**Computational Analysis of Single Cell RNA Sequencing Data from Human Livers Reveals Distinct Hepatic Cellular Populations**

**Introduction:**

The liver is responsible for a variety of metabolic processes necessary for the proper functioning of the human body including gluconeogenesis, the urea cycle and fatty acid metabolism.4 While its role as the metabolic center of an organism is well-known, the liver plays an equally important role in the immune system as it contains unique immune cells, largely in an activated state, that work to establish an organism’s peripheral immune tolerance.4

As can be expected from its functional diversity, the liver displays a great deal of cellular diversity in both parenchymal and non-parenchymal cells.3,4,5 Current understanding of this cellular diversity stems from our knowledge of the hepatic acinus, the building block of the liver.5 The acinus is composed of hepatocytes and a wide variety of non-parenchymal cells including endothelial cells, cholangiocytes, hepatic stellate cells, Kupffer cells, and a variety of lymphocytes such as B cells, conventional, and non-conventional T cells and natural killer cells.5 Though the aforementioned liver immune cells are localized in the liver following specific patterns, little is known about their cellular localization and phenotype.3,5

Additionally, the mammalian liver consists of hexagonal lobules that are polarized radially.3 The liver has been shown to be localized along this polarization in the sense that key liver genes are differentially expressed along the lobule axis in distinct layers.3 A study conducted by the Halpern et al.3 has established the existence of and transcriptional profile of 9 unique layers of mice liver cells. Little is known, however, as to how the observed layers in mice livers apply in the context of human livers.3

In this work, I take data from the MacParland et al.5, who applied a series of previously developed liver dissociation techniques to isolate individual cells from 5 donated patient livers and perform single cell RNA sequencing (scRNA-seq) to examine the unique cellular landscape of the liver. Through application of principal component analysis and k-means clustering, I identify 8 distinct hepatic cell populations from the transcriptional profiling of 8444 cells. By using known cell-specific gene markers and identification of differentially expressed genes, I find distinct populations of hepatocytes and non-parenchymal cells. I then correlate the identified populations of hepatocytes with the 9 layers of liver cells found in mice liver.

**Computational Methods:**

*Cluster Identification*

Processed scRNA-seq data was obtained from MacParland et al.5, through the NCBI GEO (accession GSE115469) and converted into a table, with columns indicating individual cells and rows indicating unique genes. Principal component analysis was employed using the default MATLAB R2019a implementation to reduce the dimensionality of the downloaded data. The number of components used was determined based on the number needed to explain 97% of the variance observed in the data. t-Distributed Stochastic Neighbor Embedding (t-SNE) was employed using the default MATLAB R2019a implementation to graphically visualize the number of clusters present in the data. This number was then used as an input in k-means clustering, using the default MATLAB R2019a implementation, to rigorously partition our data into biologically relevant clusters. The cell-type identities for each cluster were determined by normalizing average expression values (through z-score) of each cluster for a manually curated set of cell-type specific marker genes.

*Differentially Expressed Gene Identification and Pathway Analysis*

The average expression for each gene was calculated for each identified individual cluster. For each cluster, genes that were expressed at least five-fold higher in said cluster as compared to the remaining clusters were deemed to be differentially expressed (DE). DE genes for each cluster were fed into the PANTHER classification system which analyzed what pathways the genes are involved in and outputted a list of pathways suspected to be upregulated in the identified cluster.6

