Identifying Rare Subpopulations in scRNAseq Data with Missing Expression of Marker Genes

Abhinav Bichal[1], Emil Velasquez[1], Samarth Bhat[1], Ritu Gupta[1], Rohit K. Prasad[1], Dhivya Arasappan[1], Jeanne Kowalski-Muegge[2,3]

[1]Center for Biomedical Research Support, Office of the Vice President for Research; [2]Department of Oncology, Dell Medical School; [3]Live**STRONG** Cancer Institutes, Dell Medical School
The University of Texas at Austin

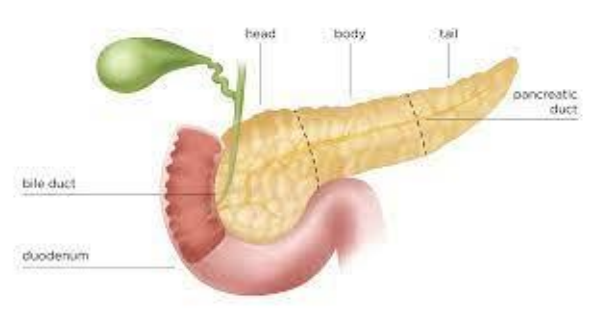
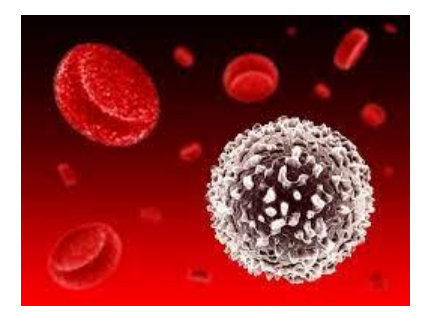
Introduction: Identifying Rare Cell Subpopulations Using Single Cell RNA-seq

In order to identify rare cell subpopulations, RNA is often profiled at a single cell level using single cell RNA seq (scRNA seq). The cells found through scRNA seq are then clustered based on key marker genes. However, scRNA seq is limited by high dropout rates, where genes are not detected in large number of cells.

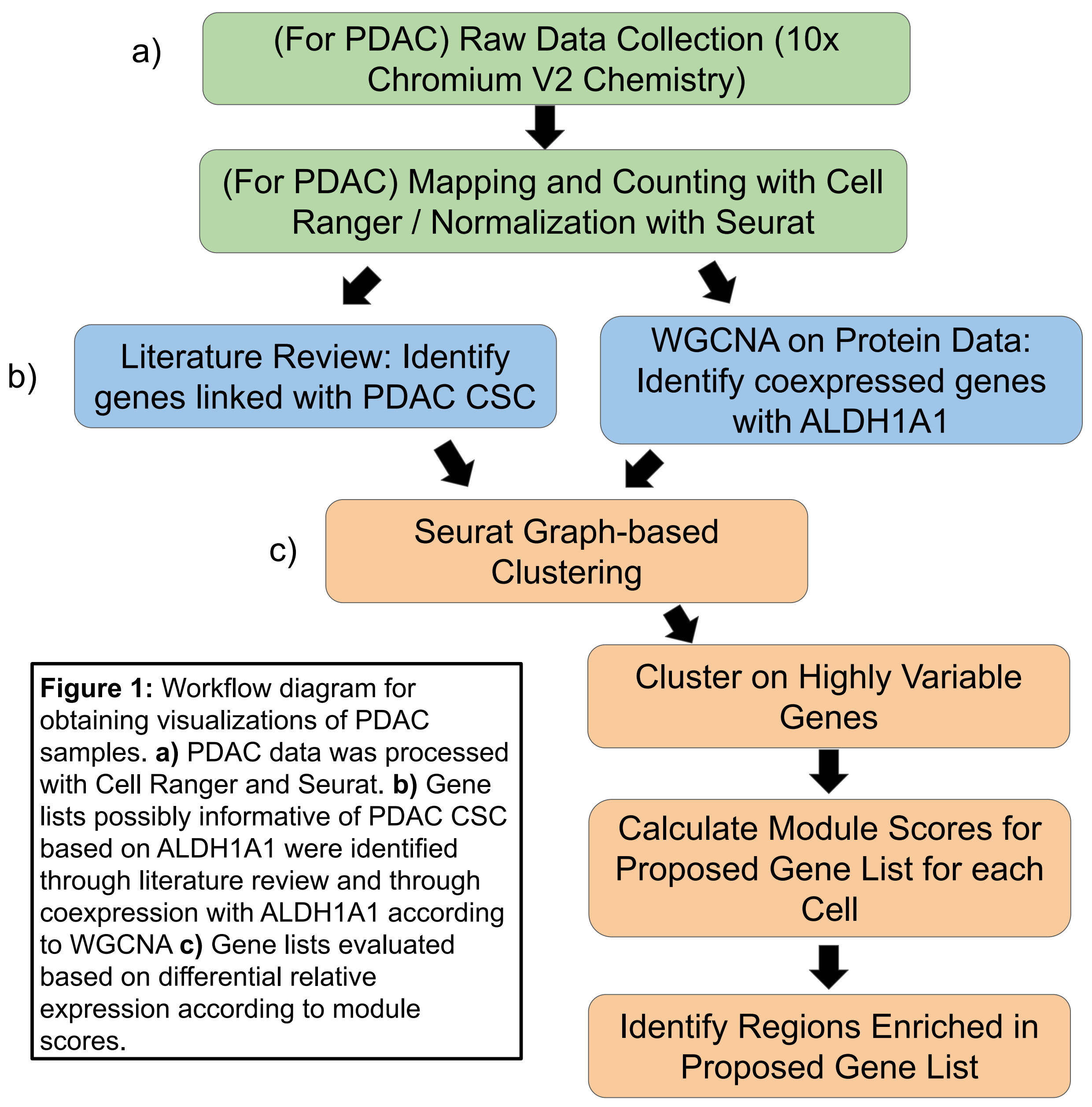
Problem: How can we identify rare cell subpopulations when the marker genes for the cell type are not detected in scRNA seq data?

