Processing of Imaging Mass Cytometry Data from Pancreatic Cancer Samples

Nathalia Kim
kim.n@queensu.ca

Abstract

The purpose of this study was to analyze imaging mass cytometry data to obtain information about the tumor microenvironment and the pancreatic cancer progression. DBSCAN clustering algorithm was employed to identify and measure the presence of several markers in the pancreatic tumors setting. The results of this study suggest that the presence of cancer cells could inhibit the infiltration of T cells in the tumor microenvironment. This could be interpreted as an inhibition of the immune system which could affect it's ability to prevent tumor progression.

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most aggressive types of cancer, having a poor prognosis. Worldwide, PDAC is the fourth most common cause of cancer-related deaths, with less than 8% of patients reaching 5-year survival [1]. Given this scenario, many efforts have been devoted to better understand this disease in order to improve clinical outcomes.

One of the relevant areas in the study of cancer is the tumor microenvironment, composed of several types of cells that cooperate with cancer cells to promote tumor progression. A state-of-the-art technique to analyze the tumor microenvironment is imaging mass cytometry (IMC), which enables the localization and visualization of multiple proteins simultaneously in tissue samples at subcellular resolution [2].

This project aims to analyze 9 pancreata of PDAC patients using IMC to investigate the spatial relationship of different cell types and better comprehend the tumor microenvironment. A density-based clustering approach was employed to identify the cells by quantifying the proteins of each sample. The clustering algorithm was applied to each image of all samples and measurements of the clusters were taken for further analysis.

Materials and Methods

Dataset

The imaging dataset for this project was acquired by performing IMC to 9 pancreata obtained from patients with PDAC. For each sample, 3 to 4 regions of interest (ROI) were identified, resulting in a total of 36 ROIs and therefore 1,656 images. Figure 1 shows representative raw ROI images. A panel of 46 antibodies was employed to identify different cell types, with a focus on immune cells. The dataset was acquired and shared by the Penn Pancreatic Cancer Research Center (PCRC) from the University of Pennsylvania.