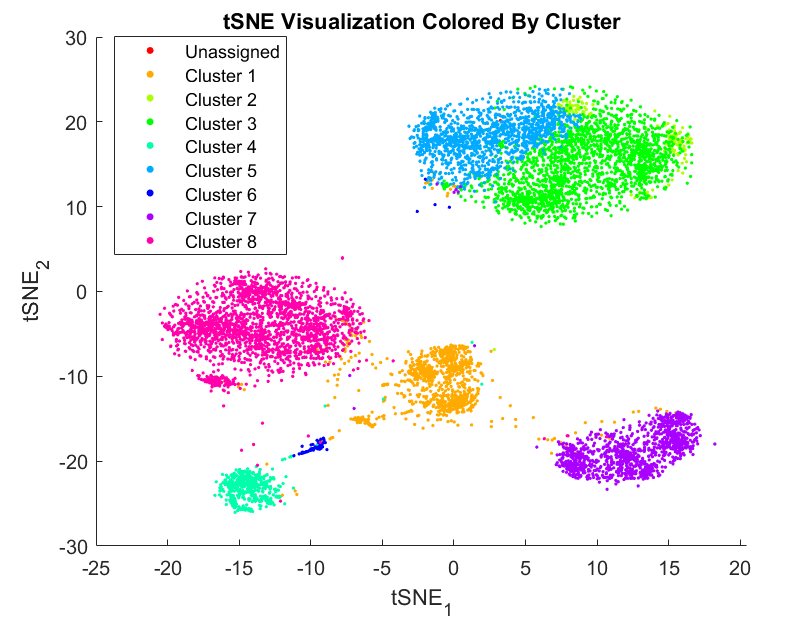
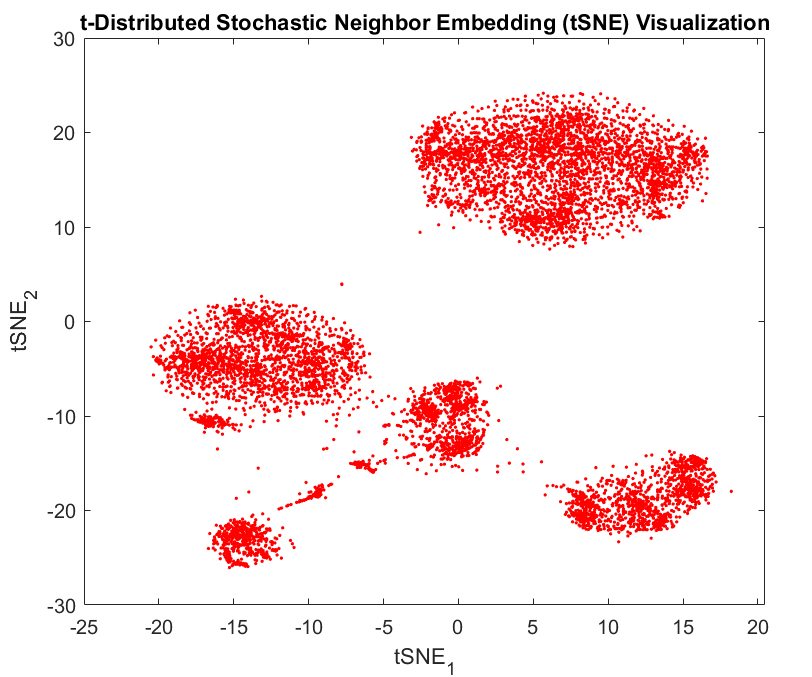
*Subcluster Identification*

If multiple cell types were detected in a single cluster, the possibility of subclusters or subpopulations within a cluster was considered. A new table composed only of cells belonging to the cluster of interest was created. t-SNE was employed on this new table using the default MATLAB R2019a implementation to graphically visualize the number of potential subclusters present in the data. This number was then used as an input in k-means clustering, using the default MATLAB R2019a implementation, to rigorously partition our data into biologically relevant clusters. The cell-type identities for each cluster were determined by normalizing average expression values (through z-score) of each cluster for a manually curated set of cell-type specific marker genes. If the identified subclusters were deemed biologically relevant, DE genes between subclusters were identified using the method described above for regular clusters. Only DE genes that were also differentially expressed in the corresponding regular clusters were considered true DE genes for the subclusters.

