SPATIAL TRANSCRIPTOMICS ANALYSIS REVEALS TRANSCRIPTOMIC AND CELLULAR TOPOLOGY ASSOCIATIONS IN BREAST AND PROSTATE CANCERS

Lujain Alsaleh

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Biostatistics and Health Data Science,
Indiana University

May 2022

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

_____

Travis S. Johnson, PhD, Chair

_____

William Fadel, PhD

_____

Wanzhu Tu, PhD

ACKNOWLEDGEMENT

Thanks to Allah, who enabled me with strength to finish this thesis successfully.

I would like to express my sincere gratitude to my thesis advisor, Dr. Johnson for his invaluable assistance leading to the complete this paper. Dr. Johnson guidance, expertise, time, and encouragement are greatly appreciated.

Finally, deepest thanks to my caring, loving, and supportive family. They always keep me confident and motivated to achieve my goals, especially during this journey. I am grateful for my parents for their unfailing love and constant support. Heartfelt thanks to my siblings for their continuous encouragement and endless support.

Lujain Alsaleh

SPATIAL TRANSCRIPTOMICS ANALYSIS REVEALS TRANSCRIPTOMIC AND CELLULAR TOPOLOGY ASSOCIATIONS IN BREAST AND PROSTATE CANCERS

Background: Cancer is the leading cause of death worldwide and as a result is one of the most studied topics in public health. Breast cancer and prostate cancer are the most common cancers among women and men respectively. Gene expression and image features are independently prognostic of patient survival. However, it is sometimes difficult to discern how the molecular profile, e.g., gene expression, of given cells relate to their spatial layout, i.e., topology, in the tumor microenvironment (TME). However, with the advent of spatial transcriptomics (ST) and integrative bioinformatics analysis techniques, we are now able to better understand the TME of common cancers.

Method: In this paper, we aim to determine the genes that are correlated with image topology features (ITFs) in common cancers which we denote topology associated genes (TAGs). To achieve this objective, we generate the correlation coefficient between genes and image features after identifying the optimal number of clusters for each of them. Applying this correlation matrix to heatmap using R package pheatmap to visualize the correlation between the two sets. The objective of this study is to identify common themes for the genes correlated with ITFs and we can pursue this using functional enrichment analysis. Moreover, we also find the similarity between gene clusters and some image features clusters using the ranking of correlation coefficient in order to identify, compare and contrast the TAGs across breast and prostate cancer ST slides.

Result: The analysis shows that there are groups of gene ontology terms that are common within breast cancer, prostate cancer, and across both cancers. Notably, extracellular matrix (ECM) related terms appeared regularly in all ST slides.

Conclusion: We identified TAGs in every ST slide regardless of cancer type. These TAGs were enriched for ontology terms that add context to the ITFs generated from ST cancer slides.

Travis S. Johnson, PhD, Chair

William Fadel, PhD

Wanzhu Tu, PhD

TABLE OF CONTENTS

List of Tables viii
List of Figures ix
Chapter One: Background 1
Chapter Two: Methods 3
Datasets 3
Image Features 4
Clustering 5
Functional Enrichment 5
Chapter Three: Results 6
Parent Visium Human Breast Cancer Slide 6
Parent Human Breast Cancer Slide 10
FFPE Human Breast Cancer Slide 14
Human Prostate Cancer Slide 19
Prostate Acinar Cell Carcinoma Slide 25
Visium FFPE Human Normal Prostate Slide 29
Chapter Four: Discussion 35
Chapter Five: Conclusion 38
References 39
Curriculum Vitae

LIST OF TABLES

Table 1: Functional enrichment for genes in Parent Visium Human Breast Cancer Slide (1st Order Image Features) with the lowest significant p-values 8
Table 2: Functional enrichment for genes in Parent Visium Human Breast Cancer Slide (0th Order Image Features) with the lowest significant p-values 10
Table 3: Functional enrichment for genes in Parent Human Breast Cancer Slide (1st Order Image Features) with the lowest significant p-values.....12
Table 4: Functional enrichment for genes in Parent Human Breast Cancer Slide (0th Order Image Features) with the lowest significant p-values 14
Table 5: Functional enrichment for genes in FFPE Human Breast Cancer Slide (1st Order Image Features) with the lowest significant p-values.....16
Table 6: Functional enrichment for genes in FFPE Human Breast Cancer Slide (0th Order Image Features) with the lowest significant p-values 18
Table 7: Functional enrichment for genes in Human Prostate Cancer Slide (1st Order Image Features) with the lowest significant p-values.....21
Table 8: Functional enrichment for genes in Human Prostate Cancer Slide (0th Order Image Features) with the lowest significant p-values.....24
Table 9: Functional enrichment for genes in Prostate Acinar Cell Carcinoma Slide (1st Order Image Features) with the lowest significant p-values.....26
Table 10: Functional enrichment for genes in Prostate Acinar Cell Carcinoma Slide (0th Order Image Features) with the lowest significant p-values.....28
Table 11: Functional enrichment for genes in Visium FFPE Human Normal Prostate Slide (1st Order Image Features) with the lowest significant p-values.....31
Table 12: Functional enrichment for genes in Visium FFPE Human Normal Prostate Slide (0th Order Image Features) with the lowest significant p-values.....33

LIST OF FIGURES

Figure 1: Workflow of the method from tissue examination to identifying gene ontology terms are associated to ITF 3
Figure 2: Pheatmap for the correlation matrix of Parent Visium Human Breast Cancer slide (1st Order Image Features) 7
Figure 3: Pheatmap for the correlation matrix of Parent Visium Human Breast Cancer slide (0th Order Image Features) 9
Figure 4: Pheatmap for the correlation matrix of Parent Human Breast Cancer slide (1st Order Image Features) 11
Figure 5: Pheatmap for the correlation matrix of Parent Human Breast Cancer slide (0th Order Image Features) 13
Figure 6: Pheatmap for the correlation matrix of FFPE Human Breast Cancer slide (1st Order Image Features) 15
Figure 7: Pheatmap for the correlation matrix of FFPE Human Breast Cancer slide (0th Order Image Features) 17
Figure 8: Pheatmap for the correlation matrix of Human Prostate Cancer slide (1st Order Image Features) 20
Figure 9: Pheatmap for the correlation matrix of Human Prostate Cancer slide (0th Order Image Features) 23
Figure 10: Pheatmap for the correlation matrix of Prostate Acinar Cell Carcinoma slide (1st Order Image Features) 25
Figure 11: Pheatmap for the correlation matrix of Prostate Acinar Cell Carcinoma Slide (0th Order Image Features) 27
Figure 12: Pheatmap for the correlation matrix of Visium FFPE Human Normal Prostate slide (1st Order Image Features) 30
Figure 13: Pheatmap for the correlation matrix of Visium FFPE Human Normal Prostate slide (0th Order Image Features) 32

Chapter One: Background

In the United States, cancer is one of the primary issues of concern in public health because it affects a high number of individuals and based on these efforts, cancer prognosis continues to improve (National Cancer Institute, 2020). According to the World Health Organization (WHO) (2022), cancer is a major cause of death across the globe. Breast cancer and prostate cancer are two of the most widespread cancer types affecting people (World Health Organization, 2022). In 2020, breast cancer affected more than 2.25 million individuals worldwide and led to the death of more than 680,000 women (World Health Organization, 2022). Additionally, WHO (2022) reported that prostate cancer affected more than 1.4 million men worldwide in 2020.

In the biology of cancer, the tumor microenvironment (TME) is the group that consists of tissues, different cell types, and the extracellular matrix (ECM), which are origins of cancer tumors (Henke et al., 2020). Henke et al. (2020) state that a mass of tumors usually contains elevated levels of ECM. Using the levels of ECM as a tumor indicator could yield many benefits in detecting cancer in early stages and helping in treatment as well (Brassart-Pasco et al., 2020). Realistically, since ECM has a function in TME as a treatment resistor, attempting to eliminate the ECM could increase cancer treatment effectiveness (Henke et al., 2020). In order to examine cell segments for the characterization of pathological proteins and gene terms related to tumors, we could use biopsy imaging as a diagnosis method to inspect tumor tissues (Versaggi & De Leucio, 2022). One way to study solid tumors, like breast and prostate cancer, is using imaging which can be better understood using image analysis techniques including but not limited to Topological Data Analysis (TDA).

TDA is a useful process that investigates the dimensional structure of biological or biomedical data to analyze the data shapes and uses the information to study resemblance and differentiation in a particular data more deeply (Loughrey et al., 2021). One of the approaches of TDA is to create a link between molecular science and mathematical basis for an application to genetics related to oncology (Loughrey et al., 2021). The advance of bioinformatics methods could be promising for cancer diagnosis and treatments in the oncology field (Singer et al., 2017).

In our data, we are interested in using gene correlation analysis, which focuses on revealing which genes have an association with biological functions and could be considered a way to detect links between genes and abnormal functions that lead to diseases (Liu et al., 2020). Determining the association between gene expression and image features in this paper is done by generating correlation matrices and visualizing these correlations using heatmaps. Furthermore, we will be applying methods that allows us to find correlations between gene expression and ITFs in breast and prostate cancer. Genes that are correlated well with ITFs we call TAGs. Functional enrichment analysis is a useful technique for distinguishing protein and gene families related to certain functions or abnormalities in the body (Thanati et al., 2021) that we will apply to TAGs. This tool will allow us to identify statistically significant genetic correlations between gene expression and ST slides. The aim of this thesis is to find gene ontology terms that are significantly associated with ITFs from either breast cancer, prostate cancer, or both.