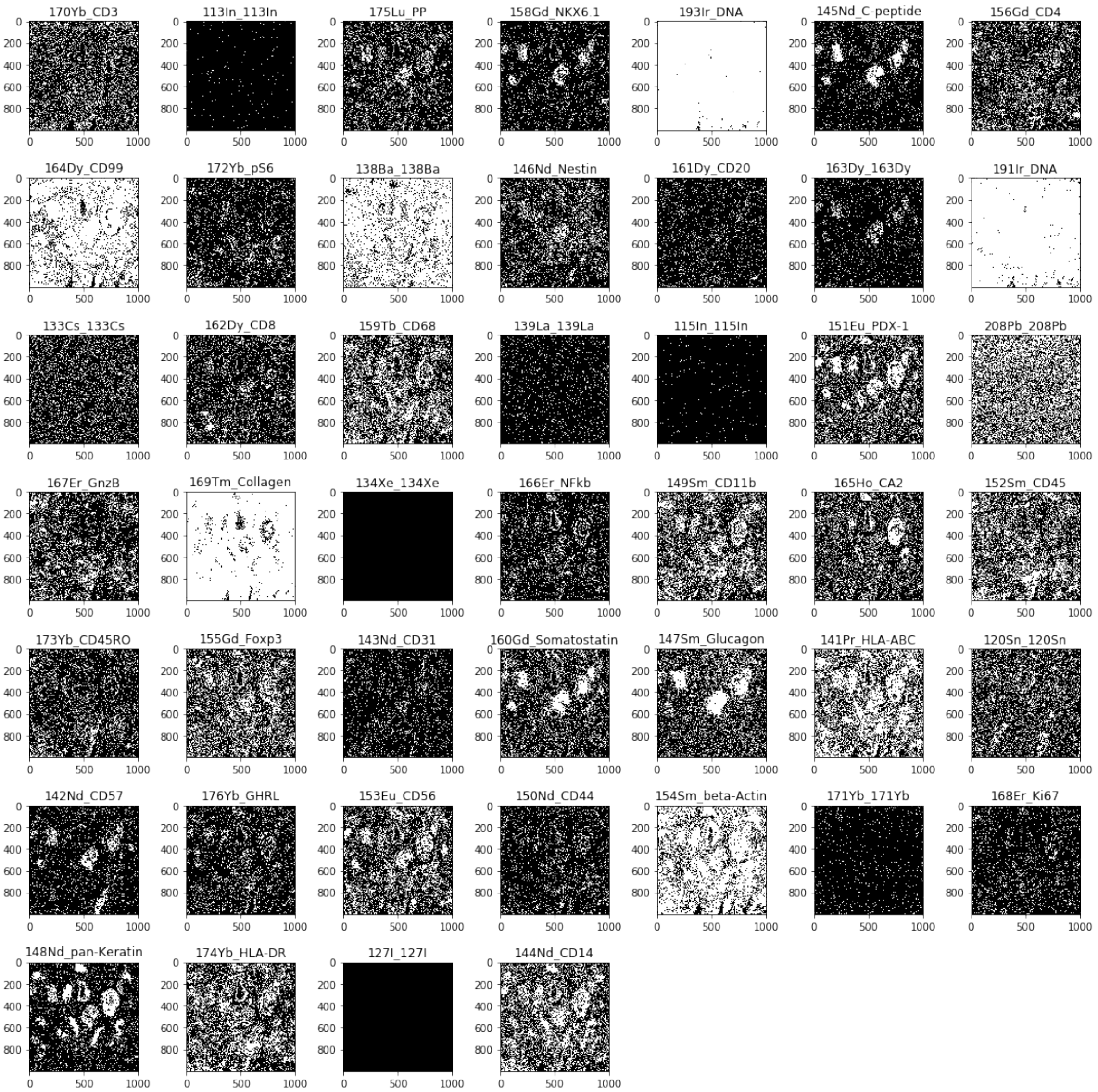
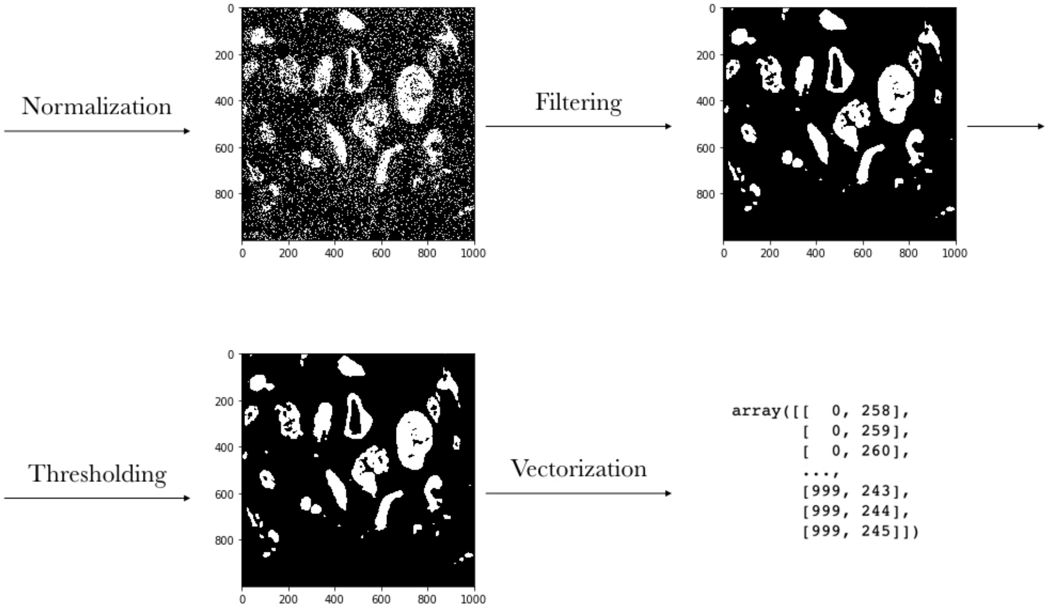


Figure 1: Representative raw ROI images. The 46 images represent the 46 markers identified in the samples. The title of each image is composed of (heavy metal used to stain the protein)_(name of the marker). The white regions represent the presence of that marker in the sample.

Preprocessing

Images were first normalized to a range of (0,1) to set them to a common scale. Then, a median filter was applied for denoising and an automatic image thresholding was perfomed to get a binary image. Otsu thresholding was chosen to preserve the edges. The last step was to transform the image to a vector size, storing in a vector the indexes of the pixels that had True values in the binary image. This vector will be the input for the clustering method next.

