cancers

MDPI

Article

Spatial Transcriptomic Analysis Reveals Associations between Genes and Cellular Topology in Breast and Prostate Cancers

Lujain Alsaleh [1], Chen Li [2], Justin L. Couetil [3], Ze Ye [2], Kun Huang [1,3,4,5], Jie Zhang [3,5], Chao Chen [2] and Travis S. Johnson [1,5,6,*]

[1] Department of Biostatistics and Health Data Science, Indiana University, Indianapolis, IN 46202, USA
[2] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA
[3] Department of Medical and Molecular Genetics, Indiana University, Indianapolis, IN 46202, USA
[4] Regenstrief Institute, Indiana University, Indianapolis, IN 46202, USA
[5] Melvin and Bren Simon Comprehensive Cancer Center, Indiana University, Indianapolis, IN 46202, USA
[6] Indiana Biosciences Research Institute, Indianapolis, IN 46202, USA
* Correspondence: johnstrs@iu.edu

Simple Summary: Tissues consist of various cell types in complex spatial arrangements. These tissues require signaling between the cells and structural proteins to maintain their organization. Spatial arrangements or the organization of cells in tissue can be readily studied via imaging analysis of dissected tissue. In cancer tissues, there is, additionally, signaling between immune cells and between immune and cancer cells. The signaling molecules as well as the structural molecules can be readily measured using molecular profiling technology. Until now, there have not been ways to have both high-resolution imaging and high-resolution molecular profiling for the same locations in tissue. Now that the technology exists to measure both molecular profile and corresponding image, we can study the relationship between molecules in the tissue and the resultant tissue architecture in order to determine functional relationships between the two in cancer. We find that image features correspond well with extracellular matrix and other mechanisms that are important in tumor structure and aggressiveness.

Abstract: Background: Cancer is the leading cause of death worldwide with breast and prostate cancer the most common among women and men, respectively. Gene expression and image features are independently prognostic of patient survival; but until the advent of spatial transcriptomics (ST), it was not possible to determine how gene expression of cells was tied to their spatial relationships (i.e., topology). Methods: We identify topology-associated genes (TAGs) that correlate with 700 image topological features (ITFs) in breast and prostate cancer ST samples. Genes and image topological features are independently clustered and correlated with each other. Themes among genes correlated with ITFs are investigated by functional enrichment analysis. Results: Overall, topology-associated genes (TAG) corresponding to extracellular matrix (ECM) and Collagen Type I Trimer gene ontology terms are common to both prostate and breast cancer. In breast cancer specifically, we identify the ZAG-PIP Complex as a TAG. In prostate cancer, we identify distinct TAGs that are enriched for GI dysmotility and the IgA immunoglobulin complex. We identified TAGs in every ST slide regardless of cancer type. Conclusions: These TAGs are enriched for ontology terms, illustrating the biological relevance to our image topology features and their potential utility in diagnostic and prognostic models.

Keywords: spatial transcriptomics; histopathological images; image analysis; breast cancer; prostate cancer; gene expression; topological data analysis; integrative analysis

check for updates

Citation: Alsaleh, L.; Li, C.; Couetil, J.L.; Ye, Z.; Huang, K.; Zhang, J.; Chen, C.; Johnson, T.S. Spatial Transcriptomic Analysis Reveals Associations between Genes and Cellular Topology in Breast and Prostate Cancers. *Cancers* **2022**, *14*, 4856. https://doi.org/10.3390/cancers14194856

Academic Editors: Yan Guo and Jinchuan Xing

Received: 2 September 2022
Accepted: 28 September 2022
Published: 4 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In the United States, cancer is one of the primary public health issues because of its high prevalence [1]. According to the World Health Organization (WHO), cancer is considered the most common cause of death across the globe, within which breast cancer and prostate cancer are the two most widespread cancer types affecting women and men, respectively [2]. Breast and prostate cancers are the most common sex-specific cancers. In 2020 alone, breast cancer affected more than 2.25 million individuals worldwide and claimed the lives of more than 680,000 women [2]. Prostate cancer affected more than 1.4 million men worldwide in 2020 [2].

In the biology of cancer, the tumor microenvironment (TME) consists of resident tissues, various inflammatory and stromal cells, and the ECM [3]. Tumors usually contain elevated levels of ECM markers [3]. Using the levels of ECM as a tumor indicator could aid in detecting cancer in its early stages and help in treatment as well [4]. The ECM has been identified as a mediator of drug resistance, therefore, attempting to modulate the ECM could improve the efficacy of cancer treatments [3]. Histopathology can be analyzed with measures of cell morphology and organization, like topological data analysis (TDA), in order to understand cellular interactions in the TME [5].

TDA [6,7] is a useful mathematical framework that assumes data points have an inherent relational structure. This makes it amenable to high-dimensional datasets like those used in biomedical research and has previously been used for image analysis, proteomics, and other omics techniques, but bioinformatic applications are relatively recent, so validated links from topological features to molecular function are still lacking [8]. Due to its efficiency in analyzing molecular profiles in cancer cells, advanced bioinformatics methods could be promising for cancer diagnosis and treatments within the field of oncology [9]. While most existing methods focus on analyzing the topology of gene expression data, our focus is on topology arising from the spatial layout of cells observed in an image (i.e., H&E-stained pathological specimens). Such image topological features have been found relevant to clinical outcomes [10,11].

To the best of our knowledge, this is the first publication linking topological features from cancer histology to spatial transcriptomics. Previously, imaging has been leveraged in spatial transcriptomics to infer gene expression at higher resolutions [12]. In prostate cancer, a convolutional neural network pre-trained on prostate needle biopsies was applied in order to extract image features from prostate spatial transcriptomic samples. These features were correlated with a factor analysis of gene expression, demonstrating how morphology colocalizes with certain gene sets in prostate spatial transcriptomic data [13]. Many methods use image features extracted from neural networks: for example, canonical correlation analysis to link bulk gene expression to embeddings from auto-encoders [14], and using a large amount of spatial transcriptomic samples to train and predict the spatial gene expression of several genes from only histopathological images [15].

In these analyses, we used correlation analysis to identify gene and ITF clusters in a similar fashion to gene co-expression module detection methods [16]. The association between gene expression and image topological features was determined by generating correlation matrices and visualization using heatmaps. Specifically, we applied methods to generate correlation matrices that allowed us to find correlations between gene expression and ITFs in breast and prostate cancer. Genes that correlate well with ITFs are called TAGs, i.e., *topology-associated genes*. Functional enrichment analysis is a useful technique for distinguishing protein and gene families related to both normal and abnormal body pathways [17]. We applied this technique to TAGs in order to find gene ontology terms that are significantly associated with ITFs from breast cancer and prostate cancer.

2. Materials and Methods

2.1. Datasets

We downloaded a total of six ST slides from the 10x Genomics Visium platform repository. Samples were either preserved by Formalin-Fixed Paraffin-Embedding (FFPE) or methanol fixation. They contain gene expression information, spatial information, and a corresponding tissue image. From these images, we extracted 1400 image topology features using topological data analysis techniques [7]. The slides that we used are FFPE Human

Breast Cancer, Parent Human Breast Cancer, Parent Visium Human Breast Cancer, Human Prostate Cancer, Prostate Acinar Cell Carcinoma, and Visium FFPE Human Normal Prostate. There are between 2000 and 4000 ST spots in each sample. Breast cancer samples were taken from the breast tissue of women older than 18 years old and the slides each contained more than 4000 genes after filtering out low-expression genes. Prostate cancer spatial transcriptomic samples contained between 2543 and 4371 genes after filtering. Summary of correlation analysis and clustering results across all datasets and image feature types can be found in Table S1.