*Human Mouse Correlation*

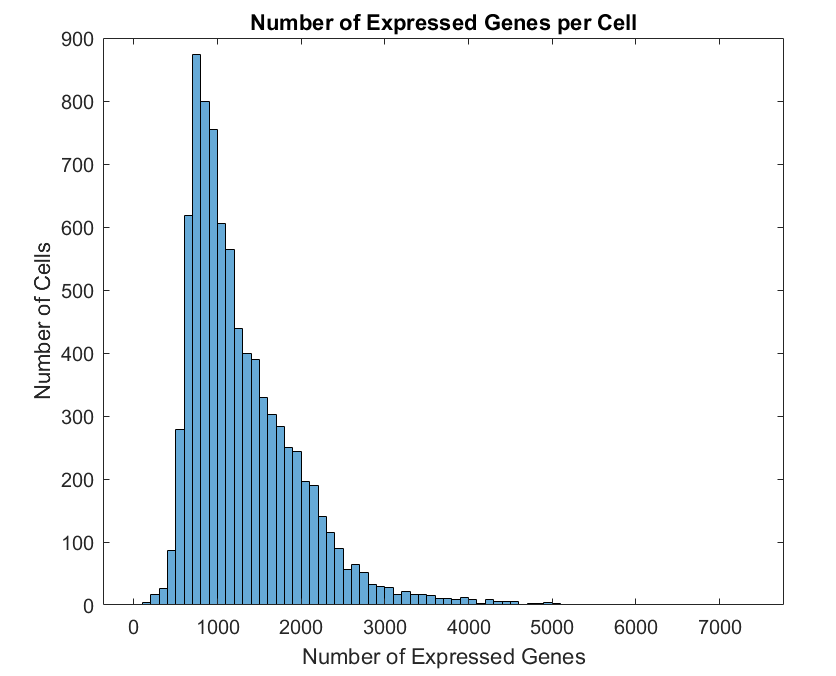
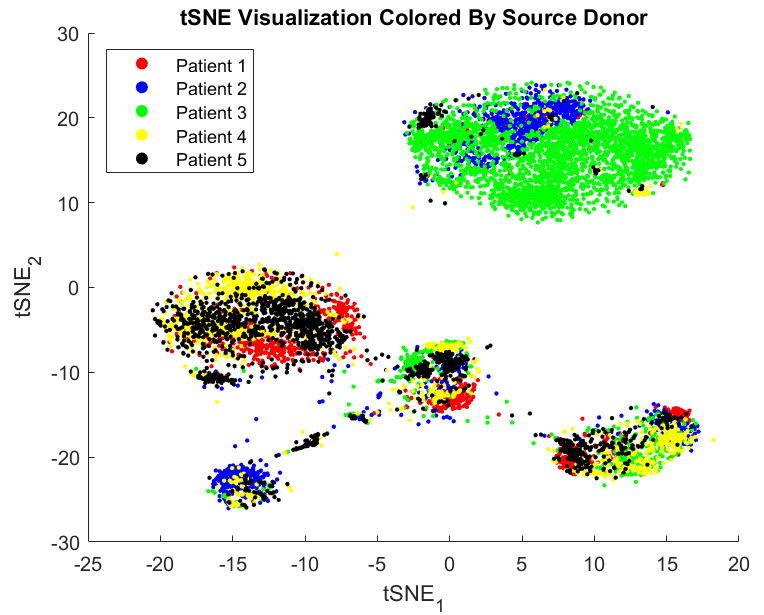
The significantly (p < 10-30) and DE genes observed by the Halpern et al.3 study for nine layers of mouse liver cells were identified. The Ensembl database was used to identify human-mouse one-to-one orthologous genes.2 Ultimately, 85 genes detected in both human and mouse were used for correlation analysis. Expression values for each gene among the three clusters of human hepatocytes were normalized by z-scores. The same procedure was repeated for the nine layers of mice liver cells. Finally, Pearson correlation coefficients, along with corresponding significance values, were calculated using the calculated z-scores across all 85 genes to correlate the three human hepatocytes clusters with the nine layers of mouse liver cells.

**Figure 1: 8 distinct cell populations were revealed in healthy human livers. a)** Histogram depicting the number of cells expressing a specified number of genes **b)** t-SNE projection of 8444 liver cells with each point representing a single cell **c)** t-SNE projection of 8444 liver cells with each point color coded according to patient liver source **d)** t-SNE projection of 8444 liver cells with each point color coded according to cluster identity assigned by k-means clustering algorithm **e)** Heat map analysis assigning the identity of each cluster by matching the cluster expression profile with a manually curated list of cell-specific marker gene expression for hepatocytes, endothelial cells and immune cells. Assignment of color to each cell of the heatmap was determined by z-score value as indicated by the legend