The pan-Keratin marker is of great interest for this project, given that it can identify cancer cells in the PDAC environment. The scheme of the preprocessing steps in a representative pan-Keratin image is shown in Figure 2.



```
array([[  0, 258],
       [  0, 259],
       [  0, 260],
       ...,
       [999, 243],
       [999, 244],
       [999, 245]])
```

Figure 2: Preprocessing steps on a representative pan-Keratin image.

Clustering

A density-based spatial clustering of applications with noise (DBSCAN) was employed to detect the structures in the images and subsequently the different cells. DBSCAN was the chosen method since it can find clusters based on the density of points, handle noise and find clusters of arbitrary shape [3]. After tuning, an epsilon value of 0.05 and minimum points of 5 were used. Figure 3 shows the result of DBSCAN on a representative pan-Keratin image, where 41 clusters were found. By visual inspection, the clusters in Figure 3 were well defined by the algorithm, since it could differentiate most of the structures of the image.

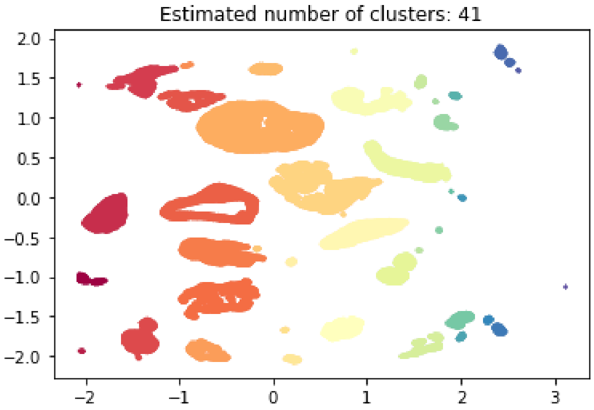


Figure 3: Resulting clusters from DBSCAN on a representative pan-Keratin image. The different clusters appear on different color tones.

After clustering, a set of measurements were taken to evaluate the model and for further analysis. Since the truth class assignments of the samples are not known, the evaluation of performance of the model used the Davies-Bouldin Index, which was chosen due to its computation being simpler than the silhouette coefficient. In addition, the number and size of the clusters were quantified for each of the markers for all corresponding ROIs.

Results and Discussion

Image Data Processing and Clustering

First, the preprocessing steps and DBSCAN clustering algorithm were applied to the entire dataset in order to identify the structures in the images. Figure 4 shows the resulting clusters on a representative ROI sample.