2.2. Statistical and Correlation Analyses

R software was used for all statistical analyses. The correlation coefficients between image topology features and gene expression were used to identify TAGs. Of the genes with no more than 25% zero-read entries, the top 150 most variable genes in the ST data were selected for analysis. We selected 150 genes because they represent more than 50% of the total variance in the dataset, and this number was also large enough to provide statistical power in the functional enrichment software. Functional enrichment can help determine which proteins and genes have relationships with cancer [17]. We calculated the log-transform of gene expression and topology image feature values and then calculated the Pearson correlation coefficients between the ITFs and the genes. The *pheatmap* package [18] was used to visualize correlations between genes and ITFs.

2.3. ITF Generation

Image features were generated by a collaborating lab that specializes in computational pathology and TDA. They were extracted from 350 × 350-pixel image patches centered around the ST spots that corresponded to the same locations from which the gene expression was measured on the ST slide. The 350-pixel size was chosen to tile the image into non-overlapping squares so that features from one ST spot did not overlap with any others. We first identified cell locations from H&E images. Next, we used TDA methods to generate topological features. The cell locations can be visualized as points on a 2-dimensional scatter plot. The diameter of these points grows until they touch one another. As the diameter grows, more and more circles touch. When this happens, lines drawn from the circle centers can create shapes like triangles, squares, and rings. This is referred to as the "birth" of the shape. Eventually, the circle diameter is so large that all points are connected, drowning out shapes. The lifespan ("persistence") of these shapes as the circles continue to grow is known as "persistent homology". By scanning the connectedness of data at many different resolutions, persistent homology is a way to identify robust patterns in complex data from a very small to a very large scale. Persistent homology can be visualized in a persistence diagram, which can then be converted to a vector for analysis [19]. We focused on 0-dimensional and 1-dimensional topological components, which are clusters and rings, respectively.

Topology also has a distinct advantage over traditional convolutional neural networks (CNN): the features extracted by CNN kernels are task-specific, whereas topologies are not. For example, a CNN that classifies humans, dogs, and cats, may not provide features that are useful for histopathological images. Topology, instead, is a generalizable mathematical framework used to describe the shape of data and can be used to study the spatial relationships of cells on a local and global/tissue-wide level. Traditional convolutional neural networks cannot be used to process an entire whole slide image, so cannot directly represent tissue-wide architecture. Finally, because topology is not task-specific it allows us to easily compare results across cancers. The complementary functions of topology and convolutions have led to them being jointly implemented for the analysis of histopathology in cancer [20].

2.4. Clustering Correlated Gene Expression and Topology to Identify Sets of TAGs

As mentioned in 2.2, we identified TAGs by first calculating the Pearson correlation coefficient (PCC) between gene expression with topology across the entire spatial transcriptomics sample. To identify clusters of tightly correlated genes and topology features, we used the R package *pheatmap* [18] to identify clusters for rows (gene expression) and columns (image topology features) with a baseline of 10 clusters, then iteratively found the optimal number of clusters using the "elbow-method". K-means and the "elbow method" are commonly used to find the optimal number of clusters [21]. The optimal number of clusters was visualized with *pheatmap*. These clusters of topology-associated genes were used for functional enrichment, providing the link between topology (i.e., tissue architecture) and the transcriptional signatures of specific biological phenomena.

2.5. Functional Enrichment Analysis

The R package *gprofiler2* is a useful tool for performing functional analysis of gene sets [22]. This package provides a list of enrichment terms for each gene set and corresponding *p*-values. After adjusting *p*-values using the holm method, we used a significance level of 0.05 to select the gene sets with the lowest *p*-values, which are the highest significant terms.

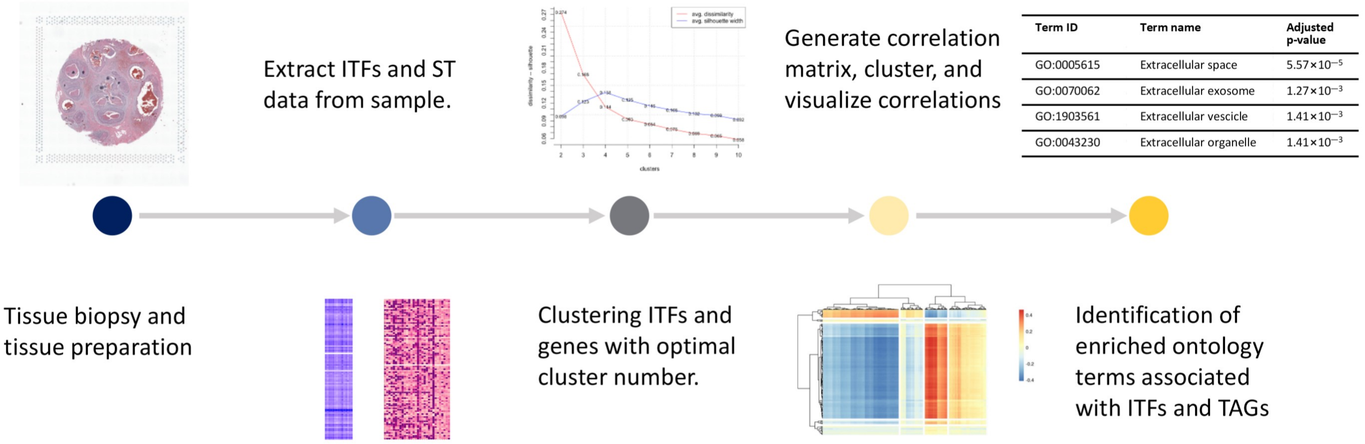
2.6. Integrative ITF Analysis in FFPE Human Breast Cancer ST Data

Based on a subset of image topology features identified through the previous correlation and functional enrichment analyses, we evaluated the relationships of the extracellular matrix ITFs with immune signaling. First, we identified a subset of 12 ITFs and 19 genes that identify ST spots related to immune signaling. This was achieved by looking at the common expression of immune-related on the ST array (e.g., major histocompatibility complex class I (MHC-I), CD8, and T-cell receptor (TCR) genes), and pathologist-defined tissue labels in order to identify ITFs most specific to these immune regions. We then used our recently developed tool *Diagnostic Evidence Gauge of Single-cells* (*DEGAS*) [23] with single-cell RNA sequencing data (scRNA-seq) from the triple-negative breast cancer (TNBC) samples [24] to estimate T-cell enrichment across the whole ST array. We also used *DEGAS* to infer ITFs in the single cells from the same TNBC dataset in order to identify which subsets of cells were more associated with specific ITFs. Finally, in the ST data, the 12 ITFs from each spot were used to train *XGBoost* [25] and *LightGBM* [26] classifiers to predict the individual expression of the 19 immune-signaling genes at that spot using leave-one-out cross-validation. The accuracy of the gene expression prediction was evaluated using the Pearson correlation coefficient between the predicted and true gene expression in the hold-out group (Figure S1).

