BIOS 611 Final Project

Spatial transcriptomic data from human breast cancer tissue

Name: Hongyu Yu PID: 730440679

1 Introduction

The human breast cancer spatial transcriptomic data used in this report was from Ståhl et al, 2016. The processed data in `.rds` format was downloaded from GitHub, which was uploaded by the authors as a supplementary for a methodology paper published in Nature Methods. The data contains the gene expression profile of approximately 15,000 genes at about 250 positions in a human breast cancer tissue, and thus each entry represents the number of detected mRNA associated with a gene in a location.

Compared to traditional mRNA sequencing data, which only has mRNA count data for each cell or cell groups, spatial transcriptomic data also contain the information about where each cell or cell group locates in the tissue. This additional dimension allows the analysis of transcription pattern across the tissue. Some questions that were hard to answer using mRNA-seq alone became easy to understand using spatial transcriptomic data.

2 Results

2.1 Expression patterns of highly variable genes

In a tissue, most genes exhibit similar expression patterns across different cells. These genes can include the house keeping genes, which maintain the basic metabolism of each cell and thus are constantly expressed, or other lowly expressed genes. Since the overall goal is to analyze the expression pattern across the tissue, the main focus is on genes differentially expressed in the experiment. Therefore, the standard deviation of the mRNA count of each gene is calculated and plotted (Figure 1 left panel). As expected, most genes showed similar expression or were sequenced insufficiently, so the majority has a standard deviation below 5. The top 10 variable genes were showed on the right panel of Figure 1, which were further visualized across the tissue.

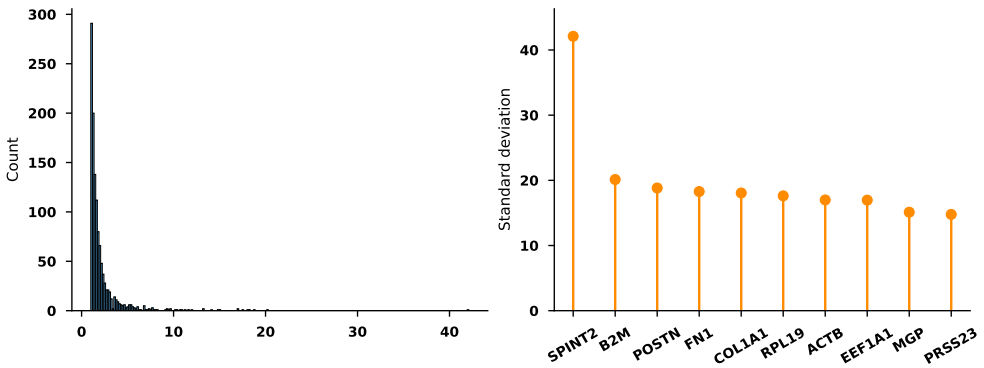

Figure 1: standard deviation of measured genes

The expression of highly variable genes is spatially correlated as shown in Figure 2. For example, the gene `COL1A1` is highly expressed in the lower part of the tissue while lowly expressed

in other regions. Genes `FN1` and `POSTN` also show similar trend. In contrast, `SPINT2` is only expressed on the left and right corner of the tissue but lowly expressed in the middle part. These together show that gene expression is related to the spatial locations of cells, and different genes have different expression patterns.

Figure 2: expression of genes across the tissue

2.2 Confounding of gene expression

The expression pattern observed in the previous section might be attributed to the variation of cell activity across the tissue, but it might also be due to the different sequencing depth at each location. To investigate this potential confounding, the mean expression count of low-variation genes (genes with standard deviation smaller than 5 across the tissue), is plotted. If all the locations were sequenced similarly well, the mean expression count is expected to be about the same across the entire tissue. However, as shown in Figure 3, the mean count is not uniform.