**b)**

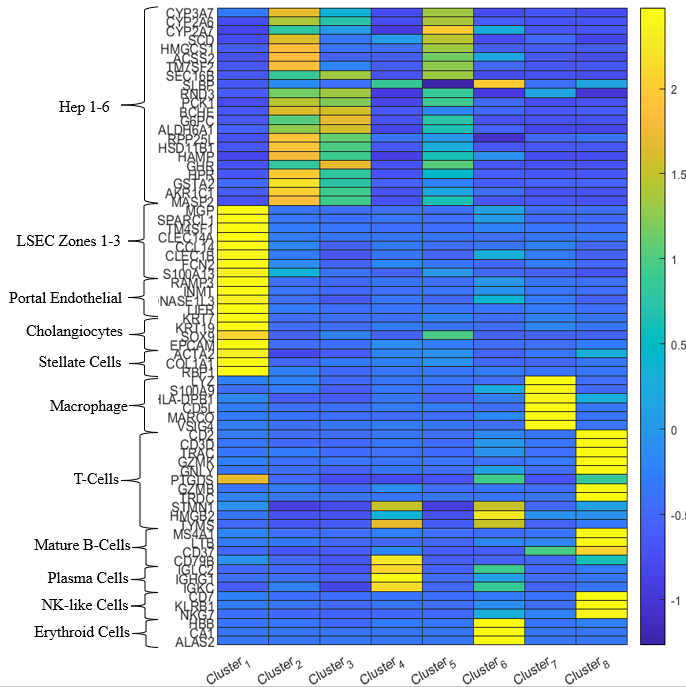
**d)**



**a)**

**c)**

**e)**



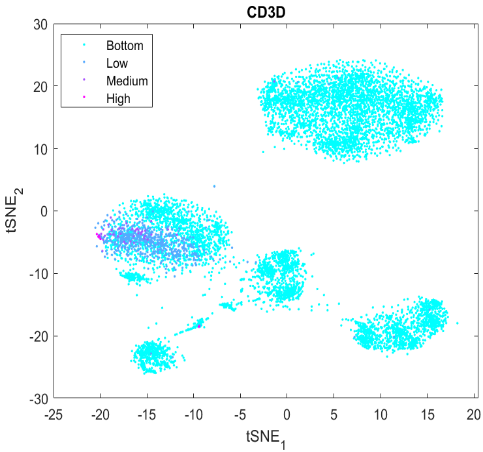
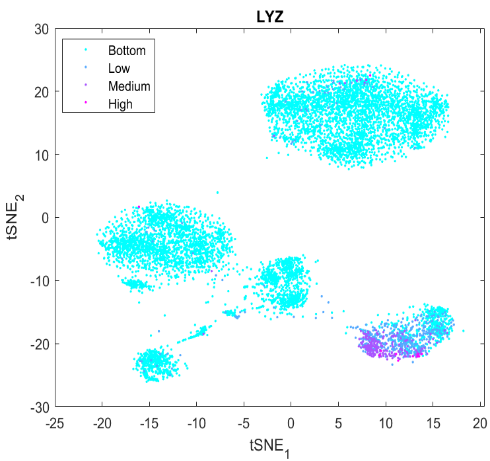
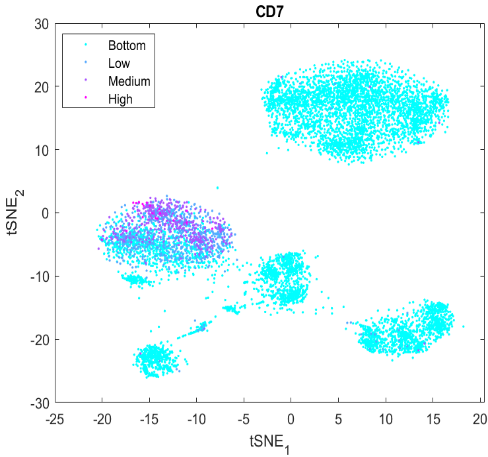
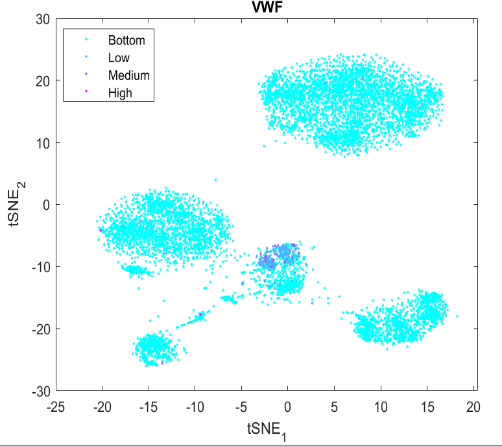
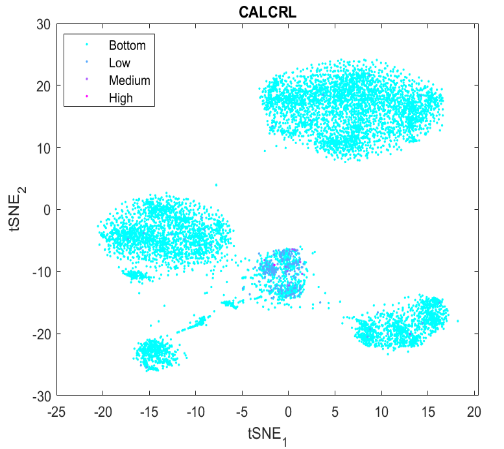
**Results:**

Altogether, the processed scRNA-seq data contained information regarding 8,444 cells with expression values for 20,007 genes. Though 20,007 genes were being tracked in the experiment, the average cell only expressed 1,313 genes with a standard deviation of 676.5 (Fig. 1a). This is a result of the low capture efficiency often present in scRNA-seq experiments and is not unexpected, though we must be cautious when interpreting the results of our data analysis as the data itself may be biased or have a great deal of unwarranted variation.1 Our t-SNE visualization of the data, after PCA was applied, suggests that 8 or 9 clusters exist in our dataset (Fig. 1b). Additionally, when we color code each point in the t-SNE plot according to the patient source, we observe that the potential cluster on the very top right is composed primarily of hepatic cells from patient 3. The homogeneity of that perceived cluster may bias our analysis.