ALDH1A1, a marker gene for a rare subpopulation of cancer stem cells in pancreatic cancer (PDAC), is not detected in our scRNA-Seq data. We propose using genes coexpressed with ALDH1A1 (coexpression based markers) and genes related to ALDH1A1 in literature (literature based markers) to identify the subpopulation of interest.

Data:

Results: PDAC 	scRNA-seq: 1 sample / 2432 cells
	Protein: 179 samples
Validation: PBMC 	scRNA-seq: 1 sample / 11591 cells
	Bulk: 7 Healthy Human PBMC Samples

Methods:



a) (For PDAC) Raw Data Collection (10x Chromium V2 Chemistry)

(For PDAC) Mapping and Counting with Cell Ranger / Normalization with Seurat

b) Literature Review: Identify genes linked with PDAC CSC

WGCNA on Protein Data: Identify coexpressed genes with ALDH1A1

c) Seurat Graph-based Clustering

Cluster on Highly Variable Genes

Calculate Module Scores for Proposed Gene List for each Cell

Identify Regions Enriched in Proposed Gene List

Figure 1: Workflow diagram for obtaining visualizations of PDAC samples. **a)** PDAC data was processed with Cell Ranger and Seurat. **b)** Gene lists possibly informative of PDAC CSC based on ALDH1A1 were identified through literature review and through coexpression with ALDH1A1 according to WGCNA **c)** Gene lists evaluated based on differential relative expression according to module scores.