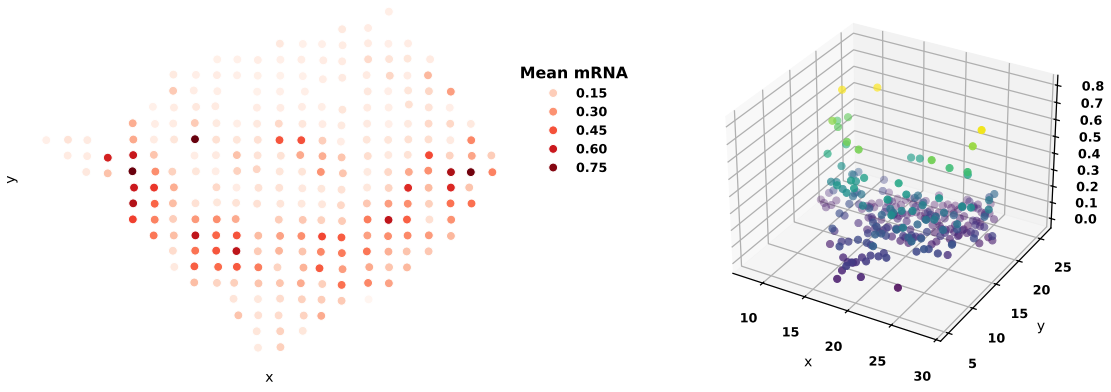

Figure 3: mean expression of low variation genes

To examine how the mean expression count affects the expression pattern, the mRNA count at

each location for each gene is divided by the average. The mean count of low variation genes is 1 after normalization (Figure 4).

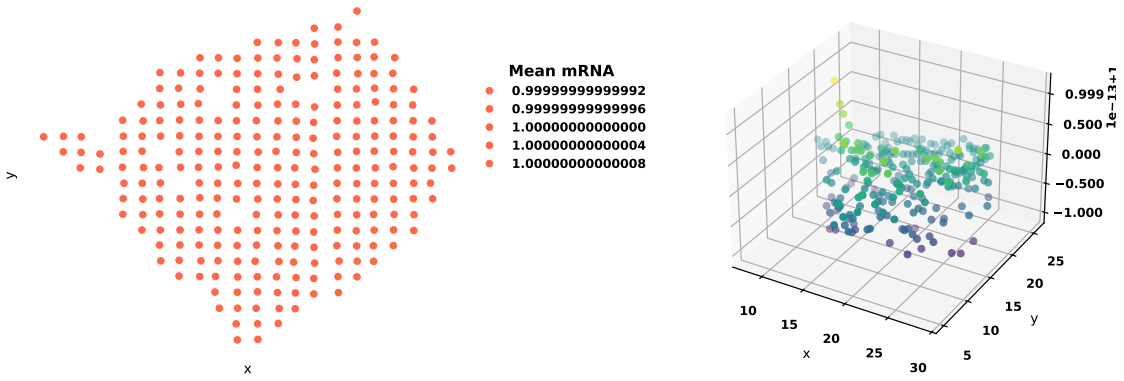

Figure 4: mean expression of low variation genes after normalization

The expression patterns of highly variable genes also change as shown in Figure 5. For example, `SPINT2` previously has expression concentrated in the left and the right edge of the tissue while it now has a nearly uniform expression in the entire region. However, the normalization did not remove all the differential expression patterns observed. For `FN1` and `COL1A1`, the expression profiles became more concentrated near the bottom of the region (Figure 5).