3. Results

3.1. 1-Dimensional ITFs Correlate with Gene Sets and Functional Enrichment

We applied the workflow outlined in Figure 1. The 1-dimensional ITFs were found to have both positive and negative correlations to the 150 highest TAGs across all six of the ST slides used in this analysis (Figure 2). Furthermore, the ITFs tended to form distinct clusters presumably due to the ITFs describing similar topologies in the image. Based on the varying levels of correlation between ITF clusters and TAG clusters in the ST arrays, these similar topology ITFs can be described in terms of the associated TAG clusters. We identified several such TAG clusters in the breast cancer ST slides.



Extract ITFs and ST data from sample.

Generate correlation matrix, cluster, and visualize correlations

Term ID	Term name	Adjusted p-value
GO:0005615	Extracellular space	5.57×10^{-5}
GO:0070062	Extracellular exosome	1.27×10^{-3}
GO:1903561	Extracellular vescicle	1.41×10^{-3}
GO:0043230	Extracellular organelle	1.41×10^{-3}

Tissue biopsy and tissue preparation

Clustering ITFs and genes with optimal cluster number.

Identification of enriched ontology terms associated with ITFs and TAGs

Figure 1. Workflow of the method from tissue preparation to identifying gene ontology terms associated with ITFs.

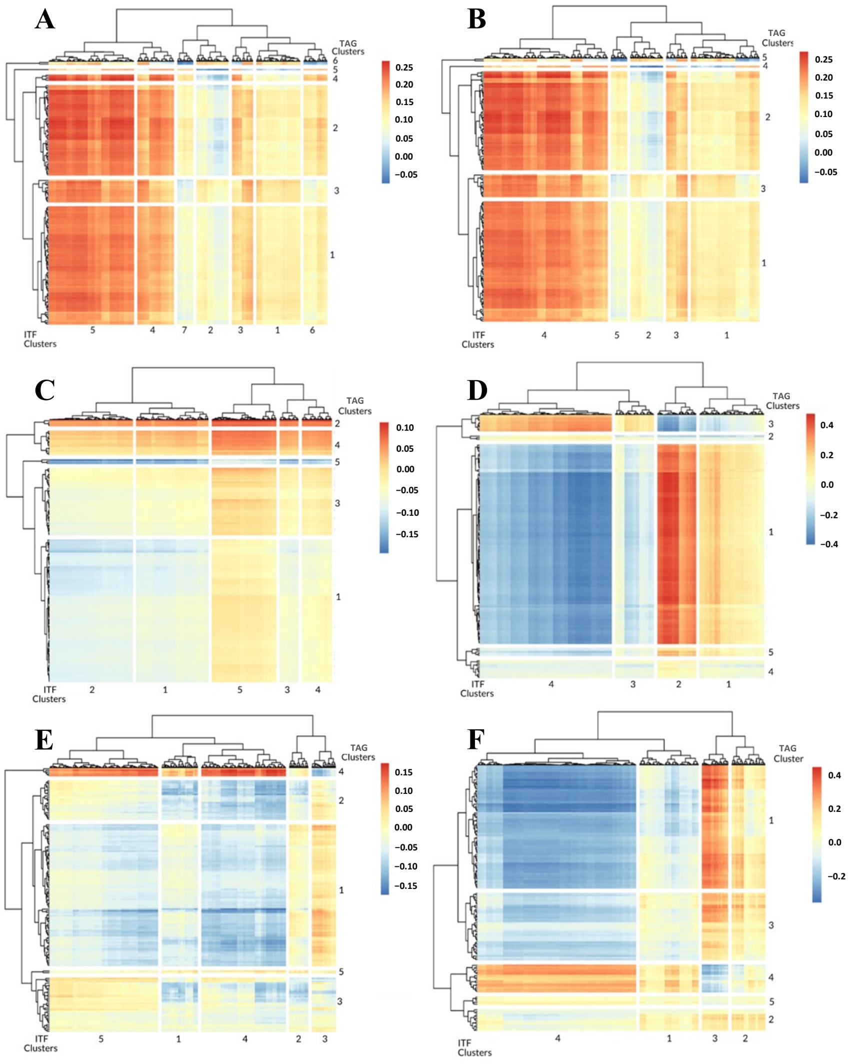


Figure 2. Heatmap of correlation matrix between 1-dimensional ITFs and TAG expression values for Parent Visium Human Breast Cancer (**A**), Parent Human Breast Cancer (**B**), FFPE Human Breast Cancer (**C**), Human Prostate Cancer (**D**), Prostate Acinar Carcinoma I (**E**), and Visium FFPE Human Normal Prostate (**F**).

From the Parent Visium Human Breast Cancer ST slide, we found that ITF clusters 1 and 2 were more correlated with TAG cluster 6 (Figure 2A, Table S2). ITF cluster 3 was the most similar to TAG cluster 2, representing the enrichment of the Mitochondrial inheritance, and ITF clusters 4 and 5 were the most similar to TAG cluster 4, representing the enrichment of the Collagen type I trimer (Figure 2A, Table S2). Lastly, ITF clusters 6 and 7 were the most similar to TAG cluster 5, representing the enrichment of the ZAG-PIP complex (Figure 2A, Table S2). In the Parent Human Breast Cancer slide, we found that ITF cluster 1 was the most similar to TAG cluster 3, representing the enrichment of the Dystroglycan binding (Figure 2B, Table S3). ITF cluster 2 was the most similar to TAG cluster 5, and ITF clusters 4 and 5 were more correlated with TAG cluster 2, representing the enrichment of Diabetic cardiomyopathy (Figure 2B, Table S3). Lastly, ITF cluster 5 was the most similar to TAG cluster 4, representing the enrichment of the ZAG-PIP complex (Figure 2B, Table S3). The FFPE Human Breast Cancer slide revealed that all of the ITFs were the most similar to TAG cluster 2 (Figure 2C, Table S4). It is also worth noting that the ECM TAG clusters were found in every one of the breast cancer ST slides and though it was not the highest ranked in terms of the PCC there were ITF clusters both positively and negatively correlated with it (Figure 2A,C). In addition to the breast cancer ST slides, there were also clear associations with TAG clusters in the prostate cancer slides.

In the Human Prostate Cancer slide, ITF clusters 1 and 2 were the most similar clusters to TAG cluster 1, representing the enrichment of the Extracellular exosome (Figure 2D, Table S5). ITF clusters 3 and 4 were more correlated with TAG cluster 3, representing the enrichment of the Smooth Muscle Contraction (Figure 2D, Table S5). Likewise, in the Prostate Acinar Carcinoma slide, ITF cluster 3 was the most similar to TAG cluster 1, representing the enrichment of the Extracellular exosome, and ITF clusters 1, 4, and 5 were the most similar to TAG cluster 4, representing the enrichment of the Smooth Muscle Contraction (Figure 2E, Table S6). ITF cluster 2 was the most similar to TAG cluster 5, representing the enrichment of the Insulin-like growth factor binding (Figure 2E, Table S6). In the normal prostate slide, Visium FFPE Human Prostate, and ITF clusters 1 and 4 were the most similar clusters to TAG cluster 4, representing the enrichment of the Smooth Muscle Contraction (Figure 2F, Table S7). ITF cluster 2 was most correlated with TAG cluster 2, representing the enrichment of the Collagen-containing extracellular matrix, and ITF cluster 3 was most correlated with TAG cluster 1, representing the enrichment of the Extracellular exosome (Figure 2F, Table S7).