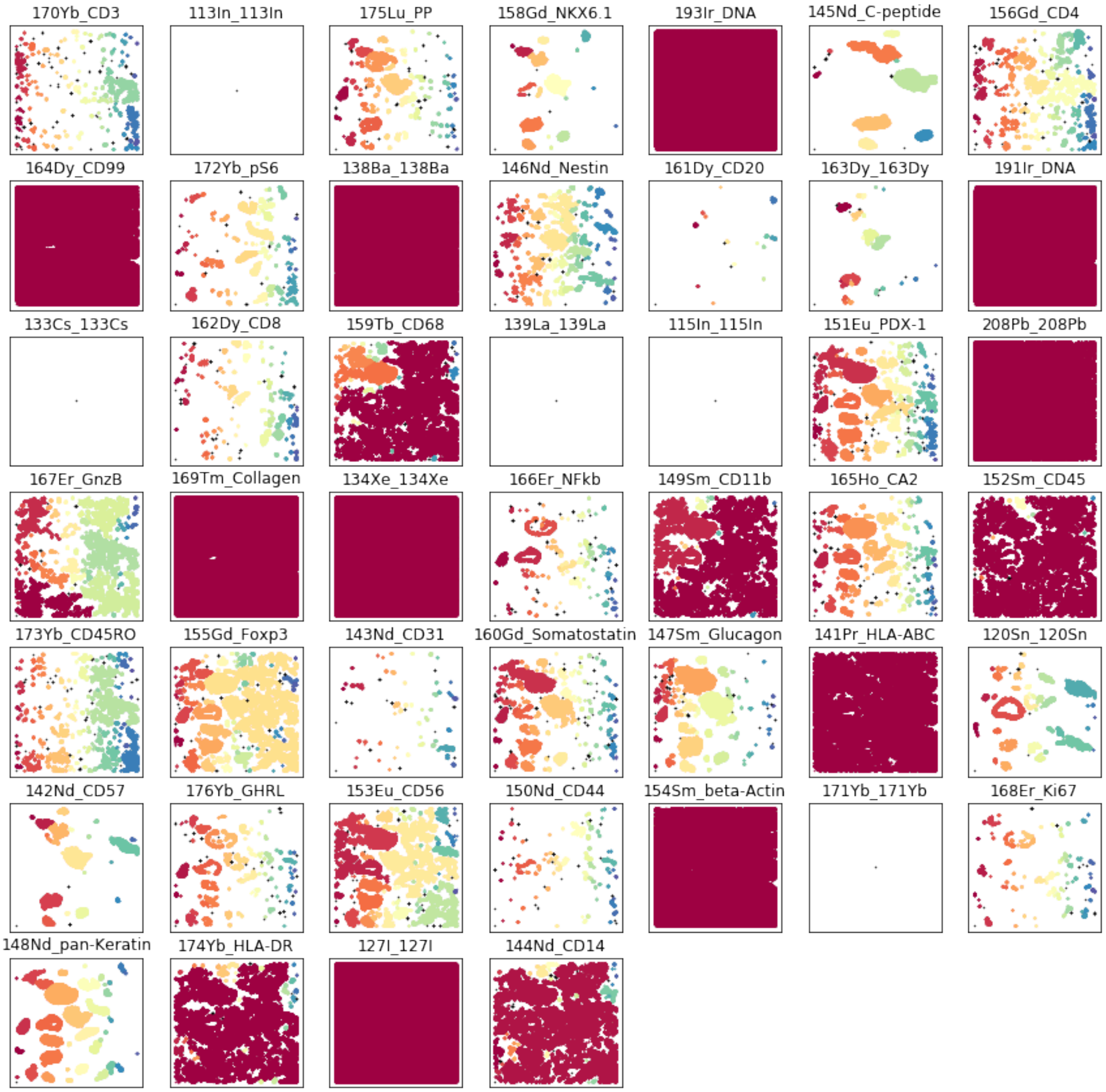


Figure 4: Resulting clusters from DBSCAN on a representative ROI sample. Each image is related to a specific marker. The different clusters appear on distinct color tones.

In Figure 4, some of the images resulted in only one cluster, identified by the predominant red or white images. For the controls (such as 138Ba_138Ba and 171Yb_171Yb) this can be explained by the fact they are not biological markers and consist of mostly noise that could not be distinguished by the clustering algorithm. However, for some markers such as 193Ir_DNA it was not expected to obtain only one cluster, since it should be able to identify the nucleus of the cells.

Clustering Measurements

Next, the number and size of clusters and Davies-Bouldin index were measured for clustering evaluation and analysis. As expected, the number of clusters had great variation between the different markers (Figure 5A).