Chapter Two: Methods

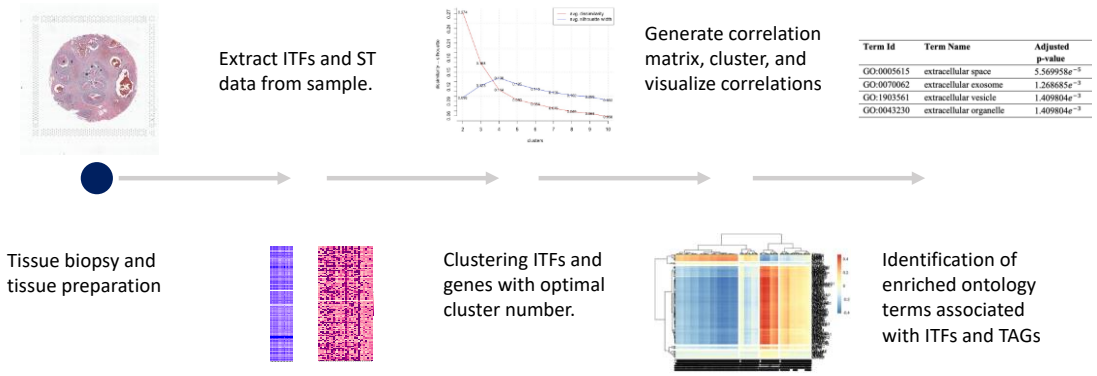

Figure 1: Workflow of the method from tissue examination to identifying gene ontology terms are associated to ITF.

Datasets

We downloaded 6 ST slides that depict breast cancer and prostate cancer from 10X genomics datasets resource. They contain gene expression information, spatial information, and corresponding tissue imaging. From these images, we extracted 1400 ITFs using TDA techniques (Abousamra et al., 2021; Aukerman et al., 2020). The slides 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. Breast cancer samples were taken from breast tissue of women older than 18 years old and the slides each contained more than 4,000 genes after filtering. The biopsy specimens for breast cancer tissues were preserved either by Formalin-Fixed Paraffin-Embedded (FFPE) or methanol fixation. Prostate cancer samples were taken from the prostate tissue of men, and the slides contained between 2543 and 4371 genes after filtering. The biopsy specimens for prostate cancer tissues

were all preserved by FFPE. The topology features for the slides were calculated from a 350×350-pixel patch of the image centered around the spot from which gene expression was calculated.

R software was used for all statistical analysis and functional enrichment was applied to find the association between the genes and the image features. According to Thanati et al. (2021), using functional enrichment can help determine which proteins and genes have relationships with cancer. Then, the correlation coefficient between image features and gene expression was calculated, as this is the fundamental method used to find the highest correlated genes to a specific ITF.

Calculating the correlation matrix starts with selecting the genes with more than 25% nonzero expression across all spots. The next step is selecting genes with top variance expression. This gives us a new gene set we can work with in the functional enrichment analysis. Next, we Log transform gene expression values and image features and calculate the correlation matrix between them.

When using the package Pheatmap (Kolde, 2018) we apply the pheatmap functions on the TAG-ITF correlation matrix to construct heatmaps in order to visualize the correlation between gene expression and image features. The color in the heatmap indicates the scale of the correlation coefficient.

Image Features

Image features were generated by a collaborating lab that specializes in computational pathology and TDA. They were extracted from image patches that corresponded to the same locations from which the gene expression was taken (Abousamra et al., 2021; Aukerman et al., 2020). Furthermore, two types of image

features were generated. The 0^{th}-order calculates persistent homology between pairs of points, whereas 1^{st}-order persistent homology searches for higher dimension patterns which include circular connections.

Clustering

We used the function pheatmap to apply ten clusters for rows (genes) and ten for columns (ITFs) as a primary number of clusters to determine the actual number of clusters in the following step. According to Sarkar (2019), k-means clustering, also known as the elbow method, is the commonly used method to find the optimal number of clusters. We used the elbow method to determine the optimal number of clusters in gene expression and image features, separately, using the correlation matrix object that we had created. From the elbow plot, we could decide from the bending elbow how many clusters we should use for genes and image features. Then, we applied the pheatmap again using the optimal number of clusters we obtained from the elbow method so we could visualize the new heatmap.

Functional Enrichment

The R package (gprofiler2) is a useful tool for performing functional analysis of genes (Kolberg et al., 2020). Gprofiler2 provides a list of gene terms for each gene set with their p-values. By using ‘gost’ function from Gprofiler2 package will apply the functional enrichment analysis on each gene ontology terms cluster. After adjusting p-values, we used a significance level of 0.05 to select the gene sets with the lowest p-values, which are the highest significant terms.

Chapter Three: Results

Parent Visium Human Breast Cancer Slide

1st Order Image Features

We selected genes with more than 25% nonzero expression and top variance expression via a quantile cutoff across all spots. We attempted to concentrate on the top 150 genes, which would have a wide range of values. We identified these genes using a quantile cutoff of 0.9959006. The correlation matrix generated from this slide has dimensions of 150×700. After applying the elbow method to determine the optimal number of clusters, we found that the optimal number of clusters for genes is six clusters and the optimal number of clusters for image features is seven clusters. The correlation coefficient for this correlation matrix heatmap is between -0.05 and 0.25. Furthermore, clear correlation differences can be seen in both the TAG and ITF clusters that can help to map TAG clusters to specific ITF clusters (Figure 2).

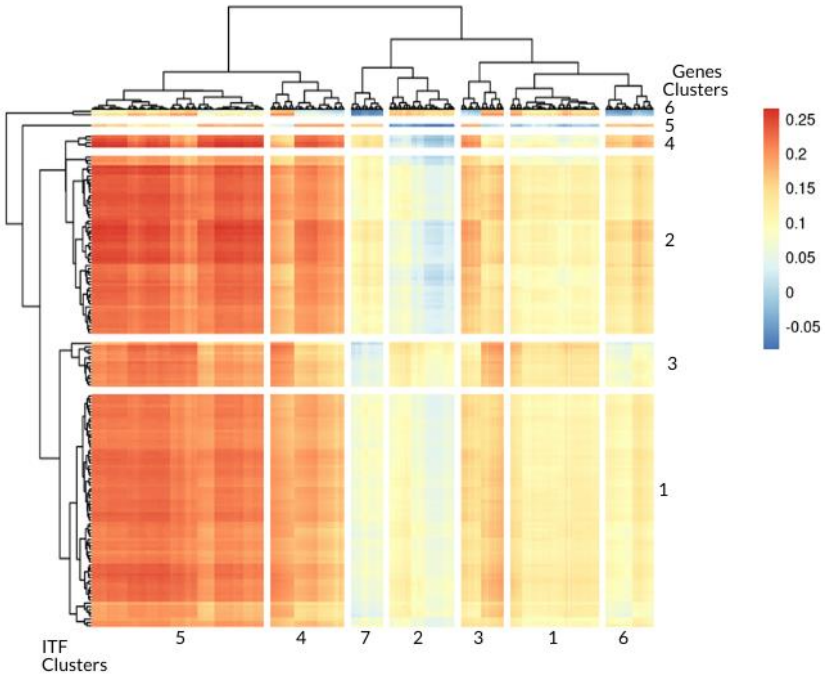

Figure 2: Pheatmap for the correlation matrix of Parent Visium Human Breast Cancer slide (1st Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 3).

Table 1: Functional enrichment for genes in Parent Visium Human Breast Cancer Slide (1st Order Image Features) with the lowest significant p-values.

Term Id	**Term Name**	**Adjusted p-value**	**Gene Cluster Number**
GO:0005615	extracellular space	$5.569958e^{-5}$	1
GO:0070062	extracellular exosome	$1.268685e^{-3}$	1
GO:1903561	extracellular vesicle	$1.409804e^{-3}$	1
GO:0043230	extracellular organelle	$1.409804ee^{-3}$	1
GO:0065010	Extracellular membrane-bounded organelle	$1.409804e^{-3}$	1
HP:0001427	Mitochondrial inheritance	$5.697496e^{-8}$	2
HP:0002572	Episodic vomiting	$3.811384e^{-7}$	2
HP:0004309	Ventricular preexcitation	$9.907111e^{-7}$	2
HP:0003200	Ragged-red muscle fibers	$1.110454e^{-6}$	2
HP:0000576	Centrocecal scotoma	$1.664157e^{-6}$	2
GO:0002162	Dystroglycan binding	0.006281256	3
GO:0005584	Collagen type I trimer	0 .0003033615	4
HP:0005623	Absent ossification of calvaria	0.0126888764	4
HP:0003321	Biconcave flattened vertebrae	0.0126888764	4
HP:0005005	Femoral bowing present at birth, straightening with time	0.0126888764	4
HP:0005897	Severe generalized osteoporosis	0.0126888764	4
CORUM:6822	ZAG-PIP complex	0.04994746	5

By measuring the similarity between genes and ITFs using correlation coefficient rank, we found that ITF clusters 1 and 2 are more correlated with gene cluster 6 (Figure 2, Table 1). ITF cluster 3 is the most similar to gene cluster 2 and ITF clusters 4 and 5 are the most similar to gene cluster 4 (Figure 2, Table 1). Lastly, ITF clusters 6 and 7 are the most similar to gene cluster 5 (Figure 2, Table 1).

0th Order Image Features

After applying the elbow method to determine the optimal number of clusters in the slide, we found that the optimal number of clusters for genes is six clusters and the optimal number of clusters for image features is four clusters. The correlation coefficient for this correlation matrix heatmap is between -0.2 and 0.2 (Figure 3).