We noted that the most significant association among all prostate ST slides was that the 1-dimensional image topology features (ITF81-ITF231) were related to the extracellular matrix. The ECM in the Human Prostate Cancer slide was gene cluster 1, which was the most similar to ITF cluster 1 (ITF1-ITF97 and ITF197-264) and ITF cluster 2 (ITF98-ITF196). In the Prostate Acinar Cell Carcinoma slide, the ECM elements were in TAG clusters 1 and 2, but only TAG cluster 1 had a high similarity with ITF cluster 3 (ITF98-ITF159). The ECM ontology terms were found in the Visium FFPE Human Normal Prostate slide TAG clusters 1, 2, and 3 as well. ITF cluster 3 was the most similar to TAG cluster 1 and ITF cluster 2 was the most similar to TAG cluster 2. Clearly, there was much overlap in 1-dimensional ITFs as they relate to the ECM.

3.2. 0-Dimensional ITFs Correlate with Gene Sets and Functional Enrichment

In the prostate ST slides, the 0-dimensional ITFs did not associate nearly as much with the ECM as the 1-dimensional ITFs and, instead, regularly associated with smooth muscle-related TAG clusters. The 0-dimensional ITFs showed many similarities when compared to the 1-dimensional ITFs in the breast cancer ST slides. For instance, in the breast tissue ST slides, the ZAG-PIP complex also appeared to be associated with the 0-dimensional ITF clusters.

From the Parent Visium Human Breast Cancer slide, among the most significant correlations, we found that ITF cluster 1 was the most similar to TAG cluster 3, representing the enrichment of Mitochondrial inheritance (Figure 3A, Table S2). ITF clusters 2, 3, and 4 were most associated with TAG cluster 5, representing the enrichment of the ZAG-PIP complex (Figure 3A, Table S2). The Parent Human Breast Cancer slide identified ITF cluster 1 as the most similar cluster to TAG cluster 3, representing the enrichment of the Dystroglycan binding. In contrast, ITF clusters 2, 3, 4, and 5 were the most correlated with TAG cluster 5, representing the ZAG-PIP complex (Figure 3B, Table S3). In the FFPE Human Breast Cancer slide, all of the ITF clusters were the most similar to TAG cluster 2, representing the enrichment of the Centrocecal scotoma (Figure 3C, Table S4). Unlike the similarity between 1-dimensional and 0-dimensional ITF functional assessment, the prostate cancer 0-dimensional ITFs reflected different functions than their 1-dimensional counterparts.

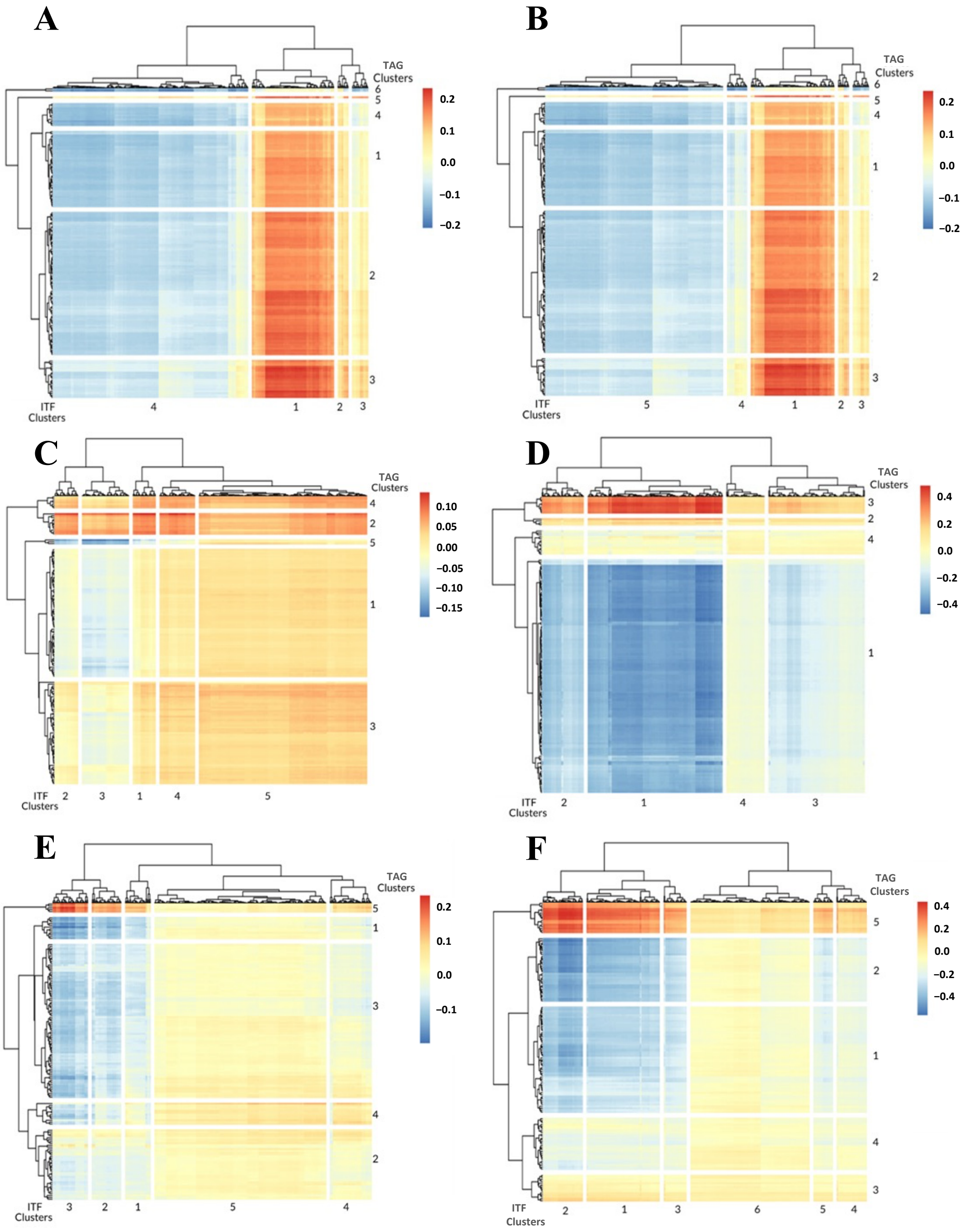


Figure 3. Heatmap of correlation matrix between 0-dimensional image features and gene expression values for Parent Visium Human Breast Cancer (**A**), Parent Human Breast Cancer (**B**), FFPE Human Breast Cancer (**C**), Human Prostate Cancer (**D**), Prostate Acinar CarcinoI (**E**), and Visium FFPE Human Normal Prostate (**F**).