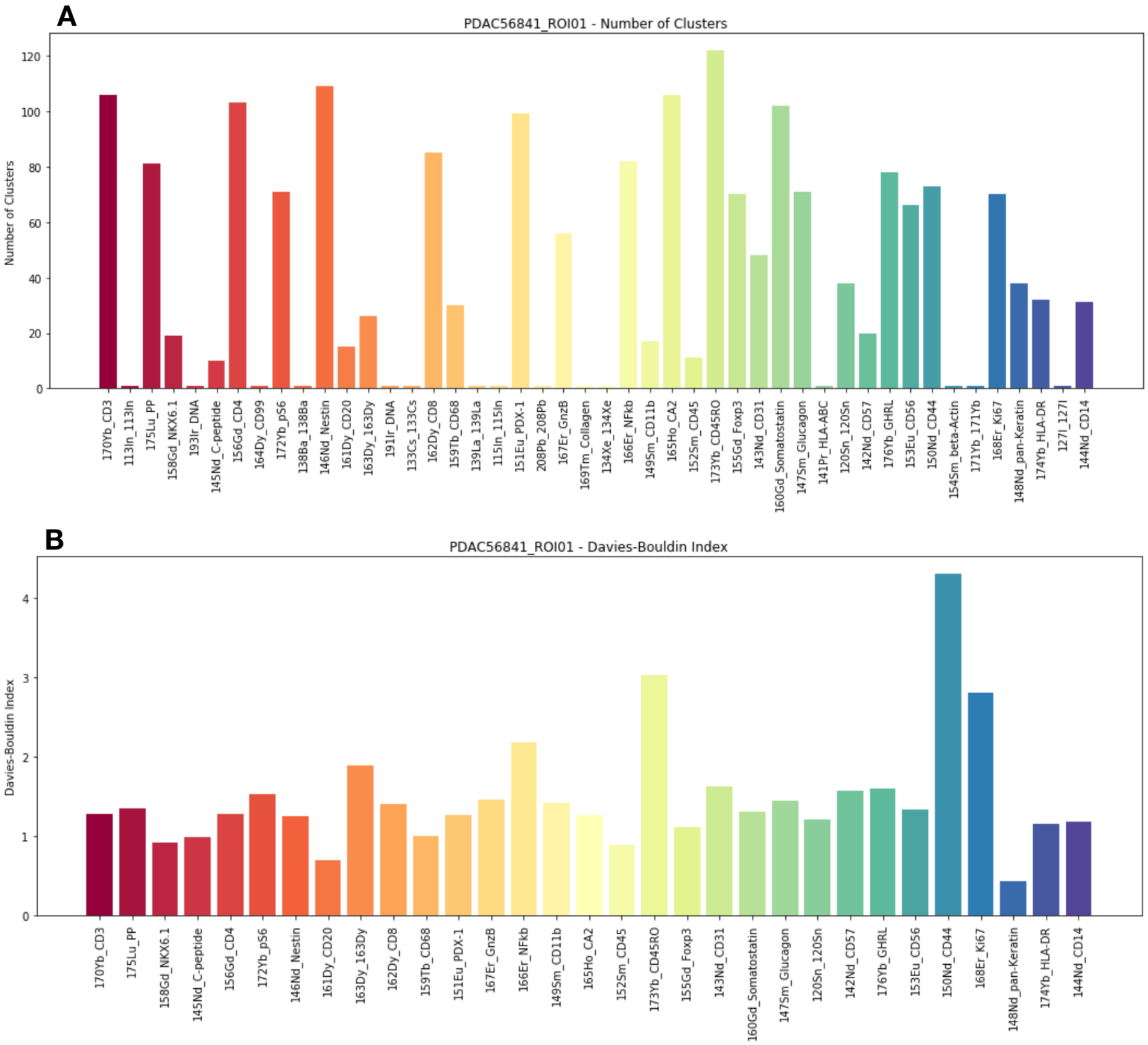


Figure 5: Resulting clustering measurements on a representative ROI sample. (A) Number of clusters for each marker of PDAC56841_ROI01. (B) Davies-Bouldin index for PDAC56841_ROI01.

Since the Davies-Bouldin index can only be measured if the number of clusters is greater than one, the measurement could only be taken for some of the markers (Figure 5B). Some of the indexes were higher than expected (such as for 150Nd_CD44, 168Er_Ki67 and 173Yb_CD45RO), however most of them fluctuate between 1.5 and 2. With the consideration that the Davies-Bouldin index is usually higher for convex clusters and that some images have structures dissimilar from each other, the resulting indexes can be considered satisfactory [4].

Data Analysis

Finally, given all the clustering measurements acquired from the entire dataset, additional analysis were made. To further evaluate the clustering model, the Davies-Bouldin index for all ROIs were averaged (Figure 6). As pictured, most of them are around 1.5 and with the same considerations in the analysis of Figure 5B, could be considered satisfactory.