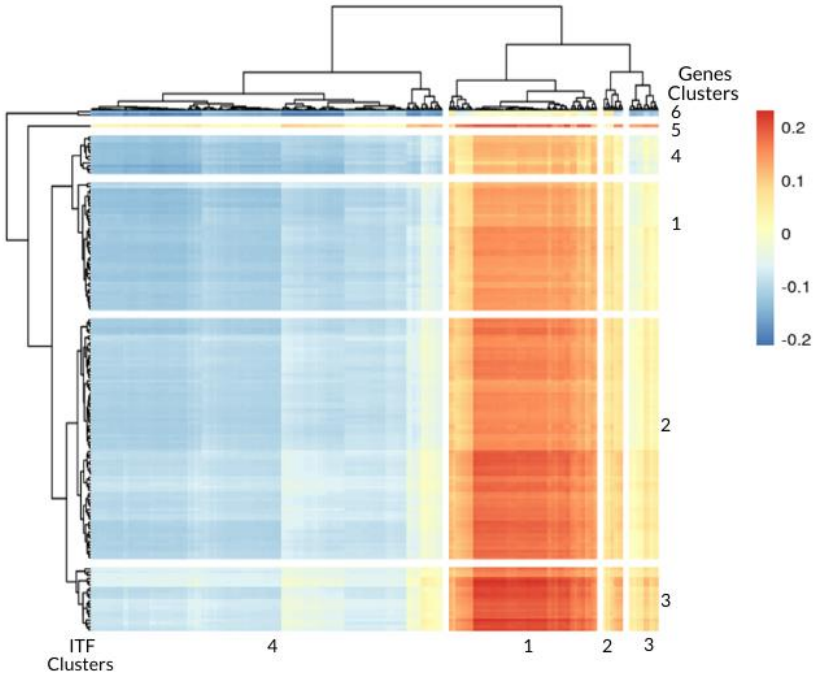

Figure 3: Pheatmap for the correlation matrix of Parent Visium Human Breast Cancer slide (0th Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 2).

Table 2: Functional enrichment for genes in Parent Visium Human Breast Cancer Slide (0^{th} Order Image Features) with the lowest significant p-values.

Term Id	Term Name	Adjusted p-value	Gene Cluster Number
GO:0005615	extracellular space	0.01631767	1
GO:0070062	extracellular exosome	$3.134382e^{-9}$	2
GO:1903561	extracellular vesicle	$3.774648e^{-9}$	2
GO:0065010	extracellular membrane-bounded organelle	$3.774648e^{-9}$	2
GO:0043230	extracellular organelle	$3.774648e^{-9}$	2
GO:0005615	extracellular space	$8.959959e^{-9}$	2
HP:0001427	Mitochondrial inheritance	$2.428127e^{-9}$	3
KEGG:05415	Diabetic cardiomyopathy	$1.142388e^{-8}$	3
HP:0004309	Ventricular preexcitation	$4.254401e^{-8}$	3
HP:0000576	Centrocecal scotoma	$1.239934e^{-7}$	3
HP:0002572	Episodic vomiting	$1.513199e^{-6}$	3
KEGG:05415	Diabetic cardiomyopathy	0.02883342	4
GO:0070069	cytochrome complex	0.03385070	4
KEGG:05012	Parkinson disease	0.03385070	4
KEGG:05010	Alzheimer disease	0.04709358	4
CORUM:6822	ZAG-PIP complex	0.04994746	5

By measuring the similarity between genes and ITFs using correlation coefficient rank, we found that ITF cluster 1 is the most similar to gene cluster 3 (Figure 3, Table 2). ITF clusters 2, 3, and 4 are the clusters more correlated with gene cluster 5 (Figure 3, Table 2).

Parent Human Breast Cancer Slide

1^{st} Order Image Features

We selected genes with more than 25% nonzero expression and top variance expression via a quantile cutoff across all spots. We attempted to concentrate on the top 150 genes, which would have a wide range of values. We identified these genes using a

quantile cutoff of 0.9959006. The correlation matrix generated from this slide has dimensions of 150×700. After applying the elbow method to determine the optimal number of clusters, we found that the optimal number of clusters for genes is five clusters and the optimal number of clusters for image features is five clusters as well. The correlation coefficient for this correlation matrix heatmap is between -0.05 and 0.25. Furthermore, clear correlation differences can be seen in both the TAG and ITF clusters that can help to map TAG clusters to specific ITF clusters (Figure 4).

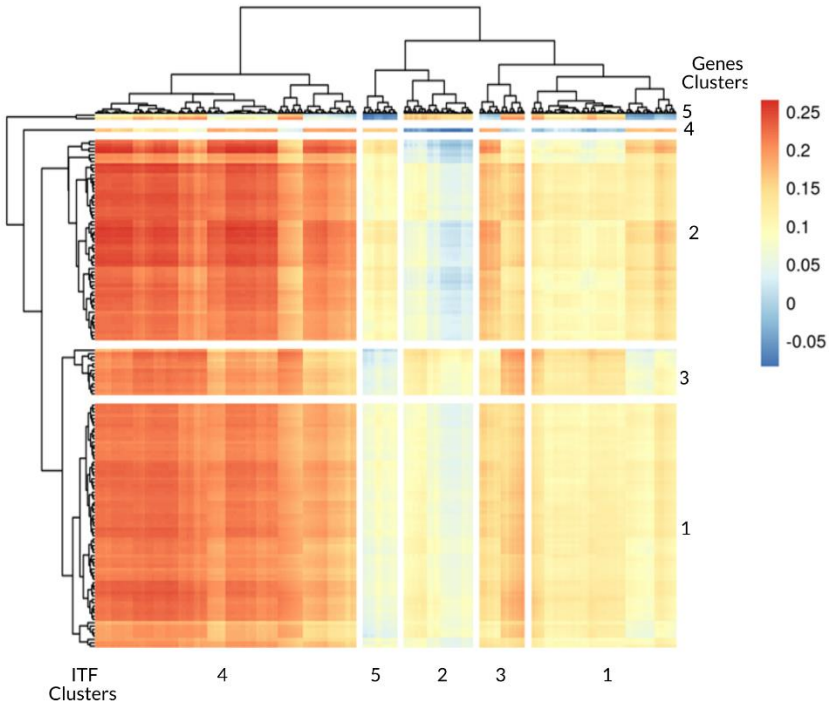

Figure 4: Pheatmap for the correlation matrix of Parent Human Breast Cancer slide (1st Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 3).

Table 3: Functional enrichment for genes in Parent Human Breast Cancer Slide (1st Order Image Features) with the lowest significant p-values.

Term Id	**Term Name**	**Adjusted p-value**	**Gene Cluster Number**
GO:0005615	extracellular space	$5.569958e^{-5}$	1
GO:0070062	extracellular exosome	$1.268685e^{-3}$	1
GO:1903561	extracellular vesicle	$1.409804e^{-3}$	1
GO:0043230	extracellular organelle	$1.409804e^{-3}$	1
GO:0065010	Extracellular membrane-bounded organelle	$1.409804e^{-3}$	1
KEGG:05415	Diabetic cardiomyopathy	$9.187864e^{-10}$	2
HP:0001427	Mitochondrial inheritance	$1.191038e^{-9}$	2
HP:0002572	Episodic vomiting	$2.432983e^{-8}$	2
HP:0000576	Centrocecal scotoma	$2.455963e^{-8}$	2
HP:0004309	Ventricular preexcitation	$3.506480e^{-8}$	2
GO:0002162	Dystroglycan binding	0.006281256	3
CORUM:6822	ZAG-PIP complex	00.04994746	4

By measuring of the similarity between genes and ITFs using the rank of the correlation coefficient, we found that ITF cluster 1 is the most similar to gene clusters 3. ITF cluster 2 is the most similar to gene cluster 5 and ITF clusters 4 and 5 are more correlated with gene cluster 2 (Figure 4, Table 3). Lastly, ITF cluster 5 is the most similar to gene clusters 4 (Figure 4, Table 3).

0^{th} Order Image Features

After applying the elbow method to determine the optimal number of clusters in the slide, we found that the optimal number of clusters for genes is six clusters and the optimal number of clusters for image features is five clusters. The correlation coefficient for this correlation matrix heatmap is between -0.05 and 0.25 (Figure 5).

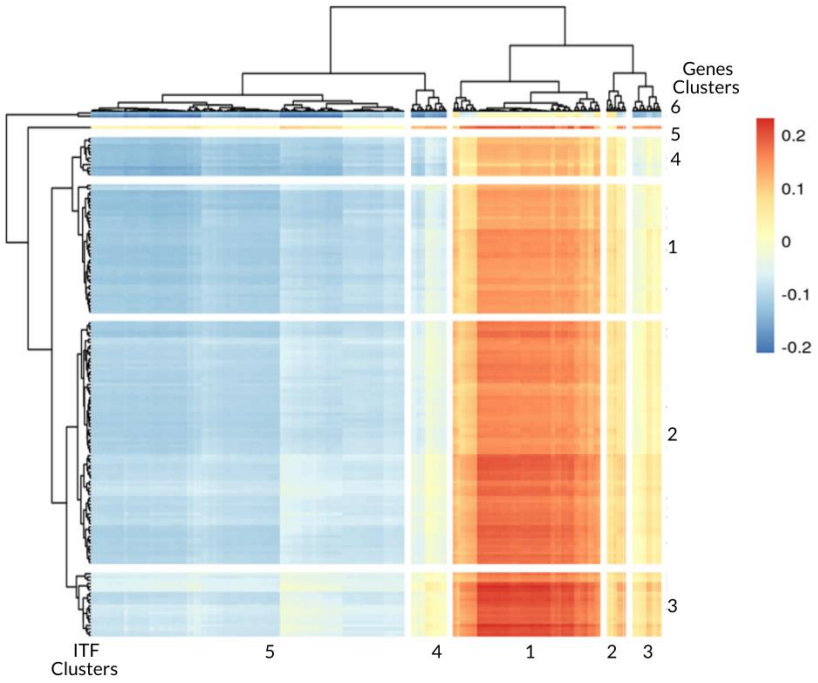

Figure 5: Pheatmap for the correlation matrix of Parent Human Breast Cancer slide (0^{th} Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 4).

Table 4: Functional enrichment for genes in Parent Human Breast Cancer Slide (0^{th} Order Image Features) with the lowest significant p-values.