In the Human Prostate Cancer slide, all four ITF clusters were the most similar to TAG cluster 3, representing the enrichment of the Smooth Muscle Contraction (Figure 3D, Table S5). In the Prostate Acinar Cell Carcinoma slide, ITF clusters 1, 2, 3, and 4 were the most similar to TAG cluster 5, representing the enrichment of the Smooth Muscle Contraction (Figure 3E, Table S6). ITF cluster 5 was most correlated with TAG cluster 4, representing the enrichment of the IgA immunoglobulin complex (Figure 3E, Table S6). Like the Human Prostate Cancer slide, the Visium FFPE Human Normal Prostate slide was readily associated with the smooth muscle because all of the ITF clusters were the most similar to TAG cluster 5 which represented Smooth Muscle Contraction (Figure 3F, Table S7). In prostate ST tissues, the 1-dimensional ITFs appeared to be more informative of biology, evaluated by the number of unique TAG ontology terms, than the 0-dimensional ITFs.

3.3. Integrative Analysis Reveals Immune Signaling in the FFPE Human Breast Cancer Slide

Due to the high prevalence of ECM-related terms found in the preceding analyses and the lack of top rank mappings between the ITFs and the ECM TAG clusters in breast cancer ST slides, we next evaluated the subset of 1-dimensional ITFs in the FFPE Human Breast Cancer ST dataset with respect to immune signaling. A detailed look at the accompanying image to the ST slide shows that some locations in the tumor have a denser immune infiltrate than others (Figure 4A). One mechanism by which tumor cells can be detected and attacked by the immune system is through antigen presentation on MHC-I, which is subsequently detected by TCR on CD8 T-cells (Figure 4B). From this ST data, we estimated the relative abundance of T-cells using our own tool, *DEGAS* [23], and scRNA-seq from TNBC samples [24] (Figure 4C). We identified these T-cells as containing many CD8 T-cells by the abundance of CD8 in those regions (Figure 4D). The adjacent tumor region expressed high levels of MHC-I (Figure 4E) and the identified T-cells expressed higher levels of TCR genes compared to other lymphocyte-rich areas (Figure 4F). These gene expression signals seem to indicate a large number of cytotoxic CD8 T-cells attacking that specific tumor region denoted by an arrow in Figure 4. We see that the topology of the cells that are measured using ITF121 is also unique in this region, possibly denoting some arrangement of cells that influences T-cell response to the HLA-I antigen presentation (Figure 4G). We also found that when we predicted the value of ITF121 in the TNBC scRNA-seq [24] using *DEGAS* [23], epithelial cells expressing higher levels of MHC-I genes tended to have a higher association with ITF121 (p-value = 3.3×10^{-5}). From these analyses, it is clear that cellular topology, measured by ITFs, is associated with different cell types and immune signaling in the TME and the ECM. The associations between tumor-infiltrating lymphocytes (TIL) and prognosis are complicated. Tumor-infiltrating lymphocytes in Ductal Carcinoma have been studied previously. High TIL densities were associated with a higher nuclear grade, more necrosis, and a higher rate of tumor recurrence. These same authors note an opposite direction of effect in triple-negative breast cancer (the most aggressive subtype) where high levels of TILs were associated with a better prognosis [27]. The topology could be applied to identify complex patterns of tumor-infiltrating lymphocytes in cancer. This is a field of active research [5,28,29].

As shown in Figure 5, we generated a correlation matrix, clustered it on the paired ITFs and TAGs using *pheatmap* (Figure 5A), then predicted the TAG expression from the ITFs by training both an AdaBoost and an *XGBoost* regression model (Table S8). The mean AdaBoost correlation (0.524) was 0.014 greater than the *XGBoost* (0.510) during leave-one-out cross-validation but was not significant (p-value = 0.07, Table S8). The ITF to TAG correlation matrix shows multiple immune-related genes (Figure 5A). Some of these genes such as HLA-E and HLA-C are directly related to MHC-I signaling by the cancer cell. The gene STAT1 is related to T-cells and immune signaling. In Figure 5B, we further performed leave-one-out cross-validation using *LightGBM* in order to predict the expression levels of each of the 19 genes based on the 12 image features. A *LightGBM* model was trained on all but one of the spots in the ST slide, reserving the last remaining spot for testing. This was repeated for each of the 19 genes. The correlations between the predicted gene expression values and the true expression values across the 19 genes ranged from 0.49 to 0.58 (Figure 5B). These results on the 10X Genomics Visium FFPE Human Breast Cancer slide showed that ITFs represent important cellular immune-tumor architectures that correspond to actual distinct molecular events and cell types and can be used to identify cell–cell interactions.

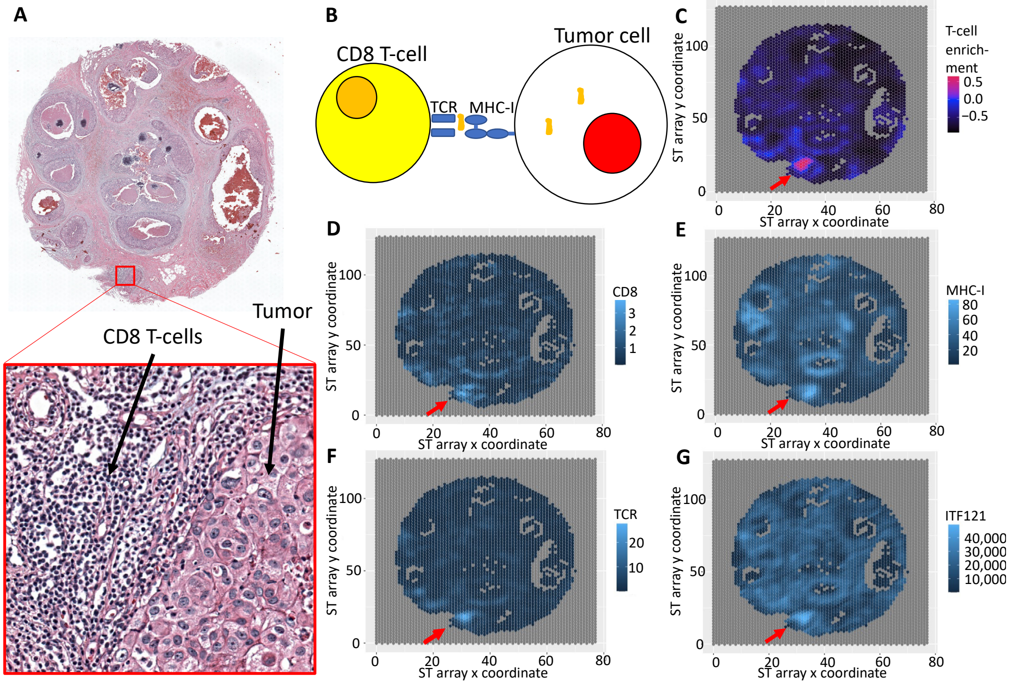


Figure 4. ITFs associated with cellular signaling. Here, 1-dimensional ITF121 is associated with CD8 T-cell interaction with tumor regions likely through MHC-I. Note that the periphery of the tumor region has greater ITF121. (**A**) H&E pathology image from an ST array. (**B**) Simplified diagram of T-cell and tumor signaling via MHC-I. (**C**) T-cell enrichment using DEGAS. Expression of marker genes (**D**) CD8, (**E**) MHC-I (HLA-C), and (**F**) TCR (TRAC). (**G**) ITF121 was calculated for an image patch around each ST spot. Red arrows indicate region enriched for CD8 T-cells interacting with the tumor.