After performing k-means clustering, we obtain 8 distinct clusters which predominantly group along the same lines as the perceived groups in the t-SNE plot, thereby validating our approach (Figure 1d). However, we do observe that clusters 2, 3 and 5 are found within the same perceived group in the t-SNE plot that is composed primarily of hepatic cells from patient 3. Consequently, our analysis of these three clusters may contain biases not representative of the average human liver. We observe the following: cluster 1 is primarily composed of endothelial cell, cholangiocytes and stellate cells; cluster 2, 3 and 5 are primarily composed of hepatocytes; cluster 4 is primarily composed of plasma cells; cluster 6 is primarily composed of erythroid cells; cluster 7 is primarily composed of macrophages; and cluster 8 is primarily composed of B-cells, T-cells and Natural Killer (NK) cells (Fig. 1e).We performed a pathway analysis of the DE genes identified for each cluster to identify which pathways are presumably being upregulated in each cluster using PANTHER. The identified pathways typically correlated with the function of the cell-type each cluster was assigned to.

As cluster 1 and 8 are made up of different cell types, we suspect that there could be subpopulations existing in these clusters. Additionally, previous literature suggests distinct populations of macrophages exist in the liver, so we suspect cluster 7 to have subpopulations as well.5 Color-coding our tSNE visualization of the overall data based on expression of specific cell-type specific gene markers suggests that distinct subpopulations do exist (Fig. 2). For example, we observe that the expression of the genes CALCRL and VWF is localized to different areas of cluster 1 (Figs. 2a-b). Similarly, expression of the genes LYZ and CD5L is localized to different areas of cluster 7 and expression of the genes CD7 and CD3D is localized to different areas of cluster 8 (Figs. 2c-f).

**Figure 2: t-SNE Plot with Each Point Color Coded by Expression of Indicated Gene. a)** CALCRL **b)** VWF **c)** LYZ **d)** CD5L **e)** CD7 **f)** CD3D Legend for relative expression of each marker from lowest expression (bright blue dots) to highest expression (purple dots) (top left).