Term Id	Term Name	Adjusted p-value	Gene Cluster Number
GO:0005615	extracellular space	$5.569958e^{-5}$	1
GO:0070062	extracellular exosome	$1.268685e^{-3}$	1
GO:1903561	extracellular vesicle	$1.409804e^{-3}$	1
GO:0043230	extracellular organelle	$1.409804e^{-3}$	1
GO:0065010	Extracellular membrane-bounded organelle	$1.409804e^{-3}$	1
HP:0001427	Mitochondrial inheritance	$5.697496e^{-8}$	2
HP:0002572	Episodic vomiting	$3.811384e^{-7}$	2
HP:0004309	Ventricular preexcitation	$9.907111e^{-7}$	2
HP:0003200	Ragged-red muscle fibers	$1.110454e^{-6}$	2
HP:0000576	Centrocecal scotoma	$1.664157e^{-6}$	2
GO:0002162	Dystroglycan binding	0.006281256	3
GO:0005584	collagen type I trimer	0.0003033615	4
HP:0005623	Absent ossification of calvaria	0.0126888764	4
HP:0003321	Biconcave flattened vertebrae	0.0126888764	4
HP:0005005	Femoral bowing present at birth, straightening with time	0.0126888764	4
HP:0005897	Severe generalized osteoporosis	0.0126888764	4
CORUM:6822	ZAG-PIP complex	0.04994746	5

By measuring the similarity between genes and ITFs using correlation coefficient rank, we found that ITF cluster 1 is the most similar cluster to gene cluster 3 and ITF cluster 2, 3,4, and 5 are the most correlated with gene cluster 5 (Figure 5, Table 4).

FFPE Human Breast Cancer Slide

1^{st} Order Image Features

We selected genes with more than 25% nonzero expression and top variance expression via a quantile cutoff across all spots. We attempted to concentrate on the top

150 genes, which would have a wide range of values. We identified these genes using a quantile cutoff of 0.9959006. The correlation matrix generated from this slide has the dimensions of 150×700. After applying the elbow method to determine the optimal number of clusters, we found that the optimal number of clusters for genes and image features are 5 clusters in both. The correlation coefficient for this correlation matrix heatmap is between -0.15 and 0.1. Furthermore, clear correlation differences can be seen in both the TAG and ITF clusters that can help to map TAG clusters to specific ITF clusters (Figure 6).

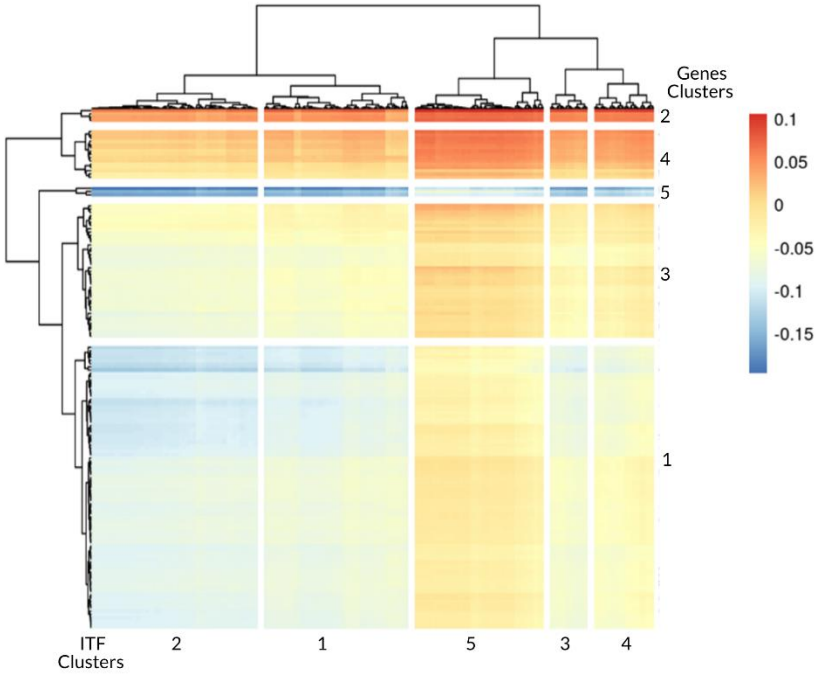

Figure 6: Pheatmap for the correlation matrix of FFPE Human Breast Cancer slide (1st Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 5).

Table 5: Functional enrichment for genes in FFPE Human Breast Cancer Slide (1st Order Image Features) with the lowest significant p-values.

Term Id	Term Name	Adjusted p-value	Gene Cluster Number
GO:0070062	extracellular exosome	$2.051973e^{-6}$	1
GO:1903561	extracellular vesicle	$2.362341e^{-6}$	1
GO:0065010	extracellular membrane-bounded organelle	$2.362341e^{-6}$	1
GO:0043230	extracellular organelle	$2.362341e^{-6}$	1
GO:0005615	extracellular space	$3.281412e^{-6}$	1
GO:0005615	extracellular space	$2.874349e^{-8}$	3
GO:0005576	extracellular region	$4.032521ee^{-6}$	3
GO:0070062	extracellular exosome	$7.746915e^{-6}$	3
GO:1903561	extracellular vesicle	$8.977896e^{-6}$	3
GO:0043230	extracellular organelle	$8.977896e^{-6}$	3
GO:0015453	oxidoreduction-driven active transmembrane transporter activity	$1.196315e^{-9}$	4
KEGG:05415	Diabetic cardiomyopathy	$3.163770e^{-9}$	4
GO:0098803	respiratory chain complex	$5.166699e^{-9}$	4
GO:0070469	respirasome	$1.073422e^{-8}$	4
HP:0002572	Episodic vomiting	$1.376587e^{-8}$	4

By measuring the similarity between genes and ITFs using correlation coefficient rank, we found that gene cluster 2 is the most similar cluster to all ITFs (Figure 6, Table 5).

0th Order Image Features

After applying the elbow method to determine the optimal number of clusters in the slide, we have found that the optimal number of clusters for both genes and image features are five clusters for both of them. The correlation coefficient for this correlation matrix heatmap is between -0.15 and 0.1 (Figure 7).

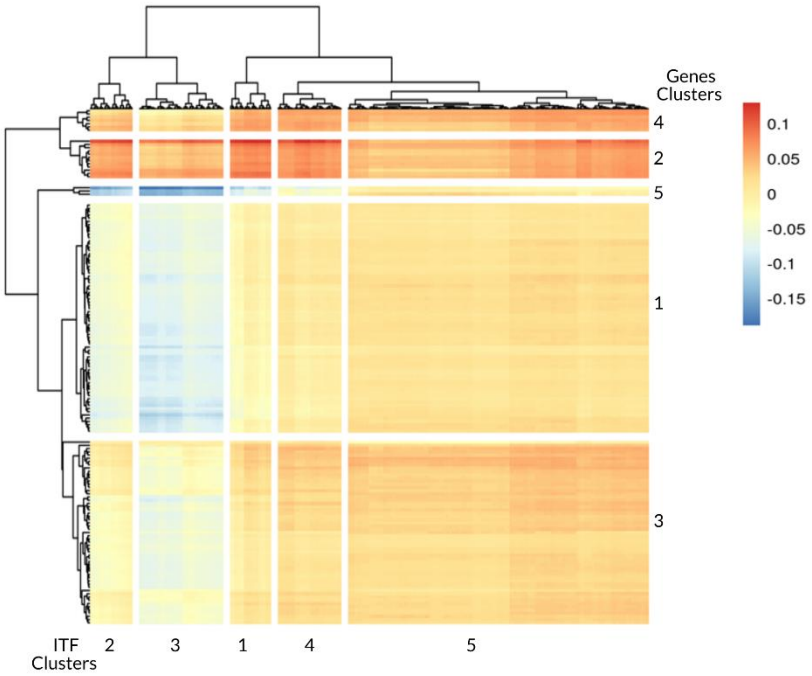

Figure 7: Pheatmap for the correlation matrix of FFPE Human Breast Cancer slide (0^{th} Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 6).

Table 6: Functional enrichment for genes in FFPE Human Breast Cancer Slide (0^{th} Order Image Features) with the lowest significant p-values.

Term Id	**Term Name**	**Adjusted p-value**	**Gene Cluster Number**
GO:0005615	extracellular space	$2.175512e^{-5}$	1
GO:0070062	extracellular exosome	$7.189842e^{-5}$	1
GO:1903561	extracellular vesicle	$7.519775e^{-5}$	1
GO:0043230	extracellular organelle	$7.519775e^{-5}$	1
GO:0065010	extracellular membrane-bounded organelle	$7.519775e^{-5}$	1
HP:0000576	Centrocecal scotoma	$2.634312e^{-11}$	2
HP:0001427	Mitochondrial inheritance	$1.891416e^{-10}$	2
HP:0200125	Mitochondrial respiratory chain defects	$5.897102e^{-10}$	2
HP:0004309	Ventricular preexcitation	$2.038573e^{-9}$	2
HP:0007763	Retinal telangiectasia	$3.471255e^{-9}$	2
GO:0005615	extracellular space	$2.918838e^{-8}$	3
GO:0005576	extracellular region	$6.894013e^{-7}$	3
GO:0070062	extracellular exosome	$8.051882e^{-7}$	3
GO:1903561	extracellular vesicle	$9.590950e^{-7}$	3
GO:0065010	extracellular membrane-bounded organelle	$9.601088e^{-7}$	3
KEGG:05415	Diabetic cardiomyopathy	0.0000012396	4
GO:0005584	collagen type I trimer	0.0014950736	4
GO:0015453	oxidoreduction-driven active transmembrane transporter activity	0.0222768962	4
CORUM:2886	Respiratory chain complex I (incomplete intermediate ND1, ND2, ND3, CIA30 assembly), mitochondrial	0.0261210401	4
GO:0098803	respiratory chain complex	0.0429815125	4

By measuring the similarity between genes and ITFs using correlation coefficient rank, we found that gene cluster 2 is the most similar cluster to all ITF clusters (Figure 7, Table 6).