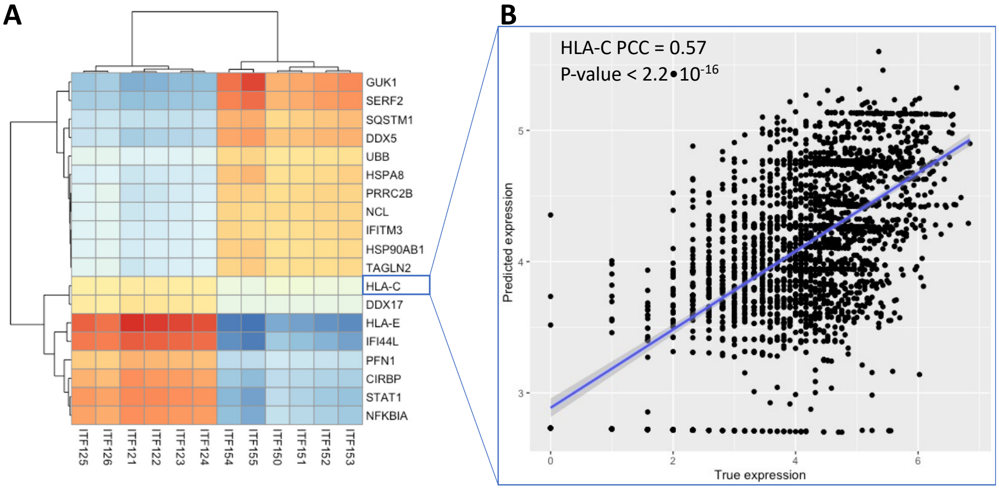


Figure 5. Results for predicting expression levels of TAGs using machine learning models. (**A**) Correlation between TAG expression and ITFs. (**B**) Prediction of MHC-I (HLA-C) expression using an *LightGBM* model trained on the 12 ITFs in (**A**).

4. Discussion

All ST slides had sets of genes significantly correlated with topology image features. These gene sets were enriched for several ontology terms. Five ECM-related gene terms are common to the samples from both cancers: *extracellular space*, *extracellular exosome*, *extracellular vesicle*, *extracellular organelle*, and *extracellular membrane-bounded organelle*. These terms had the most significant p-values of the functional enrichments.

The ECM plays a significant role in tissue and organ organization and performs various functions such as cell homeostasis, differentiation, and many other physical and biochemical functions [30]. Cancer is driven by the unregulated growth of clonal cells which can evade immune system response [31]. The ECM has been shown to enhance cancer growth by changing cell functions, which can make tissues more fertile environments for cancer cells to proliferate [32]. The ECM plays a role in the progression of cancer, where the composition, toughness, hydration, and adherence to the ECM components of cells in the tumor and microenvironment evolve during breast cancer development [33]. Extracellular vesicles and extracellular exosomes can mediate cell communication and have been shown to contribute to the spread of prostate cancers [34]. Collagen plays an important part in cancer progression, sometimes contributing to the development of cancer cells and the enhancement of tumor growth [35]. Collagen Type I Trimer was significantly enriched in our analysis. Cancer-associated fibroblasts produce collagen type I on which cancer cells can migrate using fibronectin, and drugs targeting the ECM are an evolving therapeutic avenue [35].

When the FFPE Human Breast Cancer slide was studied in more detail for immune-related signaling that could impact the TME, we identified image features that appear to be predictive of immune-related genes (e.g., MHC-I). Furthermore, the localization of some image features appears to closely mirror immune signaling molecules: CD8, TCR, and MHC-I expression have a very similar distribution to the 1-dimensional ITF121. It is worth noting that ITF121 was also commonly associated with ECM in the prostate ST slides, along with ITF81-ITF231. In the Prostate Acinar Cell Carcinoma sample, we found enrichment for the IgA immunoglobulin complex. The prostate excretes IgA, and changes in IgA expression could be indicative of abnormal prostate immunity [36]. Immunoglobulins can appear at high levels in some cancers [37] and have been shown to be markers for an immunosuppressive subpopulation of B-cells in prostate cancer [38].

We identified the enrichment for the ZAG-PIP complex in the topology-associated genes from the Parent Visium Human Breast Cancer and Parent Human Breast Cancer ST slides. This is a complex of Zinc alpha2-glycoprotein (ZAG) with Prolactin inducible protein (PIP). ZAG has been studied more in prostate cancer, where this secretory protein has androgen-response enhancer elements and pushes the G1/S cell cycle transition [39]. PIP is also inducible by androgens, plays a role in the adaptive immune response, and is a marker that can be used to identify metastatic foci as having originated from breast tissue [40]. In Pip deficient mice, researchers found abnormal lymphoid proliferations in prostatic tissues and hypertrophy of the thymus medulla [41].

The chief advantage of this study is the novel use of ITFs to describe the cellular topologies associated with gene expression using spatial transcriptomics. Unlike convolutional neural networks, topological features can explicitly describe tissue architecture. We use the term tissue "architecture" because topology describes the shape of networks of cells through the study of "persistent homology" (described previously), providing a complement to measuring the morphology of individual cells which our group has studied extensively previously [42,43]. The chief limitation of our study is in not having visualized topological features for interpretation by a pathologist. This is of great interest to us, as we plan to study it using the deep-transfer-learning framework *DEGAS* [23] to identify high-risk topologies and visualize them for histopathological interpretation.

5. Conclusions

In this study, we took advantage of ST in order to determine how gene expression is related to cellular topology in breast and prostate cancers. We have found TAGs that are common in both cancers and some that are unique to each cancer. ECM-related enrichment terms are the most common terms in both the breast cancer and prostate cancer samples. In the prostate ST slides, 1-dimensional Image Topological Features ITF81-ITF231 were the most tightly correlated with ECM gene clusters. Among the breast cancer slides, ZAG-PIP Complex gene expression was shared. In the prostate cancer slides, enrichment for Gastrointestinal Dysmotility gene expression and IgA immunoglobin gene sets were shared across samples. The significance of this work is in demonstrating the utility of topology features in order to describe tissue architecture, and that ITFs can be correlated with gene expression. Gene sets that are correlated with ITFs are significantly enriched for biological functions. We do not contend that these biological functions are guaranteed to be present in the tissue. Instead, we introduced the important benefit of topology as means to rigorously quantify tissue architecture. Pathological assessment is usually largely qualitative (or, at best, semi-quantitative), therefore, quantification of tissue architecture is necessary in order to take advantage of histopathological information for integrative omics research. The histopathology of cancer is crucially important for the diagnosis and prognosis of patients–the amount and configuration of reactive tumor stroma, the spatial arrangement of TILs, and the density of tumor cells all have prognostic implications for patients and underlying genetic correlates. Here, we have demonstrated that topology is a promising technique to quantify the complex, high-dimensional, and highly variable characteristics of histopathological specimens, that can be effectively integrated with other biomedical data.