**a)**

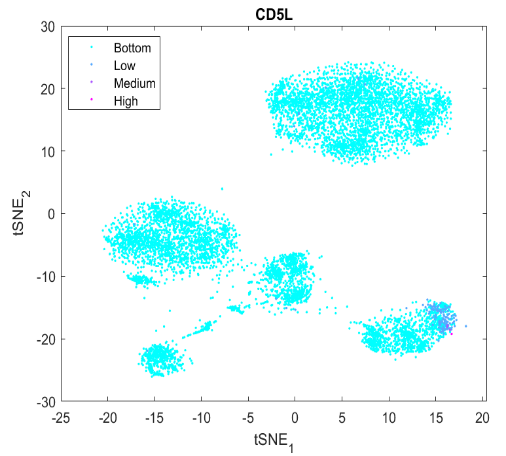
**b)**

**c)**

**d)**

**e)**

**f)**

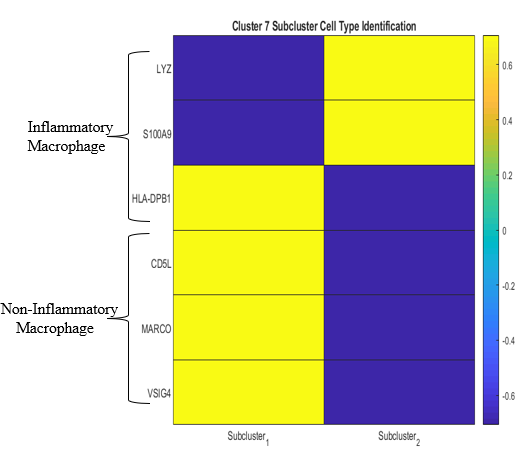
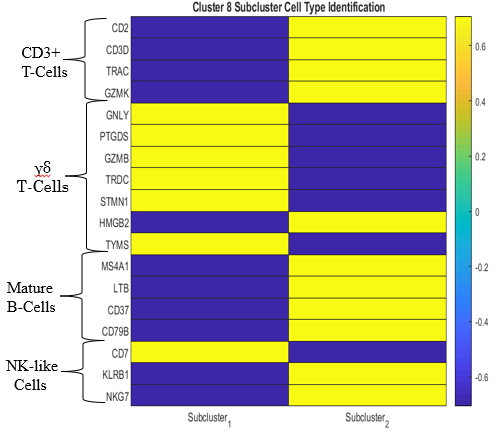
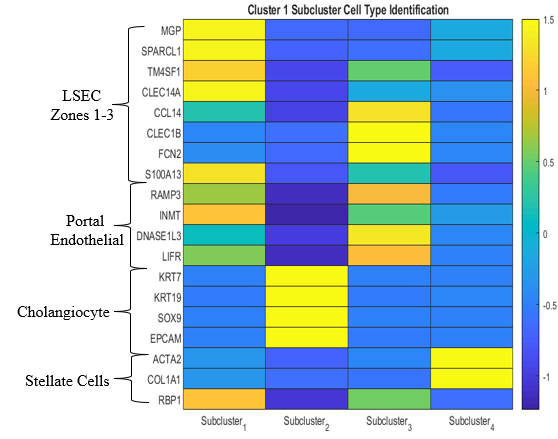
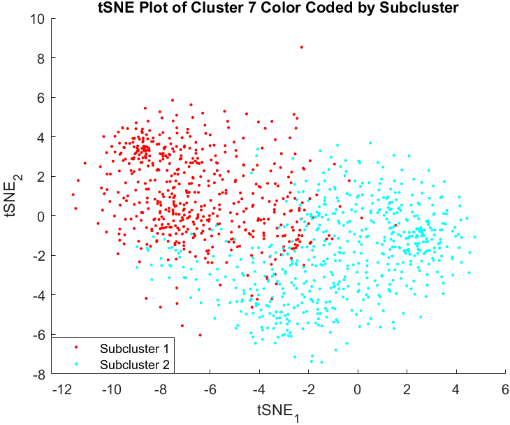
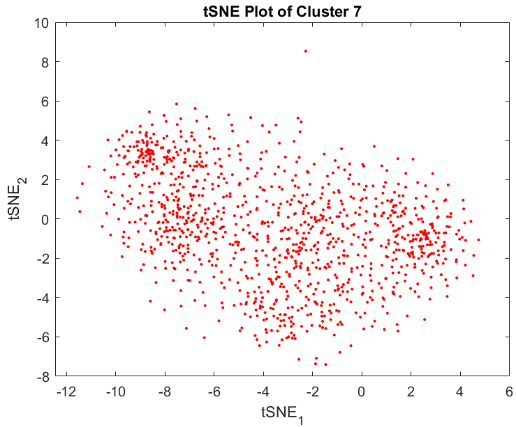
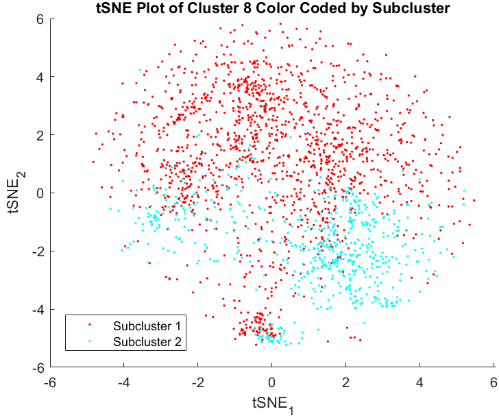
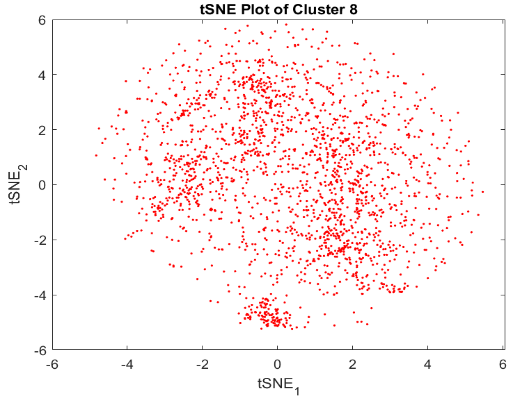
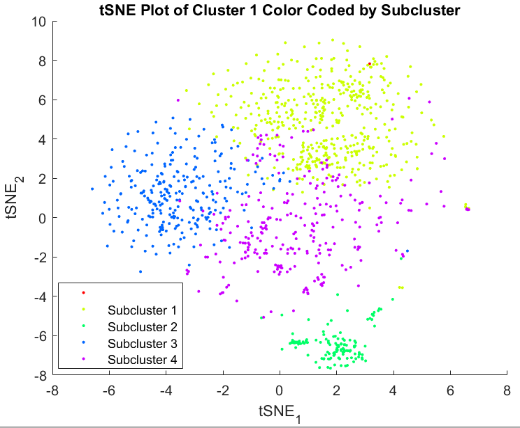
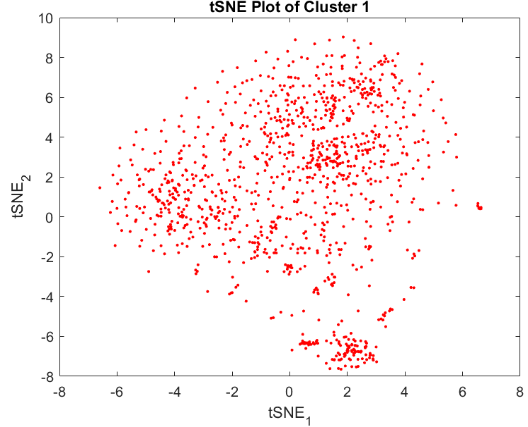