Human Prostate Cancer Slide

1st Order Image Features

We selected genes with more than 25% nonzero expression and top variance expression via a quantile cutoff across all spots. We attempted to concentrate on the top 150 genes, which would have a wide range of values. We identified these genes using a quantile cutoff of 0.9916393. The correlation matrix generated from this slide has dimensions of 150×700. After applying the elbow method to determine the optimal number of clusters, we found that the optimal number of clusters for genes is five clusters and for image features is four clusters. The correlation coefficient for this correlation matrix heatmap is between -0.4 and 0.4. Furthermore, clear correlation differences can be seen in both the TAG and ITF clusters that can help to map TAG clusters to specific ITF clusters (Figure 8).

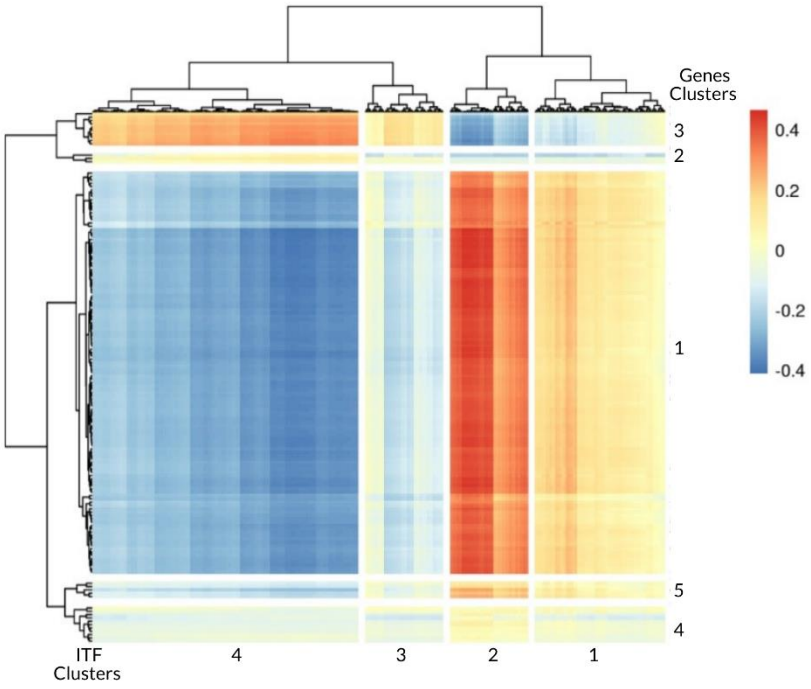

Figure 8: Pheatmap for the correlation matrix of Human Prostate Cancer slide (1^{st} Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 7).

Table 7: Functional enrichment for genes in Human Prostate Cancer Slide (1st Order Image Features) with the lowest significant p-values.

Term Id	**Term Name**	**Adjusted p-value**	**Gene Cluster Number**
GO:0070062	extracellular exosome	$8.522469e^{-21}$	1
GO:1903561	extracellular vesicle	$1.427739e^{-20}$	1
GO:0065010	extracellular membrane-bounded organelle	$1.457071e^{-20}$	1
GO:0043230	extracellular organelle	$1.457071e^{-20}$	1
GO:0031982	vesicle	$1.842410e^{-20}$	1
MIRNA:hsa-miR-489-5p	hsa-miR-489-5p	0.01172711	2
GO:0010595	positive regulation of endothelial cell migration	0.03729422	2
REAC:R-HSA-445355	Smooth Muscle Contraction	$6.325223e^{-10}$	3
GO:0006936	muscle contraction	$1.717889ee^{-8}$	3
REAC:R-HSA-397014	Muscle contraction	$7.828103e^{-8}$	3
GO:0003012	muscle system process	$9.496920e^{-8}$	3
HPA:0490693	smooth muscle; smooth muscle cells[High]	$3.967884e^{-7}$	3
GO:0005584	collagen type I trimer	0.001492507	4
HP:0005623	Absent ossification of calvaria	0.003113332	4
HP:0005897	Severe generalized osteoporosis	0.003113332	4
HP:0003321	Biconcave flattened vertebrae	0.003113332	4
HP:0005005	Femoral bowing present at birth, straightening with time	0.003113332	4
TF:M07618_1	Factor: SRF; motif: CCTTWTATGGNN; match class: 1	0.007103503	5
TF:M01304_1	Factor: SRF; motif: NNCCAWAWAAGGV; match class: 1	0.007103503	5
TF:M12669_1	Factor: SRF; motif: TTNCCTTATWTGGNC; match class: 1	0.016334781	5

GO:0050998	nitric-oxide synthase binding	0.016997056	5
TF:M01007_1	Factor: SRF; motif: CNKNKCCTTATWTGGNNNN; match class: 1	0.044725161	5

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we found that ITF cluster 1 and 2 is the most similar cluster to gene cluster 1 (Figure 8, Table 7). ITF clusters 3 and 4 are more correlated with the cluster 3 of genes (Figure 8, Table 7).

0^{th} Order Image Features

After applying the elbow method to determine the optimal number of clusters in the slide, we found that the optimal number of clusters for genes and for image features clusters is four for both of them. The correlation coefficient for this correlation matrix heatmap is between

-0.4 between 0.4 (Figure 9).

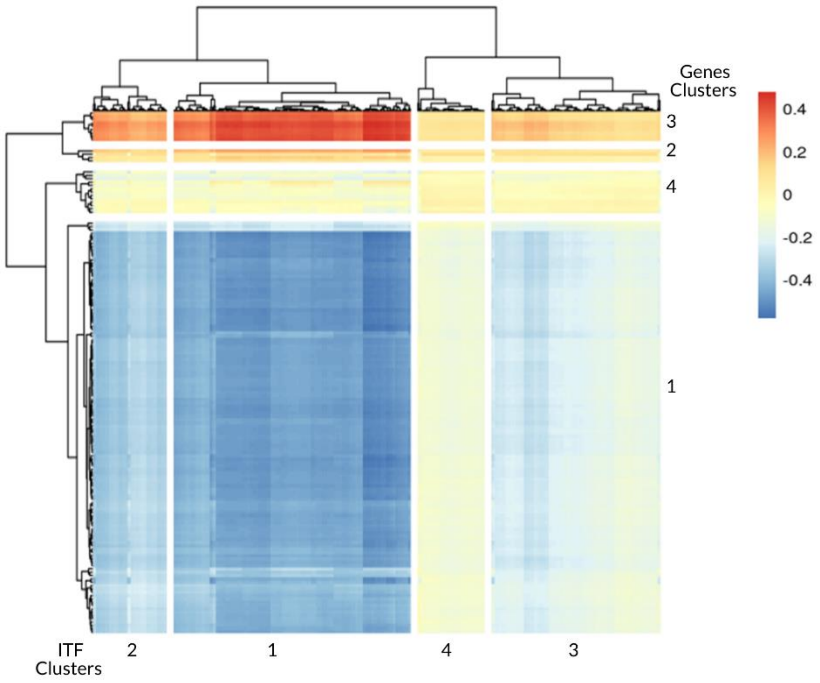

Figure 9: Pheatmap for the correlation matrix of Human Prostate Cancer slide (0^{th} Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 8).

Table 8: Functional enrichment for genes in Human Prostate Cancer Slide (0^{th} Order Image Features) with the lowest significant p-values.

Term Id	**Term Name**	**Adjusted p-value**	**Gene Cluster Number**
GO:0070062	extracellular exosome	$6.479362e^{-22}$	1
GO:1903561	extracellular vesicle	$1.107932e^{-21}$	1
GO:0065010	extracellular membrane-bounded organelle	$1.132104e^{-21}$	1
GO:0043230	extracellular organelle	$1.132104e^{-21}$	1
GO:0031982	vesicle	$3.814630e^{-17}$	1
GO:0034599	cellular response to oxidative stress	0.01761702	2
GO:0062197	cellular response to chemical stress	0.01927010	2
GO:0006979	response to oxidative stress	0.02181185	2
REAC:R-HSA-445355	Smooth Muscle Contraction	$1.115693e^{-7}$	3
GO:0006936	muscle contraction	$6.454091e^{-7}$	3
REAC:R-HSA-397014	Muscle contraction	$2.791592e^{-6}$	3
GO:0003012	muscle system process	$2.840664e^{-6}$	3
HPA:0490693	smooth muscle; smooth muscle cells[High]	$4.850593e^{-6}$	3
GO:0005584	collagen type I trimer	0.001540888	4
HP:0003321	Biconcave flattened vertebrae	0.011601258	4
HP:0005005	Femoral bowing present at birth, straightening with time	0.011601258	4
HP:0005623	Absent ossification of calvaria	0.011601258	4
HP:0005897	Severe generalized osteoporosis	0.011601258	4

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we found all four ITF clusters is the most similar to gene cluster number 3 (Figure 9, Table 8).

Prostate Acinar Cell Carcinoma Slide

1st Order Image Features

We selected genes with more than 25% nonzero expression and top variance expression via a quantile cutoff across all spots. We attempted to concentrate on the top 150 genes, which would have a wide range of values. We identified these genes using a quantile cutoff of 0.9916393. The correlation matrix generated from this slide has dimensions of 150×700. After applying the elbow method to determine the optimal number of clusters, we found that the optimal number of clusters is five clusters for genes and image features is five clusters, too. The correlation coefficient for this correlation matrix heatmap is between -0.15 and 0.15. Furthermore, clear correlation differences can be seen in both the TAG and ITF clusters that can help to map TAG clusters to specific ITF clusters (Figure 10).