We can conclude that cellular topology has clear associations with the gene expression profiles from breast and prostate cancers. Understanding the biological meaning behind image features describing cellular topology can lead to improved image feature extraction methods, predictive models based on histopathological images, and integrative omics and image analysis, which may better describe the relationships between molecular events and the spatial arrangements of cells in the tumor and microenvironment.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14194856/s1, Figure S1: Pearson correlation between *XGBoost*-predicted gene expression based on image topology features and measured gene expression, Table S1: Summary of correlation analysis and clustering results across all datasets and image feature types, Table S2: Functional enrichment for TAGs in Parent Visium Human Breast Cancer Slide (1-dimensional Image Topological Features and 0-dimensional Image Topological Features) with the lowest significant *p*-values, Table S3: Functional enrichment for TAGs in Parent Human Breast Cancer Slide (1-dimensional Image Topological Features and 0-dimensional Image Topological Features) with the lowest significant *p*-values, Table S4: Functional enrichment for TAGs in FFPE Human Breast Cancer Slide (1-dimensional Image Topological Features and 0-dimensional Image Topological Features) with the lowest significant *p*-values, Table S5: Functional enrichment for TAGs in Human Prostate Cancer Slide (1-dimensional Image Topological Features and 0-dimensional Image Topological Features) with the lowest significant *p*-values, Table S6: Functional enrichment for TAGs in Prostate Acinar Cell Carcinoma Slide (1-dimensional Image Topological Features and 0-dimensional Image Topological Features) with the lowest significant *p*-values, Table S7: Functional enrichment for TAGs in Visium FFPE Human Normal Prostate Slide (1-dimensional Image Topological Features and 0-dimensional Image Topological Features) with the lowest significant *p*-values, and Table S8: Pearson correlation coefficients for predicted gene expression vs. true gene expression from *XGBoost* and *LightGBM* models.

Author Contributions: T.S.J. conceptualized the manuscript; C.C. conceptualized the image features; T.S.J., C.C., J.Z. and K.H. conceptualized the integration methodologies; T.S.J. and C.C. supervised the experiments and the manuscript preparation; Z.Y. and C.L. extracted image features; L.A. performed all integrative analyses; L.A., J.L.C., K.H., J.Z., C.C. and T.S.J. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this research included: Indiana University School of Medicine Precision Health Initiative to T.S.J., J.Z., J.L.C. and K.H.; American Cancer Society (19-144-34) to T.S.J.; National Cancer Institute R21264339 to T.S.J. and J.Z.; National Institute of General Medical Sciences R01GM148970 to C.C., J.Z. and T.S.J.; National Science Foundation IIS-1909038 to C.C.; National Cancer Institute R01CA253368 to C.C.; and Indiana Biosciences Research Institute to T.S.J.; fellowship funding UL1TR002529 (CTSI TL1, NIH, S. Moe, S. Wiehe) to J.L.C.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All of the spatial transcriptomics data and related images are freely available on the 10X Genomics data commons.

Acknowledgments: We would like to acknowledge the help and support of scientists at the Indiana Biosciences Research Institute, the Indiana University School of Medicine Center for Medical Genomics, Center for Computational Biology and Bioinformatics, and the Melvin and Bren Simon Comprehensive Cancer Center.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. NCI. *Public Health Research and Cancer*; National Cancer Institute: Bethesda, MD, USA, 2020.
2. WHO. *Cancer*; World Health Organization: Geneva, Switzerland, 2022.
3. Henke, E.; Nandigama, R.; Ergün, S. Extracellular Matrix in the Tumor Microenvironment and Its Impact on Cancer Therapy. *Front. Mol. Biosci.* **2020**, *6*, 160. [CrossRef] [PubMed]
4. Brassart-Pasco, S.; Brézillon, S.; Brassart, B.; Ramont, L.; Oudart, J.-B.; Monboisse, J.C. Tumor Microenvironment: Extracellular Matrix Alterations Influence Tumor Progression. *Front. Oncol.* **2020**, *10*, 397. [CrossRef] [PubMed]
5. Lu, Z.; Xu, S.; Shao, W.; Wu, Y.; Zhang, J.; Han, Z.; Feng, Q.; Huang, K. Deep-Learning–Based Characterization of Tumor-Infiltrating Lymphocytes in Breast Cancers From Histopathology Images and Multiomics Data. *JCO Clin. Cancer Inform.* **2020**, *4*, 480–490. [CrossRef] [PubMed]
6. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **2009**, *46*, 255–308. [CrossRef]
7. Edelsbrunner, H.; Harer, J.L. *Computational Topology: An Introduction*; American Mathematical Society: Providence, RI, USA, 2022.
8. Loughrey, C.F.; Fitzpatrick, P.; Orr, N.; Jurek-Loughrey, A. The topology of data: Opportunities for cancer research. *Bioinformatics* **2021**, *37*, 3091–3098. [CrossRef]
9. Singer, J.; Irmisch, A.; Ruscheweyh, H.-J.; Singer, F.; Toussaint, N.C.; Levesque, M.P.; Stekhoven, D.J.; Beerenwinkel, N. Bioinformatics for precision oncology. *Brief. Bioinform.* **2019**, *20*, 778–788. [CrossRef]
10. Aukerman, A.; Carrière, M.; Chen, C.; Gardner, K.; Rabadán, R.; Vanguri, R. Persistent homology based characterization of the breast cancer immune microenvironment: A feasibility study. In Proceedings of the 36th International Symposium on Computational Geometry (SoCG), Zürich, Switzerland, 23–26 June 2020.
11. Lawson, P.; Sholl, A.B.; Brown, J.Q.; Fasy, B.T.; Wenk, C. Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology. *Sci. Rep.* **2019**, *9*, 1139. [CrossRef]
12. Bergenstråhle, L.; He, B.; Bergenstråhle, J.; Abalo, X.; Mirzazadeh, R.; Thrane, K.; Ji, A.L.; Andersson, A.; Larsson, L.; Stakenborg, N.; et al. Super-resolved spatial transcriptomics by deep data fusion. *Nat. Biotechnol.* **2022**, *40*, 476–479. [CrossRef]
13. Chelebian, E.; Avenel, C.; Kartasalo, K.; Marklund, M.; Tanoglidi, A.; Mirtti, T.; Colling, R.; Erickson, A.; Lamb, A.D.; Lundeberg, J.; et al. Morphological Features Extracted by AI Associated with Spatial Transcriptomics in Prostate Cancer. *Cancers* **2021**, *13*, 4837. [CrossRef]
14. Ash, J.T.; Darnell, G.; Munro, D.; Engelhardt, B.E. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nat. Commun.* **2021**, *12*, 1609. [CrossRef]
15. He, B.; Bergenstråhle, L.; Stenbeck, L.; Abid, A.; Andersson, A.; Borg, Å.; Maaskola, J.; Lundeberg, J.; Zou, J. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **2020**, *4*, 827–834. [CrossRef] [PubMed]
16. Liu, Y.; Ye, X.; Yu, C.Y.; Shao, W.; Hou, J.; Feng, W.; Zhang, J.; Huang, K. TPSC: A module detection method based on topology potential and spectral clustering in weighted networks and its application in gene co-expression module discovery. *BMC Bioinform.* **2021**, *22*, 111. [CrossRef] [PubMed]
17. Thanati, F.; Karatzas, E.; Baltoumas, F.; Stravopodis, D.; Eliopoulos, A.; Pavlopoulos, G. FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists. *Biology* **2021**, *10*, 665. [CrossRef] [PubMed]
18. Kolde, R. *Pheatmap: Pretty Heatmaps*; CRAN: Vienna, Austria, 2019.
19. GUDHI. GUDHI Library for TDA. 2021. Available online: https://gudhi.inria.fr/ (accessed on 1 September 2022).
20. Levy, J.; Haudenschild, C.; Barwick, C.; Christensen, B.; Vaickus, L. Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks. In *Biocomputing 2021, Proceedings of the Pacific Symposium, Kohala Coast, HI, USA, 3–7 January 2021*; World Scientific: Singapore, 2021.