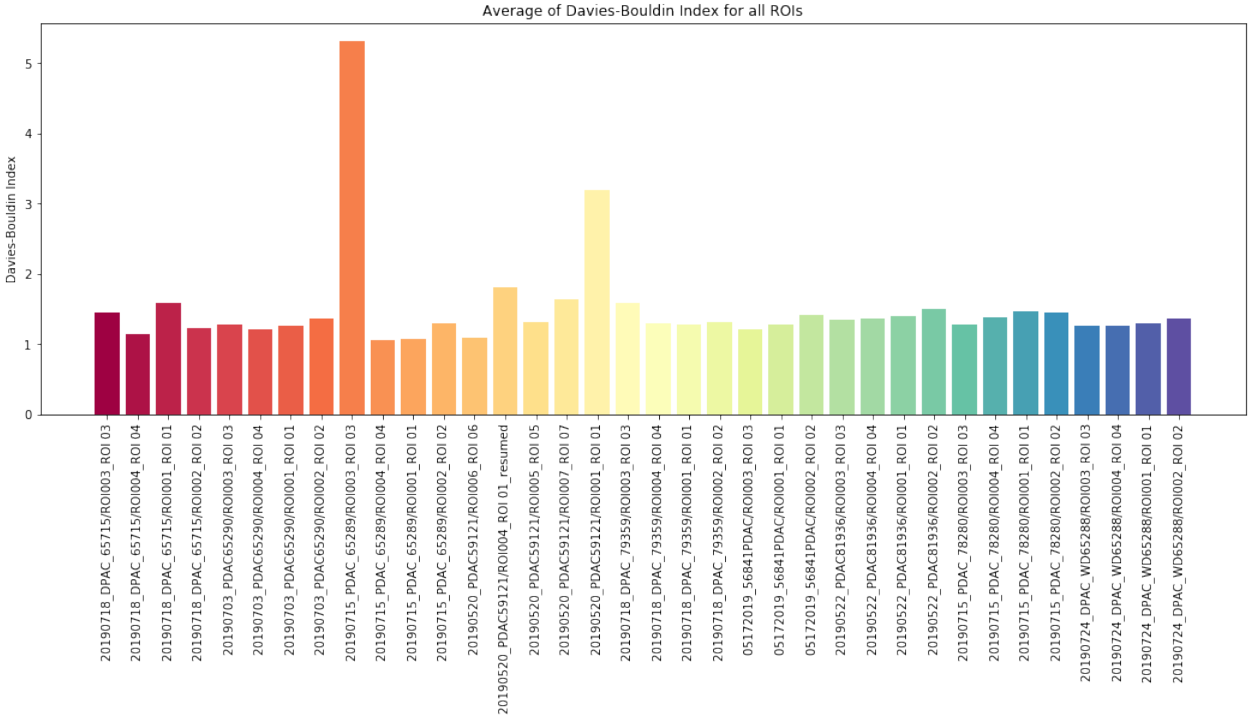


Figure 6: Average Davies-Bouldin index for all ROIs samples.

As previously stated, the pan-Keratin marker can identify cancer cells in PDAC. One of the relevant areas in the study of the tumor microenvironment is the level of infiltration by immune cells. In this matter, a comparison between the pan-Keratin and CD3 markers was performed for all ROIs (Figure 7), since the CD3 marker could indicate the presence of T cells in the tissue.

As pictured in Figure 7, the number of clusters of CD3 is higher than of pan-Keratin for the majority of ROIs (Figure 7A). However, the total area of clusters of pan-Keratin is usually greater than CD3 (Figure 7B). This could indicate that the presence of CD3 markers is more sparse in PDAC tissue while the pan-Keratin marker represent regions of high density of tumor cells.