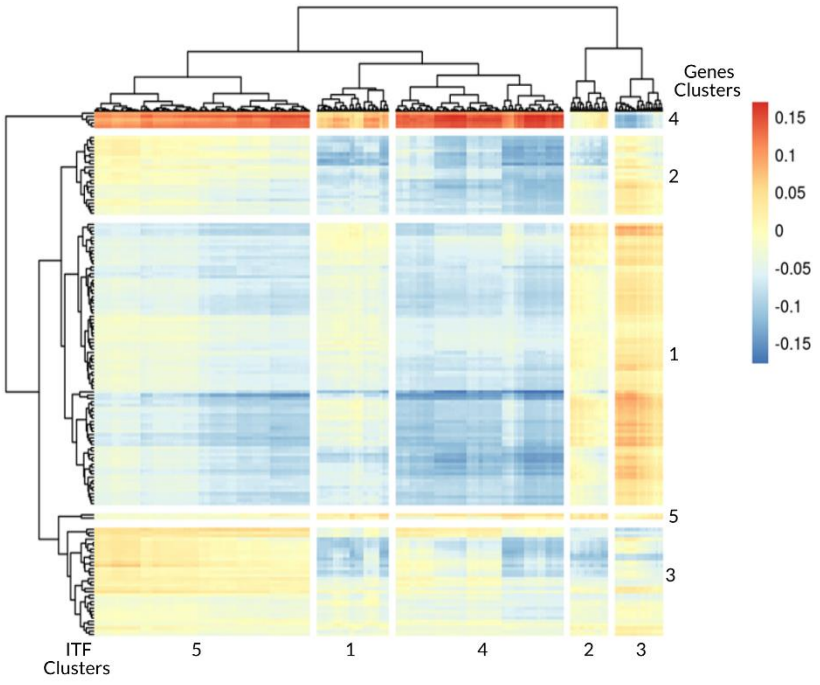

Figure 10: Pheatmap for the correlation matrix of Prostate Acinar Cell Carcinoma slide (1st Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 9).

Table 9: Functional enrichment for genes in Prostate Acinar Cell Carcinoma Slide (1st Order Image Features) with the lowest significant p-values.

Term Id	Term Name	Adjusted p-value	Gene Cluster Number
GO:0070062	extracellular exosome	$1.093617e^{-17}$	1
GO:1903561	extracellular vesicle	$1.652514e^{-17}$	1
GO:0065010	extracellular membrane-bounded organelle	$1.678243e^{-17}$	1
GO:0043230	extracellular organelle	$1.678243e^{-17}$	1
GO:0031982	vesicle	$2.652779e^{-15}$	2
GO:0005615	extracellular space	0.0002065117	2
GO:0005576	extracellular region	0.0005853319	2
HPA:0471443	skin 2; fibrohistiocytic cells[High]	0.0134575912	2
GO:0070062	extracellular exosome	0.0139763283	2
GO:1903561	extracellular vesicle	0.0153151399	2
GO:0071751	secretory IgA immunoglobulin complex	0.0001962954	3
GO:0071749	polymeric IgA immunoglobulin complex	0.0001962954	3
GO:0071746	IgA immunoglobulin complex, circulating	0.0001962954	3
GO:0071745	IgA immunoglobulin complex	0.0001962954	3
GO:0070062	extracellular exosome	0.0004214256	3
REAC:R-HSA-445355	Smooth Muscle Contraction	0.001364539	4
HPA:0401322	rectum; enterocytes - Microvilli[≥Medium]	0.001378253	4
HPA:0401321	rectum; enterocytes - Microvilli[≥Low]	0.002072836	4
HP:0002579	Gastrointestinal dysmotility	0.002825241	4
HP:0030895	Abnormal gastrointestinal motility	0.002913245	4

GO:0005520	insulin-like growth factor binding	0.008985672	5

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we found ITF cluster 3 is the most similar to gene cluster 1 and ITF clusters 1, 4 and 5 are the most similar to genes cluster 4 (Figure 10, Table 9). ITF cluster 2 is the most similar to genes cluster 5 (Figure 10, Table 9).

0^{th} Order Image Features

After applying the elbow method to determine the optimal number of clusters in the slide, we found that the optimal number of clusters for the genes is five and for image features clusters is five as well. The correlation coefficient for this correlation matrix heatmap is between -0.1 and 0.2 (Figure 11).

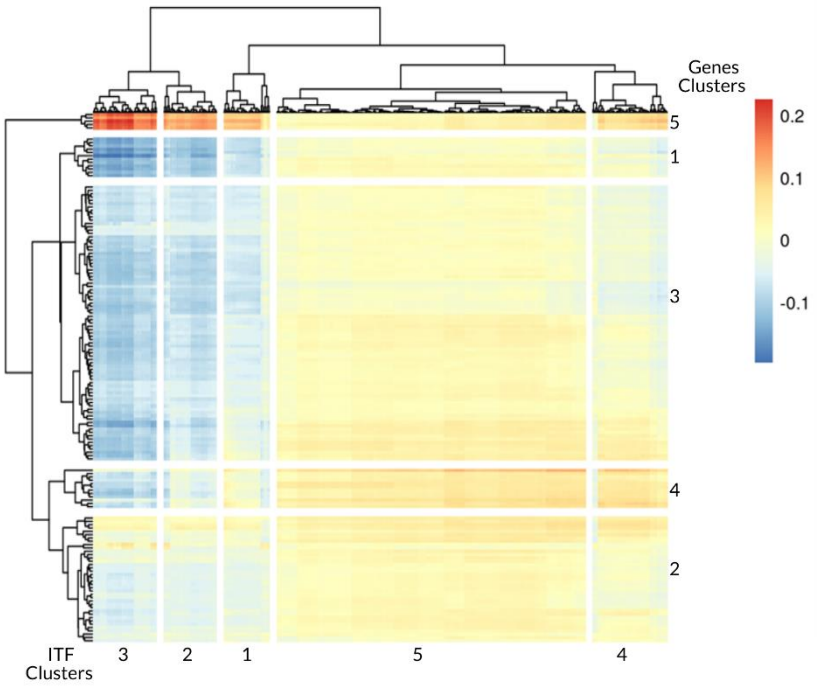

Figure 11: Pheatmap for the correlation matrix of Prostate Acinar Cell Carcinoma Slide (0^{th} Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 10).

Table 10: Functional enrichment for genes in Prostate Acinar Cell Carcinoma Slide (0^{th} Order Image Features) with the lowest significant p-values.

Term Id	Term Name	Adjusted p-value	Gene Cluster Number
GO:0044754	Autolysosome	0.0002318339	1
GO:0000323	lytic vacuole	0.0003481053	1
GO:0005764	Lysosome	0.0003481053	1
GO:0005773	Vacuole	0.0008109995	1
GO:0005767	secondary lysosome	0.0009158836	1
GO:0031982	Vesicle	$6.295632e^{-5}$	2
GO:0070062	extracellular exosome	$2.288019e^{-4}$	2
GO:1903561	extracellular vesicle	$2.636669e^{-4}$	2
GO:0043230	extracellular organelle	$2.636669e^{-4}$	2
GO:0065010	extracellular membrane-bounded organelle	$2.636669e^{-4}$	2
GO:0070062	extracellular exosome	$1.861833e^{-13}$	3
GO:1903561	extracellular vesicle	$2.639846e^{-13}$	3
GO:0065010	extracellular membrane-bounded organelle	$2.672183e^{-13}$	3
GO:0043230	extracellular organelle	$2.672183e^{-13}$	3
GO:0005615	extracellular space	$2.010480e^{-12}$	3
GO:0071745	IgA immunoglobulin complex	$4.459986e^{-6}$	4
GO:0071751	secretory IgA immunoglobulin complex	$4.459986e^{-6}$	4
GO:0071749	polymeric IgA immunoglobulin complex	$4.459986e^{-6}$	4
GO:0071746	IgA immunoglobulin complex, circulating	$4.459986e^{-6}$	4
GO:0005615	extracellular space	$1.141363e^{-3}$	4
REAC:R-HSA-445355	Smooth Muscle Contraction	0.001364539	5
HPA:0401322	rectum; enterocytes - Microvilli[≥Medium]	0.001378253	5

HPA:0401321	rectum; enterocytes - Microvilli[≥Low]	0.002072836	5
HP:0002579	Gastrointestinal dysmotility	0.002825241	5
HP:0030895	Abnormal gastrointestinal motility	0.002913245	5

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we found that ITF cluster 1, 2, 3, and 4 is the most similar cluster to gene cluster 5 (Figure 11, Table 10). ITF clusters 5 is more correlated with gene cluster 4 (Figure 11, Table 10).

Visium FFPE Human Normal Prostate Slide

1st Order Image Features

We selected genes with more than 25% nonzero expression and top variance expression via a quantile cutoff across all spots. We attempted to concentrate on the top genes, which would have a wide range of values. We identified these genes using a quantile cutoff of 0.9916393. The correlation matrix generated from this slide has dimensions of 145×700. After applying the elbow method to determine the optimal number of clusters, we found that the optimal number of clusters is five clusters for genes and image features is four clusters. The correlation coefficient for this correlation matrix heatmap is between -0.2 and 0.4. Furthermore, clear correlation differences can be seen in both the TAG and ITF clusters that can help to map TAG clusters to specific ITF clusters (Figure 12).