ALDH1A1 Marker Dropout in PDAC Sample

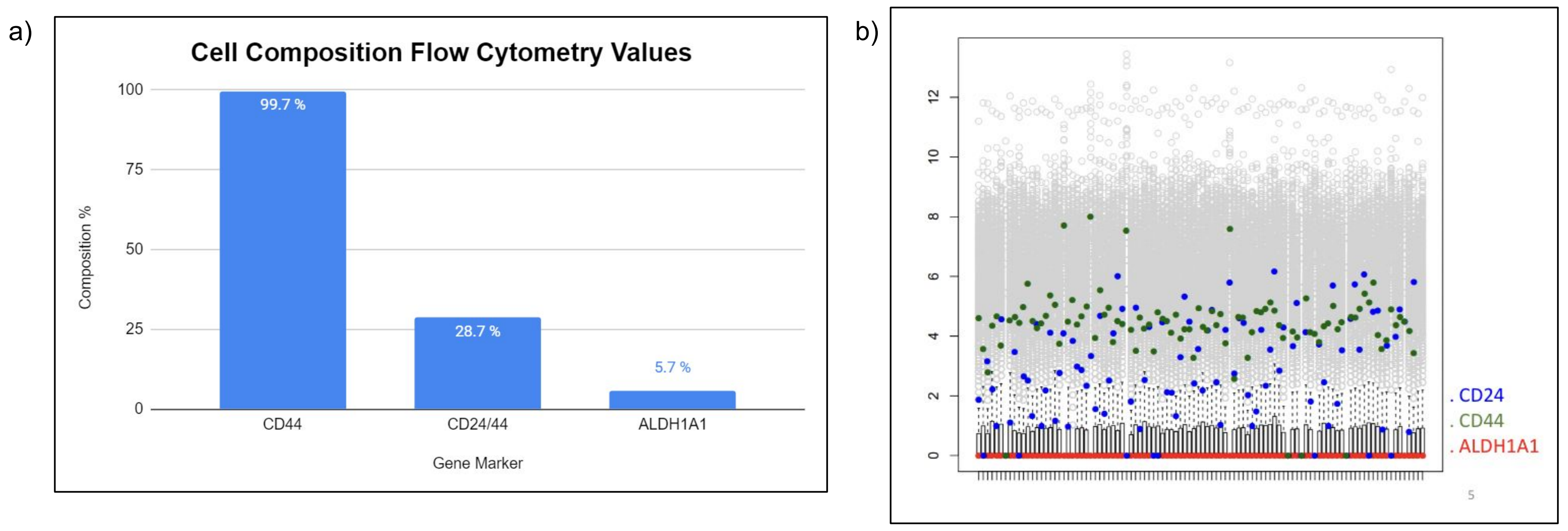


Figure 2: a) Bar graph showing expected composition of scRNA-seq data according to flow cytometry. ALDH1A1+ cells make up 5.7% of the sample. **b)** Measured expression of PDAC CSC markers. ALDH1A1 was not expressed in the sample.