The t-SNE plot for just cluster 1 suggests the presence of 4 clusters; our subsequent k-means clustering output creates subclusters that fall roughly within the visually suggested groupings (Fig. 3a). Cell-type identification of each subcluster suggests that subcluster 2 is the collection of cholangiocytes, epithelial cells which line bile ducts, present in the liver and that subcluster 4 is the collection of stellate cells (Fig. 3a). As a transcriptional profile of human cholangiocytes has been not been described before the collection of this data, the 140 DE genes for subcluster 2 I have identified serve as a novel transcriptional profile for the cell type. Additionally, this subcluster displays upregulated expression of genes encoding secretory and inflammatory pathways such as TFF1 and KRT19. The 43 DE genes identified for subcluster 4 are mostly genes that have been observed to be upregulated during hepatic stellate cell (HSC) activation in human liver, giving further evidence to suggest this cluster is the collection of HSCs.

The t-SNE plot for cluster 7 suggests the presence of 2 subclusters and the k-means output clusters them in a visually agreeable manner; our cell-type identification heatmap suggests subcluster 1 is composed of non-inflammatory macrophages while subcluster 2 is the collection of inflammatory macrophages (Fig. 3b). Despite observing this difference in function from known cell-type specific gene markers, pathway analysis of identified DE genes suggests that the functional difference between the 2 subpopulations stems from something other than the inflammatory or non-inflammatory properties of a macrophage, as the implicated pathways do not seem to serve any role in an immune response (Fig. 3b).

The t-SNE plot for cluster 8 suggests the presence of at least 2 subclusters as there appears to be a distinct population towards the x-axis of the plot, however, our k-means output clusters our data very differently (Fig. 3c). Our cell-type identification heatmap suggests subcluster 1 is composed of γδ T-cells while subcluster 2 is the collection of CD3+ T-cells, Mature B-cells and NK-like cells and suggests that these clusters, while unexpected, are biologically significant (Fig. 3c). Pathway analysis of identified DE genes from subcluster 1 suggest that subcluster has upregulated expression of genes involved in the p53 pathway suggesting that γδ T-cells play a role in detecting or battling cancer. Similar analysis of DE genes from subcluster 2 reveal that these cells upregulate expression of genes involved in inflammation mediated by chemokine and cytokine signaling as expected.

I then compared the gene expression patterns in the three identified human hepatocyte clusters with the gene expression patterns previously shown in mouse layers (Fig. 4). Human cluster 2 was correlated with mouse layers 4, 5 and 6, which are interzonal in nature (Fig. 4).



**a)**

**b)**

**c)**

**Figure 3. Subcluster Identification for Cluster 1,7 and 8. a)** t-SNE projection of 1136 liver cells in cluster 1 with each point representing a single cell; t-SNE projection of 1136 cells with each point color coded according to subcluster identity assigned by k-means clustering algorithm; Heat map analysis assigning the identity of each subcluster was done by matching the subcluster expression profile with established cell-specific marker gene expression for endothelial cells, cholangiocytes, and stellate cells. Assignment of color to each cell of the heatmap was determine by z-score value as indicated by the legend **b)** Subcluster identification for cluster 7 follows the exact same procedure as that documented in Fig. 3a but we begin with the 1125 liver cells in cluster 7 ; Heat map analysis was performed with established cell-specific marker gene expression for macrophages. **c)** Subcluster identification for cluster 8 follows the exact same procedure as that documented in Fig. 3a but we begin with the 2096 liver cells in cluster 8; Heat map analysis was performed with established cell-specific marker gene expression for immune cells.