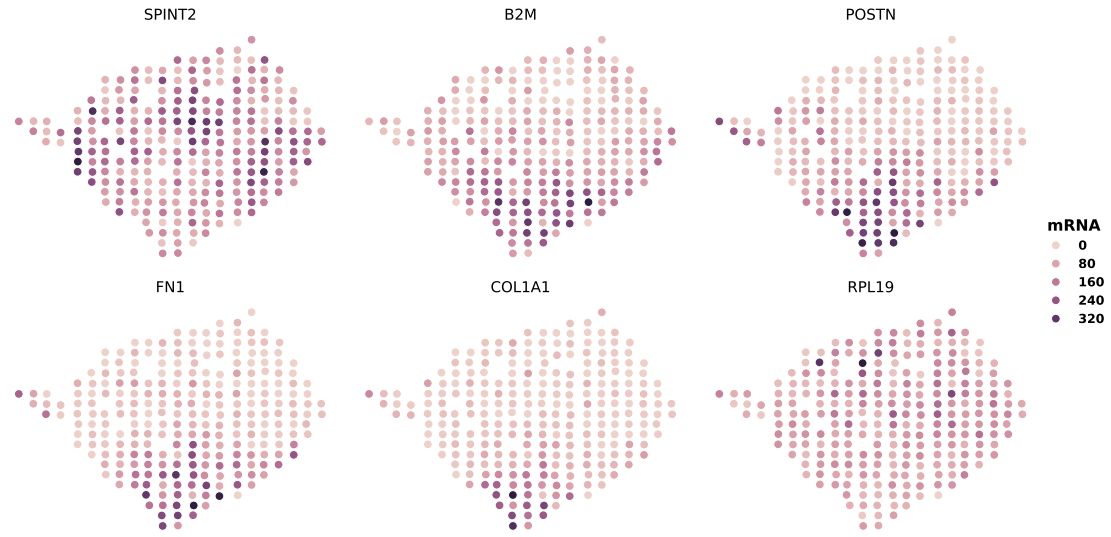

Figure 5: normalized expression of genes across the tissue

2.3 Dimension reduction and clustering

Another important task in spatial transcriptomics is to identify spatial domains, regions with distinct functions and microenvironments. This can be viewed as a clustering task. Due to the high-dimensionality of transcriptomic data (large number of genes sequenced compared to the number of cells or cell groups), the expression profile are rarely used directly for clustering. Instead, they

are first projected to a lower dimensional space, and then clustering is performed on this low dimensional embedding.

In this section, two different dimension reduction methods were tested. The first is a usual PCA, formulated in its variational form as

$$\min_{\mathbf{Z},\mathbf{W}} \left\| \mathbf{X} - \mathbf{Z}\mathbf{W}^\top \right\|_2^2 \quad \text{s.t.} \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I},$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the gene expression data at each location with X_{ij} being the mRNA count of the jth gene in the ith cell, $\mathbf{W} \in \mathbb{R}^{k\times p}$ is an orthognal eigenvector matrix, and $\mathbf{Z} \in \mathbb{R}^{n\times k}$ is the k-dimensional embedding of all n cells. Write $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. The solution is $\mathbf{Z} = \mathbf{U}_k\mathbf{\Sigma}_k$ and $\mathbf{W} = \mathbf{V}_k$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ represent the first k left and right singular vectors and $\mathbf{\Sigma}_k$ denote a $k \times k$ diagonal matrix with diagonal elements being the largest k singular values.

The embedding and clustering result of PCA is shown in Figure 6, left panel. The bottom region of the tissue is identified as one cluster, coherent with previous observations that some highly variable genes show clustered expression at the bottom.

Another dimension reduction method is the `GraphPCA` proposed in a Genome Biology paper. To incorporate the spatial dependency observed in typical spatial transcriptomic data, `GraphPCA` includes an additional adjancency penalty into PCA. The optimization problem is

$$\min_{\mathbf{Z},\mathbf{W}} \left\| \mathbf{X} - \mathbf{Z}\mathbf{W}^\top \right\|_2^2 + \lambda \mathrm{Tr}(\mathbf{Z}^\top\mathbf{L}\mathbf{Z}) \quad \text{s.t.} \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I},$$

where $\mathbf{L}$ is the Graph Laplacian defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the degree of each nodes and $\mathbf{A}$ is the adjacency matrix of a graph. The graph used here is a k-NN graph, which, for each spatial location, identifies the nearest k-locations and connect them as neighbors.

Like PCA, `GraphPCA` has closed form solution. First, it can be shown that the solution $\mathbf{Z} = \mathbf{K}\mathbf{X}\mathbf{W}$, where $\mathbf{K} = (\mathbf{I} - \lambda\mathbf{L})^{-1}$. Then the problem reduces to

$$\min_{\mathbf{W}} \mathrm{Tr}(-\mathbf{W}^\top\mathbf{X}^\top\mathbf{K}\mathbf{X}\mathbf{W}) \quad \text{s.t.} \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I},$$

which gives that $\mathbf{W} = \mathbf{V}_k$, where $\mathbf{V}_k$ is the k eigenvectors of $\mathbf{X}^\top\mathbf{K}\mathbf{X}$ associated with the largest k eigenvalues. Therefore, the embedding is $\mathbf{Z} = \mathbf{K}\mathbf{X}\mathbf{V}_k$.