21. Clustering Metrics Better than the Elbow-Method. Available online: https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6 (accessed on 1 September 2022).
22. Kolberg, L.; Raudvere, U.; Kuzmin, I.; Vilo, J.; Peterson, H. Gprofiler2—An R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. *F1000Research* **2020**, *9*, ELIXIR-709. [CrossRef]
23. Johnson, T.S.; Yu, C.Y.; Huang, Z.; Xu, S.; Wang, T.; Dong, C.; Shao, W.; Zaid, M.A.; Huang, X.; Wang, Y.; et al. Diagnostic Evidence GAuge of Single cells (DEGAS): A flexible deep transfer learning framework for prioritizing cells in relation to disease. *Genome Med.* **2022**, *14*, 11. [CrossRef]
24. Karaayvaz, M.; Cristea, S.; Gillespie, S.M.; Patel, A.P.; Mylvaganam, R.; Luo, C.C.; Specht, M.C.; Bernstein, B.E.; Michor, F.; Ellisen, L.W. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **2018**, *9*, 3588. [CrossRef]
25. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
26. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
27. Darvishian, F.; Ozerdem, U.; Adams, S.; Chun, J.; Pirraglia, E.; Kaplowitz, E.; Guth, A.; Axelrod, D.; Shapiro, R.; Price, A.; et al. Tumor-Infiltrating Lymphocytes in a Contemporary Cohort of Women with Ductal Carcinoma In Situ (DCIS). *Ann. Surg. Oncol.* **2019**, *26*, 3337–3343. [CrossRef]
28. Acs, B.; Ahmed, F.S.; Gupta, S.; Wong, P.F.; Gartrell, R.D.; Pradhan, J.S.; Rizk, E.M.; Rothberg, B.G.; Saenger, Y.; Rimm, D.L. An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma. *Nat. Commun.* **2019**, *10*, 5440. [CrossRef]
29. Failmezger, H.; Muralidhar, S.; Rullan, A.; de Andrea, C.E.; Sahai, E.; Yuan, Y. Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immuno-suppression in Melanoma Histology. *Cancer Res* **2020**, *80*, 1199–1209. [CrossRef]
30. Frantz, C.; Stewart, K.M.; Weaver, V.M. The extracellular matrix at a glance. *J. Cell Sci.* **2010**, *123*, 4195–4200. [CrossRef]
31. National Cancer Institute. *What Is Cancer?* National Cancer Institute: Bethesda, MD, USA, 2021.
32. Nallanthighal, S.; Heiserman, J.P.; Cheon, D.-J. The Role of the Extracellular Matrix in Cancer Stemness. *Front. Cell Dev. Biol.* **2019**, *7*, 86. [CrossRef] [PubMed]
33. Zhao, Y.; Zheng, X.; Zheng, Y.; Chen, Y.; Fei, W.; Wang, F.; Zheng, C. Extracellular Matrix: Emerging Roles and Potential Therapeutic Targets for Breast Cancer. *Front. Oncol.* **2021**, *11*, 650453. [CrossRef] [PubMed]
34. Vlaeminck-Guillem, V. Extracellular Vesicles in Prostate Cancer Carcinogenesis, Diagnosis, and Management. *Front. Oncol.* **2018**, *8*, 222. [CrossRef] [PubMed]
35. Xu, S.; Xu, H.; Wang, W.; Li, S.; Li, H.; Li, T.; Zhang, W.; Yu, X.; Liu, L. The role of collagen in cancer: From bench to bedside. *J. Transl. Med.* **2019**, *17*, 309. [CrossRef] [PubMed]
36. Silva, J.A.F.; Biancardi, M.F.; Stach-Machado, D.R.; Reis, L.O.; Sant'Anna, O.A.; Carvalho, H.F. The origin of prostate gland-secreted IgA and IgG. *Sci. Rep.* **2017**, *7*, 16488. [CrossRef]
37. Cui, M.; Huang, J.; Zhang, S.; Liu, Q.; Liao, Q.; Qiu, X. Immunoglobulin Expression in Cancer Cells and Its Critical Roles in Tumorigenesis. *Front. Immunol.* **2021**, *12*, 613530. [CrossRef]
38. Zhong, Z.; Nan, K.; Weng, M.; Yue, Y.; Zhou, W.; Wang, Z.; Chu, Y.; Liu, R.; Miao, C. Pro- and Anti- Effects of Immunoglobulin A-Producing B Cell in Tumors and Its Triggers. *Front. Immunol.* **2021**, *12*, 765044. [CrossRef]
39. Cao, R.; Ke, M.; Wu, Q.; Tian, Q.; Liu, L.; Dai, Z.; Lu, S.; Liu, P. AZGP1 is androgen responsive and involved in AR-induced prostate cancer cell proliferation and metastasis. *J. Cell Physiol.* **2019**, *234*, 17444–17458. [CrossRef]
40. Edechi, C.A.; Ikeogu, N.M.; Akaluka, G.N.; Terceiro, L.E.L.; Machado, M.; Salako, E.S.; Barazandeh, A.F.; Kung, S.K.P.; Uzonna, J.E.; Myal, Y. The Prolactin Inducible Protein Modulates Antitumor Immune Responses and Metastasis in a Mouse Model of Triple Negative Breast Cancer. *Front. Oncol.* **2021**, *11*, 639859. [CrossRef]
41. Blanchard, A.; Nistor, A.; Castaneda, F.; Martin, D.; Hicks, G.; Amara, F.; Shiu, R.; Myal, Y. Generation and initial characterization of the prolactin-inducible protein (PIP) null mouse: Accompanying global changes in gene expression in the submandibular gland. *Can. J. Physiol. Pharmacol.* **2009**, *87*, 859–872. [CrossRef]
42. Cheng, J.; Han, Z.; Mehra, R.; Shao, W.; Cheng, M.; Feng, Q.; Ni, D.; Huang, K.; Cheng, L.; Zhang, J. Computational analysis of pathological images enables a better diagnosis of TFE3 Xp11.2 translocation renal cell carcinoma. *Nat. Commun.* **2020**, *11*, 1778. [CrossRef]
43. Cheng, J.; Zhang, J.; Han, Y.; Wang, X.; Ye, X.; Meng, Y.; Parwani, A.; Han, Z.; Feng, Q.; Huang, K. Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. *Cancer Res.* **2017**, *77*, e91–e100. [CrossRef]