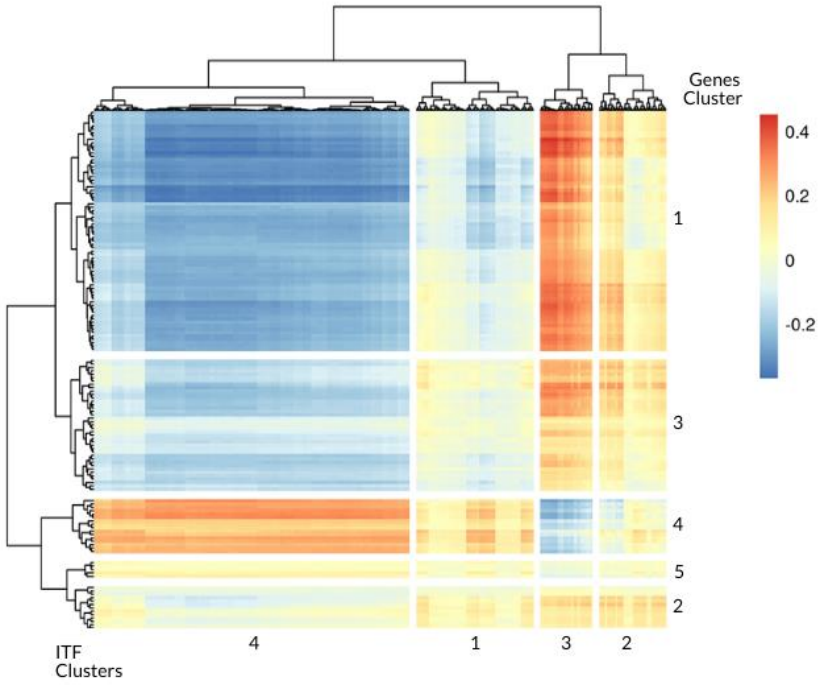

Figure 12: Pheatmap for the correlation matrix of Visium FFPE Human Normal Prostate slide (1st Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 11).

Table 11: Functional enrichment for genes in Visium FFPE Human Normal Prostate Slide (1st Order Image Features) with the lowest significant p-values.

Term Id	**Term Name**	**Adjusted p-value**	**Gene Cluster Number**
GO:0070062	extracellular exosome	$1.603815e^{-13}$	1
GO:1903561	extracellular vesicle	$2.219204e^{-13}$	1
GO:0043230	extracellular organelle	$2.240562e^{-13}$	1
GO:0065010	extracellular membrane-bounded organelle	$2.240562e^{-13}$	1
GO:0005615	extracellular space	$8.853403e^{-10}$	1
GO:0062023	collagen-containing extracellular matrix	$6.700992e^{-5}$	2
GO:0031012	extracellular matrix	$3.520194e^{-4}$	2
GO:0030312	external encapsulating structure	$3.520194e^{-4}$	2
CORUM:2254	CTGF/Hcs24-actin(beta/gamma) complex	$2.231109e^{-3}$	2
GO:0005615	extracellular space	$5.403000e^{-3}$	2
GO:0043230	extracellular organelle	$9.806696e^{-8}$	3
GO:0005576	extracellular region	$3.037689e^{-6}$	3
GO:0070887	cellular response to chemical stimulus	$8.624522e^{-4}$	3
GO:0010033	response to organic substance	$8.624522e^{-4}$	3
GO:0071310	cellular response to organic substance	$8.624522e^{-4}$	3
REAC:R-HSA-445355	Smooth Muscle Contraction	$1.111645e^{-14}$	4
HPA:0490693	smooth muscle; smooth muscle cells[High]	$3.736946e^{-12}$	4
GO:0006936	muscle contraction	$8.389078e^{-12}$	4
REAC:R-HSA-397014	Muscle contraction	$3.763786e^{-11}$	4
GO:0015629	actin cytoskeleton	$4.049754e^{-11}$	4

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we found that ITF cluster 1 and 4 is the most similar cluster to gene cluster 4 (Figure 12, Table 11). ITF cluster 2 is more correlated with gene cluster 2 and ITF cluster 3 is more correlated with gene cluster 1 (Figure 12, Table 11).

0th Order Image Features

After applying the elbow method to determine the optimal number of clusters in the slide, we found that the optimal number of clusters for the genes is five and for image features clusters is six. The correlation coefficient for this correlation matrix heatmap is between -0.4 and 0.4 (Figure 13).

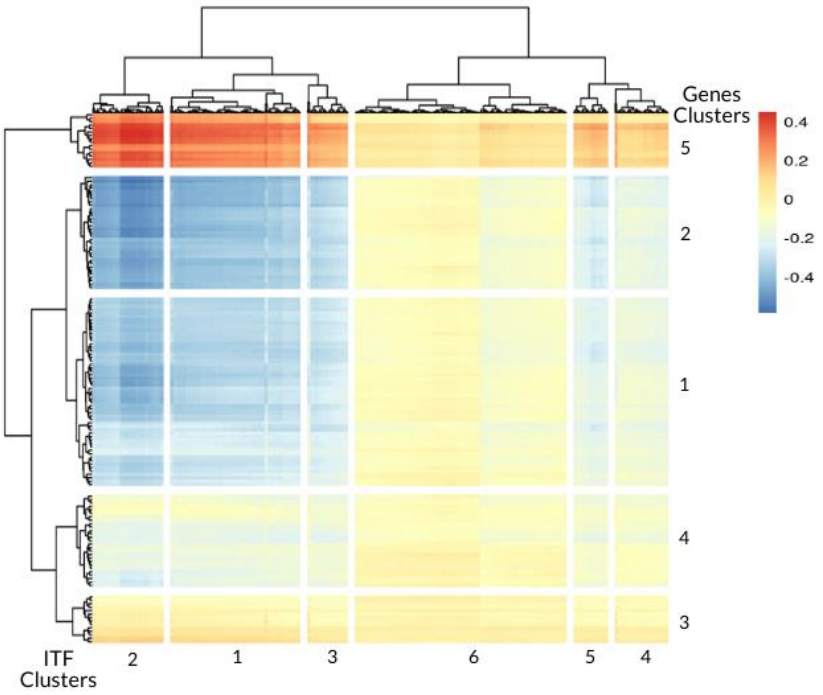

Figure 13: Pheatmap for the correlation matrix of Visium FFPE Human Normal Prostate slide (0th Order Image Features).

After performing functional enrichment for each gene cluster, we selected the top 5 gene ontology terms for each cluster based on lowest p-value (Table 12).

Table 12: Functional enrichment for genes in Visium FFPE Human Normal Prostate Slide (0^{th} Order Image Features) with the lowest significant p-values.

Term Id	Term Name	Adjusted p-value	Gene Cluster Number
GO:0070062	extracellular exosome	$6.127363e^{-10}$	1
GO:1903561	extracellular vesicle	$7.797512e^{-10}$	1
GO:0065010	extracellular membrane-bounded organelle	$7.797512e^{-10}$	1
GO:0043230	extracellular organelle	$7.797512e^{-10}$	1
GO:0005615	extracellular space	$1.210568e^{-7}$	1
GO:0070062	extracellular exosome	$1.870788e^{-6}$	2
GO:1903561	extracellular vesicle	$2.204394e^{-6}$	2
GO:0065010	extracellular membrane-bounded organelle	$2.204394e^{-6}$	2
GO:0043230	extracellular organelle	$2.204394e^{-6}$	2
GO:0005615	extracellular space	$4.329608e^{-4}$	2
GO:0062023	collagen-containing extracellular matrix	0.001950355	3
GO:0031012	extracellular matrix	0.007442751	3
GO:0030312	external encapsulating structure	0.007442751	3
TF:M07618_1	Factor: SRF; motif: CCTTWTATGGNN; match class: 1	0.043463755	3
TF:M01304_1	Factor: SRF; motif: NNCCAWAWAAGGV; match class: 1	0.043463755	3
GO:0005615	extracellular space	$3.635334e^{-8}$	4
GO:0005576	extracellular region	$4.175812e^{-6}$	4
GO:0070062	extracellular exosome	$2.636948e^{-4}$	4
GO:1903561	extracellular vesicle	$2.993143e^{-4}$	4
GO:0065010	extracellular membrane-bounded organelle	$2.993143e^{-4}$	4
REAC:R-HSA-445355	Smooth Muscle Contraction	$1.111645e^{-14}$	5
HPA:0490693	smooth muscle; smooth muscle cells[High]	$3.736946e^{-12}$	5

GO:0006936	muscle contraction	$8.389078e^{-12}$	5
REAC:R-HSA-397014	Muscle contraction	$3.763786e^{-11}$	5
GO:0015629	actin cytoskeleton	$4.049754e^{-11}$	5

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we found that gene cluster 5 is the most cluster correlated with all ITF clusters (Figure 13, Table 12).

Chapter Four: Discussion

From the results presented above, we can conclude that all ST slides have TAGs that represent significantly enriched ontology terms. Five extracellular gene terms are the most common terms in those slides: extracellular space, extracellular exosome, extracellular vesicle, extracellular organelle and extracellular membrane-bounded organelle. Both breast cancer slides, and prostate cancer slides contain these genes terms as these gene terms appear to have the most significant p-values among genes in functional enrichment.

By measuring of the similarity between genes and ITFs using the correlation coefficient rank, we have noticed that in the ST slides (1st Order Image Features): Human Prostate Cancer, Prostate Acinar Cell Carcinoma and Visium FFPE Human Normal Prostate that the ECM elements are existing in gene clusters that have the most similarity with specific ITFs. ECM in Human Prostate Cancer ST slide is in gene cluster 1, which is the most similar to ITF clusters 1 and 2. ITF cluster 1 contains ITF1-ITF97 and ITF197-264, and ITF cluster 2 contains ITF98-ITF196. In Prostate Acinar Cell Carcinoma, ECM elements are in gene clusters 1 and 2, but only gene cluster 1 has the most similarity with ITF cluster 3, which contains ITF98-ITF159. ECM in Visium FFPE Human Normal Prostate ST slide is in gene clusters 1, 2, and 3, but the similarity with ITF is only in gene cluster 1, which is the most similar to ITF cluster 3 and gene cluster 2, which is the most similar to ITF cluster 2. ITF cluster 2 contains ITF81-ITF231and ITF cluster 3 contains ITF104-ITF169. Clearly, there is an overlap in ITFs among these ST slides. By distinguishing the overlapping in ITFs between these ST slides, we can see that ITF81-

ITF231 are regularly appeared to share the similarity with ECM in those prostate ST slides.