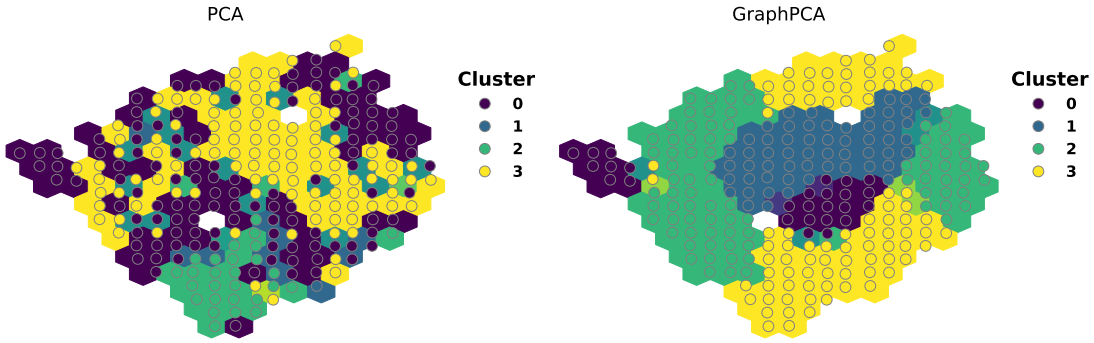

Figure 6: clustering based on PCA and GraphPCA

The `GraphPCA` clustering result is shown in the right panel of Figure 6. The λ (Graph Laplacian penalization coefficient) and k (number of neighbors) are set as 3 and 5, respectively. As expected, `GraphPCA` yields more spatially continuous clusters compared to plain PCA.

However, a limitation of `GraphPCA` is that the user has to identify the hyperparameters λ and k, and as shown in Figure 7. As the penalty term increases, a stronger spatial correlation is expected in the final clustering result though the specific effect also depends on k, the number of neighbors in the k-NN graph constructed.

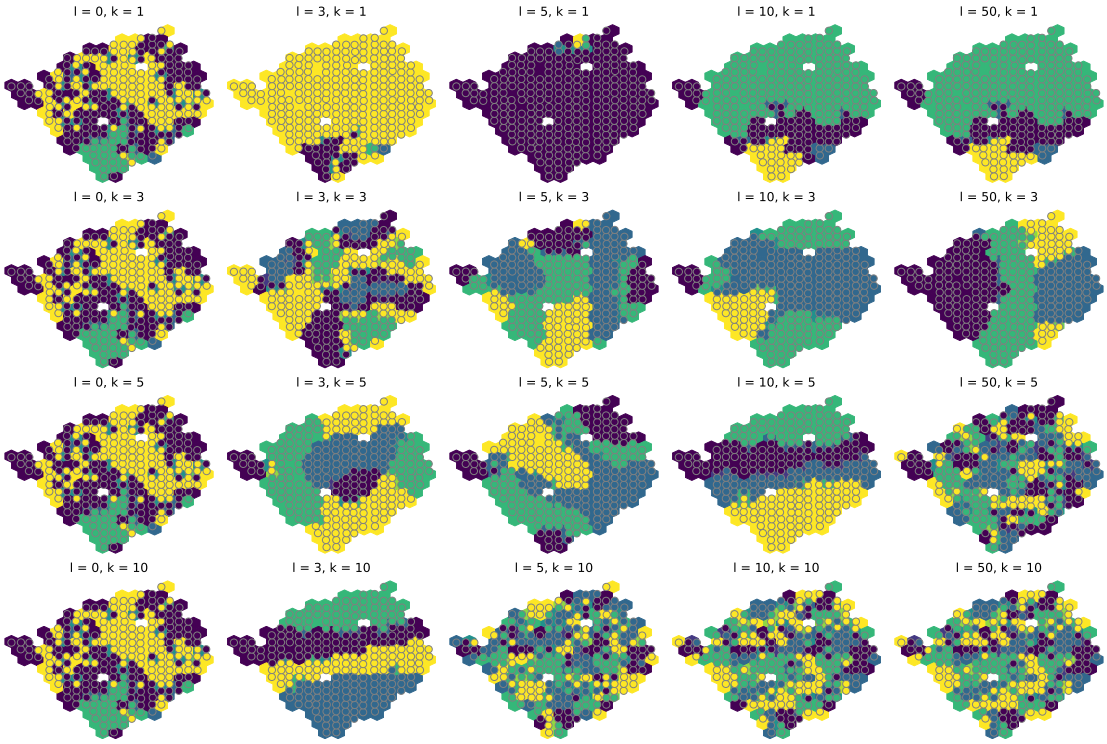

Figure 7: effects of different values of hyperparameters

3 Conclusion

In this report, the spatial transcriptomic data of human breast cancer is re-analyzed. In summary, highly variable genes show distinct expression patterns across the tissue, but the difference in sequencing depth at each spatial location also affects the pattern observed. In addition, clustering results based on PCA and `GraphPCA` are presented. Though `GraphPCA` better captures the spatial dependency, it is sensitive to the choice of hyperparameters.