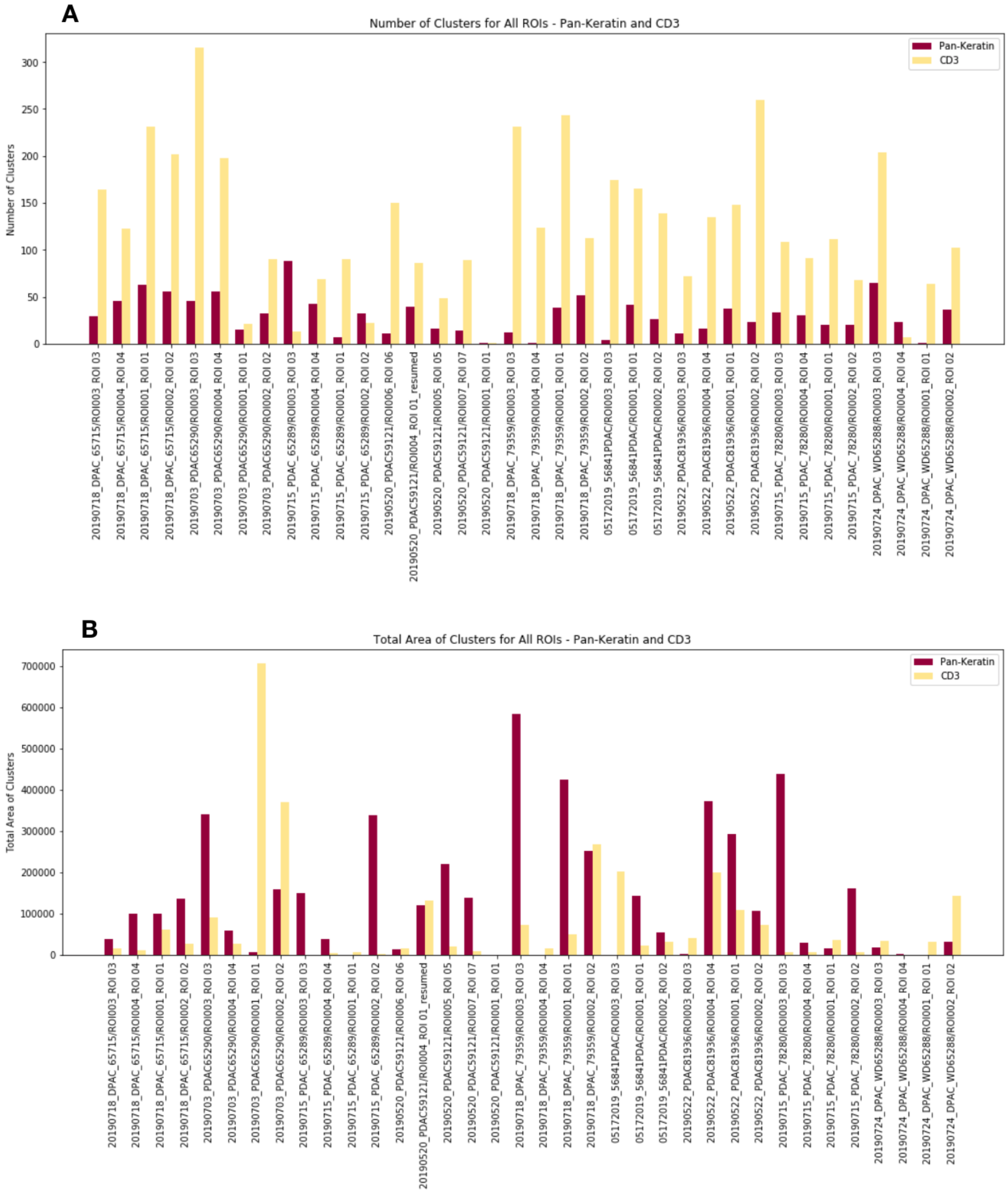


Figure 7: Comparison between pan-Keratin and CD3 markers for all ROIs. (A) Comparison of number of clusters. (B) Comparison of total area of clusters.

In Figure 7B, it is also noteworthy that the ROIs that have the highest total area of clusters of CD3 (such as 20190703_PDAC65290/ROI001_ROI 01 and 20190703_PDAC65290/ROI002_ROI 02) also have low total area of pan-Keratin. This suggests that an inhibition of the infiltration of T cells could be produced by the presence of tumor cells. This inhibition could affect the immune system's ability to prevent tumor progression.

Conclusions

Imaging mass cytometry is a valuable technique that provide the means to analyze several markers simultaneously. The application of DBSCAN clustering on IMC data has proven capable of identifying the different structures of the markers.

It is acknowledged there are limitations to this study and room for improvements. For instance, the size of the dataset restricts the degree of analysis that can be made. Regarding the clustering results, for some of the markers it was expected to obtain various clusters but only one was acquired (for example, 193Ir_DNA marker should identify the nucleus of the cells). The reason for this could be the filtering and thresholding steps of preprocessing that removed desired information from the images.

Concerning model evaluation, the Davies-Bouldin index had levels higher than expected for some samples. This could be explained by the nature of the data, where each marker consists of different shapes and sizes of clusters and therefore could present high dissimilarity.

With regard to the infiltration of the tumor microenvironment by immune cells, the results suggest that the presence of cancer cells could inhibit the infiltration of T cells and thus affect tumor progression.

Finally, it was confirmed that IMC data can provide valuable information about the tumor microenvironment. With the measurements resulting from this work several subsequent analysis can be carried. Moreover, the data can be additionally interrogated employing different data analytics methods, which is intended to be addressed in a future work.

References

[1] M. Orth. "Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches". In: *Radiation Oncology* 14.1 (2019), p. 141. DOI: 10.1186/s13014-019-1345-6.

[2] Fluidigm. *Imaging Mass Cytometry: Visualize pathology and disease with high-multiplex imaging.* URL: https://www.fluidigm.com/applications/imaging-mass-cytometry. (accessed: 13.04.2020).

[3] S. Ding. *CISC 372 Clustering.* URL: https://github.com/L1NNA/L1nna.github.io/blob/master/372/L13%5C%20Clustering.pdf. (accessed: 13.04.2020).

[4] Scikit-learn. *2.3. Clustering.* URL: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation. (accessed: 13.04.2020).