According to Frantz et al. (2010), ECM elements exist in tissues and body organs and are identified as non-cellular parts. The matrix plays a significant role in tissues and organs as it helps in cell balancing, cell differentiating and many other physical and biochemical functions (Frantz et al., 2010). Our understanding of cancer is that it is unreasonable growth of cells that can evade immune system response and defense (National Cancer Institute, 2021). This ECM could also enhance cancer growth by changing cell functions, so they become abnormal and make the tissues a more fertile environment for cancer cells (Nallanthighal et al., 2019). In essence, developing breast cancer cells could begin as cells with abnormal homeostasis due to an alteration caused by the ECM (Zhao et al., 2021). Furthermore, extracellular vesicle and extracellular exosomes help in cell communication, and they contribute to the spread of cancer cells in the prostate, in particular (Vlaeminck-Guillem, 2018).

Another significant gene term is shared between the breast cancer slides and prostate cancer slides: Collagen Type I Trimer. Collagen plays an important part in cancer progression, especially when it contributes to the activities of the ECM in developing cancer cells via biological signals and enhancing tumors growth (Xu et al., 2019). Even though collagen is considered cancer-promoting, some studies have shown that collagen could help in cancer treatment because a patient could benefit from its resistance ability, especially in the case of chemotherapy treatment (Xu et al., 2019).

In the slides Parent Visium Human Breast Cancer and Parent Human Breast Cancer, we found the ZAG-PIP Complex gene term. According to Urbaniak et al. (2018),

ZAG-PIP is a complex of Zinc alpha2-glycoprotein (ZAG) with Prolactin inducible protein (PIP), and it appears that ZAG-PIP should be at very low levels in normal breast cells and high levels in breast cancer cells. However, ZAG-PIP can also be at lower levels when breast cancer is in a late-stage (Urbaniak et al., 2018).

Another expression with a significant p-value in the Prostate Acinar Cell Carcinoma slide is IgA immunoglobulin complex. Prostates excrete IgA immunoglobulin and when a change in the component of IgA could lead to a prospect of abnormality of prostate immunity behavior (Silva et al., 2017). In fact, immunoglobulin should be playing a role as part of the immune system in defending against diseases as it is an antibody (Cui et al., 2021). However, Cui et al. (2021) state that immunoglobulin can appear in high levels in cancer cases. According to Zhong et al. (2021), immunoglobulin IgA is shown in elevated levels when the individual has a tumor and will be in more elevated levels when they are in the late stages of cancer.

Chapter Five: Conclusion

The objective of this thesis was to determine the gene expression terms that are correlated significantly with the TAGs in each breast and prostate cancer ST slide. We have found some gene terms that are common in all cancer slides. In addition, we have found some gene terms are common between slides of the same cancer type. These results were analyzed by generating a correlation matrix, and then we used functional enrichment analysis to determine the gene expression terms that were most highly correlated with each gene cluster. We conclude that extracellular gene terms are the most common terms in breast cancer and prostate cancer slides, and these terms are extracellular space, extracellular exosome, extracellular vesicle, extracellular organelle and extracellular membrane-bounded organelle. Also, Collagen Type I Trimer has a significant p-value in breast cancer and prostate cancer slides. In breast cancer slides, ZAG-PIP Complex gene expression was common within these slides. In prostate cancer slides, IgA immunoglobin was common within these slides, which shows that this expression has a relationship with cellular topology in prostate tumors. In ST slides for prostates, ITF81-ITF231 are having the most similarity with ECM and ITFs using the correlation coefficient rank. We can conclude that cellular topology has some clear association to the gene expression profiles from breast and prostate cancers. Identifying the biological meaning behind cellular topologies can lead to improved image feature extraction and potential improvement of immunotherapeutic that rely on proper topologies between cancer and immune cells.

References

1. Abousamra, S., Belinsky, D., Van Arnam, J., Allard, F., Yee, E., Gupta, R., Kurc, T., Samaras, D., Saltz, J., & Chen, C. (2021). *Multi-Class Cell Detection Using Spatial Context Representation.* Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4005-4014.
2. Aukerman, A., Carrière, M., Chen, C., Gardner, K., Rabadán, R., & Vanguri, R. (2020). Persistent *Homology Based Characterization of the Breast Cancer Immune Microenvironment: A Feasibility Study.* 36th International Symposium on Computational Geometry (SoCG 2020).
3. Brassart-Pasco, S., Brézillon, S., Brassart, B., Ramont, L., Oudart, J. B., & Monboisse, J. C. (2020). *Tumor Microenvironment: Extracellular Matrix Alterations Influence Tumor Progression.* Frontiers in oncology, 10, 397. https://doi.org/10.3389/fonc.2020.00397
4. Cui, M., Huang, J., Zhang, S., Liu, Q., Liao, Q., & Qiu, X. (2021). *Immunoglobulin Expression in Cancer Cells and Its Critical Roles in Tumorigenesis.* Frontiers in immunology, 12, 613530. https://doi.org/10.3389/fimmu.2021.613530
5. Frantz, C., Stewart, K. M., & Weaver, V. M. (2010). *The extracellular matrix at a glance.* Journal of cell science, 123(Pt 24), 4195–4200. https://doi.org/10.1242/jcs.023820
6. Henke, E., Nandigama, R., & Ergün, S. (2020). *Extracellular Matrix in the Tumor Microenvironment and Its Impact on Cancer Therapy*. Frontiers in molecular biosciences, 6, 160. https://doi.org/10.3389/fmolb.2019.00160

7. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., & Peterson, H. (2020). *gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler.* F1000Research, 9, ELIXIR-709. https://doi.org/10.12688/f1000research.24956.2
8. Kolde, R. (2018). *Raivokolde/pheatmap*. Github. Retrieved from https://github.com/raivokolde/pheatmap/blob/master/DESCRIPTION
9. Liu, Y., Ye, X., Yu, C., Shao, W., Hou, J., Feng, W., Zhang, J., & Huang, K. (2020). *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 Bioinformatics 2021, 22(Suppl 4):111 https://doi.org/10.1186/s12859-021-03964-5
10. Loughrey, C., Fitzpatrick, P., Orr, N., & Jurek-Loughrey, A. (2021). *The topology of data: opportunities for cancer research*. Bioinformatics. Volume 37, Issue 19, Pages 3091–3098, https://doi.org/10.1093/bioinformatics/btab553
11. Nallanthighal S, Heiserman JP and Cheon D-J. (2019). *The Role of the Extracellular Matrix in Cancer Stemness.* Front. Cell Dev. Biol. 7:86. doi: 10.3389/fcell.2019.00086.
12. National Cancer Institute. (2020). *Public Health Research and Cancer.* National Cancer Institute. Retrieved from https://www.cancer.gov/research/areas/public-health
13. National Cancer Institute. (2021). *What Is Cancer?.* National Cancer Institute. Retrieved from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

14. Sarkar, T. (2019). *Clustering metrics better than the elbow-method.* Towards Data Science. Retrieved from https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6

15. Silva, J.A.F., Biancardi, M.F., Stach-Machado, D.R. et al. *The origin of prostate gland-secreted IgA and IgG.* Sci Rep 7, 16488 (2017). https://doi.org/10.1038/s41598-017-16717-3

16. Singer, J., Irmisch, A., Ruscheweyh, H., Singer, F., Toussaint, N., Levesque, M., Stekhoven, D., & Beerenwinkel, N. (2017). *Bioinformatics for precision oncology.* Briefings in Bioinformatics. Volume 20, Issue 3, Pages 778–788. https://doi.org/10.1093/bib/bbx143

17. Thanati, F., Karatzas, E., Baltoumas, F. A., Stravopodis, D. J., Eliopoulos, A. G., & Pavlopoulos, G. A. (2021). *FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists*. Biology, 10(7), 665. https://doi.org/10.3390/biology10070665

18. Urbaniak, A., Jablonska, K., Podhorska-Okolow, M., Ugorski, M., & Dziegiel, P. (2018). *Prolactin-induced protein (PIP)-characterization and role in breast cancer progression.* American journal of cancer research, 8(11), 2150–2164.

19. Versaggi, S., & De Leucio, A. *Breast Biopsy.* (2022). National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK559147/

20. Vlaeminck-Guillem V. (2018). *Extracellular Vesicles in Prostate Cancer Carcinogenesis, Diagnosis, and Management.* Frontiers in oncology, 8, 222. https://doi.org/10.3389/fonc.2018.00222

21. World Health Organization. (2022). *Cancer*. World Health Organization. Retrieved from Cancer (who.int)

22. Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., Zhang, W., Yu, X., & Liu, L. (2019). *The role of collagen in cancer: from bench to bedside*. Journal of translational medicine, 17(1), 309. https://doi.org/10.1186/s12967-019-2058-1

23. Zhao, Y., Zheng, X., Zheng, Y., Chen, Y., Fei, W., Wang, F., & Zheng, C. (2021). *Extracellular Matrix: Emerging Roles and Potential Therapeutic Targets for Breast Cancer.* Frontiers in oncology, 11, 650453. https://doi.org/10.3389/fonc.2021.650453

24. Zhong, Z., Nan, K., Weng, M., Yue, Y., Zhou, W., Wang, Z., Chu, Y., Liu, R., & Miao, C. (2021). *Pro- and Anti- Effects of Immunoglobulin A- Producing B Cell in Tumors and Its Triggers*. Frontiers in immunology, 12, 765044. https://doi.org/10.3389/fimmu.2021.765044

Curriculum Vitae

Lujain Alsaleh

Education

- Master of Science in Biostatistics, earned at Indiana University-Purdue University Indianapolis, (May 2022).
- Bachelor of Science in Biochemistry, earned at King Abdulaziz University, (2015).

Professional Experience

Payroll officer - Enaya Company, (2015-2017)

Publications

Alsaleh, L. A., & Gull, M. (2016). Growth, origin and applied potential of embryonic stemcells: A recent fact sheet about stem cell therapies as a promising option for incurable diseases in humans. The Pharma Innovation 5(4), 09-13.