****

**Figure 4: Human Hepatocyte Cluster and Mice Hepatocyte Layer Correlation.** Using the average values of nine layers of mouse liver cells provided by Halpern et al.3,85 significantly DE genes were selected for correlation. Expression values of each gene for the three human hepatocyte clusters and nine layers of mouse liver cells were normalized by z-scores. Pearson Correlation Coefficients were calculated using the z-score to determine relatedness of human clusters to mice layers. Assignment of color to each cell of the heatmap was determined by the Pearson Correlation Coefficient value as indicated by the legend

Cluster 3 correlated well with mouse layer 1, the most central venous mouse liver layer responsible for drug metabolism in mice (Fig. 4). Cluster 5 did not appear to correlate strongly with any mouse layers, most likely because the DE genes defining this cluster were not members of the 85 genes defining zonation in mouse livers (Fig. 4). Only the positive correlation of cluster 2 with layers 4, 5 and 6 was deemed to be statistically significant according to our p-values (0.006, 0.005 and 0.0466 respectively).

**Discussion:**

While my analysis identified 8 unique clusters, analysis performed by MacParland et al. on the same data yielded 20 unique clusters. The 20 clusters they identified were better partitioned according to cell type; while my clusters 1 and 8 were clearly composed of several cell types, each cluster identified by MacParland et al.5 corresponded to a single cell type and some clusters corresponded to the same cell type indicating the presence of distinct subpopulations of the cell type in the liver. In short, their analysis was much more powerful than mine at identifying unique cell populations from the same data. One potential reason for this may be our use of PCA to reduce the dimensionality of the data. As our data contains 20,007 variables for just 8,444 observations, using PCA is not recommended.1Alternative techniques such as sparse PCA may be better suited for our dataset and may improve the robustness of our analysis.

As a consequence of our vastly different clustering results, our downstream analysis results are quite different as well. Many of the DE genes identified in my analysis were not identified as such by MacParland et al.5 and vice-versa. This also resulted in the identification of starkly different upregulated pathways for presumably the same population. For example, MacParland et al.5 functionally categorize the non-inflammatory macrophage population found in both of our analysis as having a tolerogenic function in the immune system, while my analysis suggests it is somehow involved with dealing with cancer and other diseases.

One major issue in my analysis was in identifying differentially expressed genes. Due to the low capture rate present in scRNA-seq experiments, most expression values in a given cell will be 0.1 This makes identifying differentially expressed genes through a fold change criterion difficult; if a cluster has an expression value of 0.000001 for a gene, it can still be considered differentially expressed if the other clusters have an expression value of 0.1 Additionally, my fold change method was extremely poor at identifying genes differentially expressed lower in a cluster. For example, my algorithm identified 9,739 genes that cluster 2 expressed five-fold lower than the other clusters; it is difficult to believe that all these genes are truly differentially expressed and worth analyzing.

There are many avenues through which we can continue exploring our data. An obvious one would be to create a more robust algorithm for DE gene identification through implementation of more advanced techniques like the rank product method.1 Additionally, I identified many pathways that were seemingly upregulated, but my lack of familiarity with liver cell types made it difficult to validate the biological relevance of these pathways in the context of liver functioning. Finally, single-molecule RNA FISH experiments should be performed to confirm the validity of the DE genes I identified for each cluster.

**References:**

1. Chengzhong Ye, Terence P Speed, Agus Salim, DECENT: differential expression with

Capture efficiency adjustment for single-cell RNA-seq data, Bioinformatics (2019).

2. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration

of genomic datasets with the R/Bioconductor package biomart. *Nat. Protoc.* **4**, 1184–

1191 (2009).

3. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the

mammalian liver. *Nature* **542**, 352–356 (2017).

4. Jakab, L. "The liver and the immune system." *Orvosi hetilap* 156, no. 30, 1203-1213 (2015)

5. MacParland, S.A., Liu, J.C., Ma, X. et al. Single cell RNA sequencing of human liver reveals

distinct intrahepatic macrophage populations. *Nat Commun* **9**, 4383 (2018).

6. Mi, H., Muruganujan, A., Huang, X. et al. Protocol Update for large-scale genome and gene

function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703–

721 (2019).

**Supplementary Figures and Results:**

All code used for analysis can be found in <https://github.com/pree1199/BIOC470_Final>.

The genes that were identified to be differentially expressed for each cluster can be found in Differentially\_Exppressed\_Genes.xlsx. The folder Implicated\_Pathways contains a series of .txt files containing the identified pathways of interest for each cluster. The folders Clust1\_Analysis, Clust7\_Analysis and Clust8\_Analysis contain the data from the subpopulation analysis that was performed.