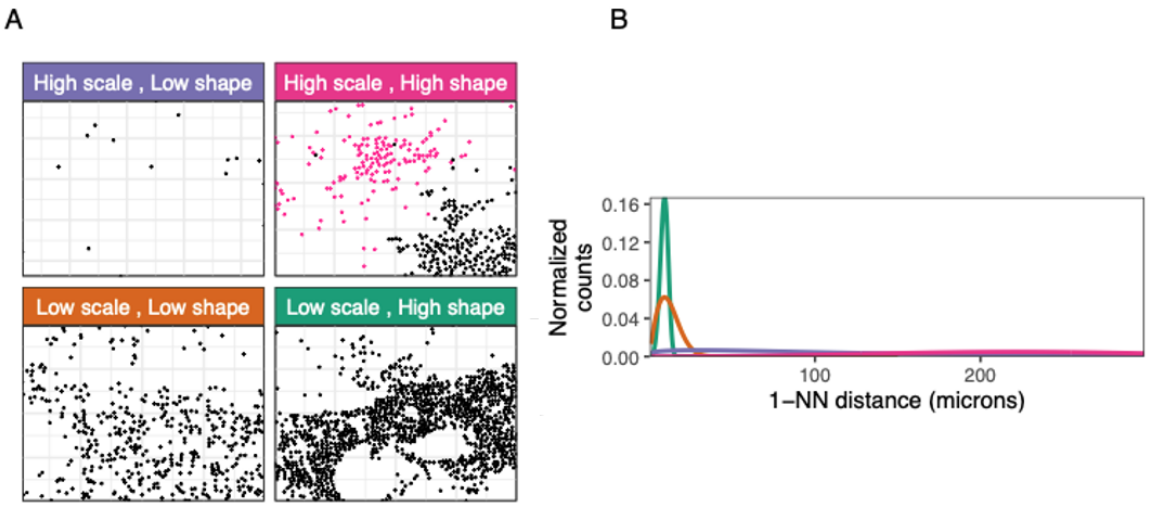

Figure 2: Four examples of spatial relationships (A) and 1-NN distance distributions (B) of spatial relationships in extremes of the parameter space.