Ideal Marker Genes Would Separate Out Cell Subpopulation of Interest

a) Ideal Clustering Pattern Based on Gene List

b) Ideal Module Score Pattern Based on Gene List

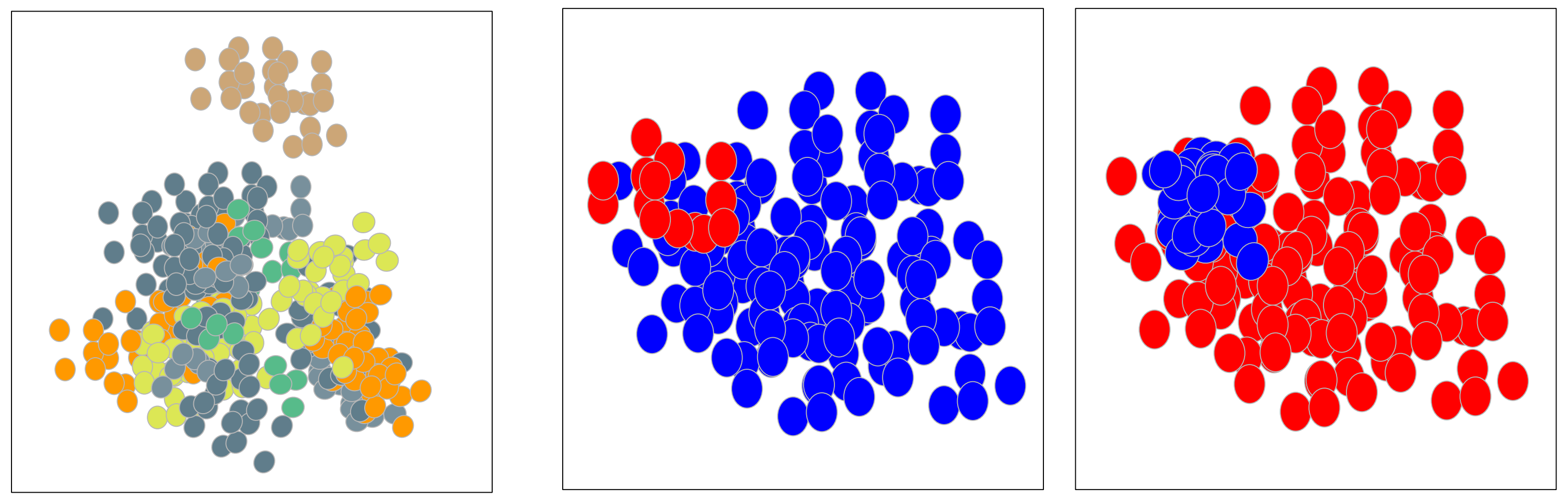


Figure 3: a) Ideal result when using the proposed gene list as the features for clustering. Here, a cluster of the expected proportion separates itself from the rest of the sample. **b)** Ideal result when using module score to visualize relative expression of the proposed gene list. Here, a cluster of cells of the correct proportion over express (red) or under express (blue) the gene list.

Clustering the PDAC Cells Using Highly Variable Genes Identifies Cluster 6 as a Cluster of Interest

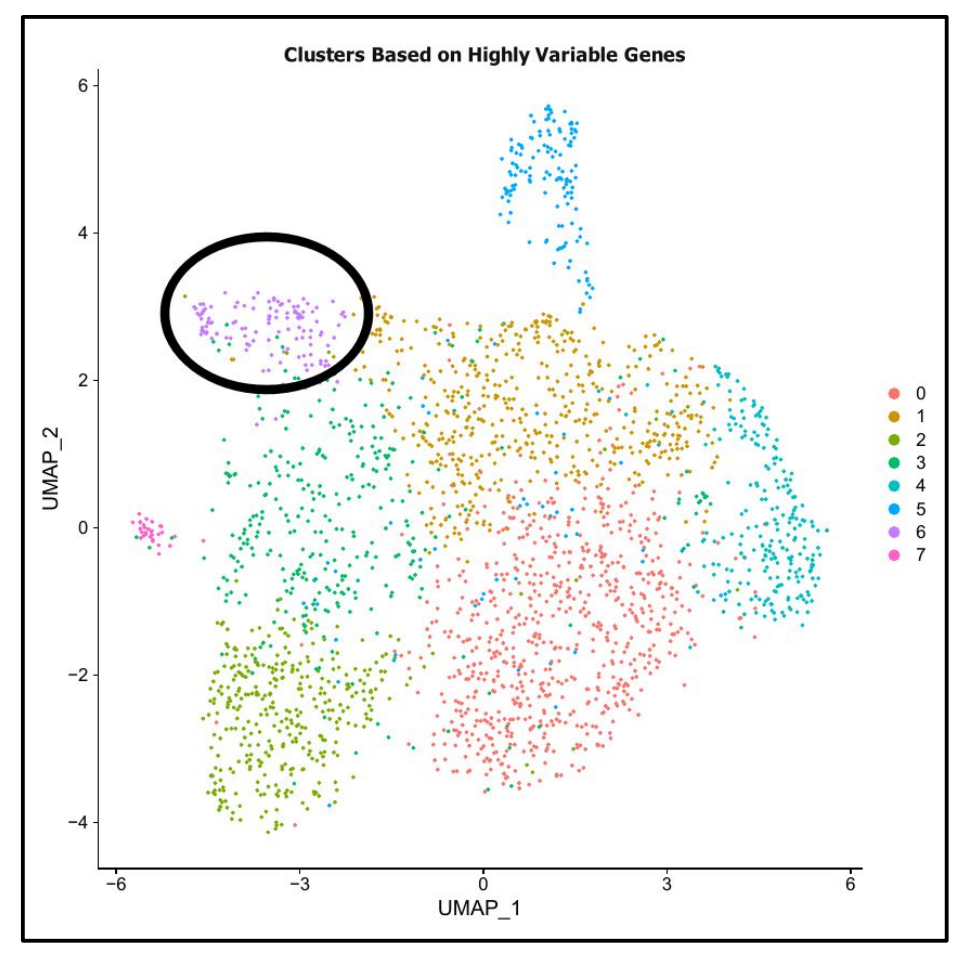
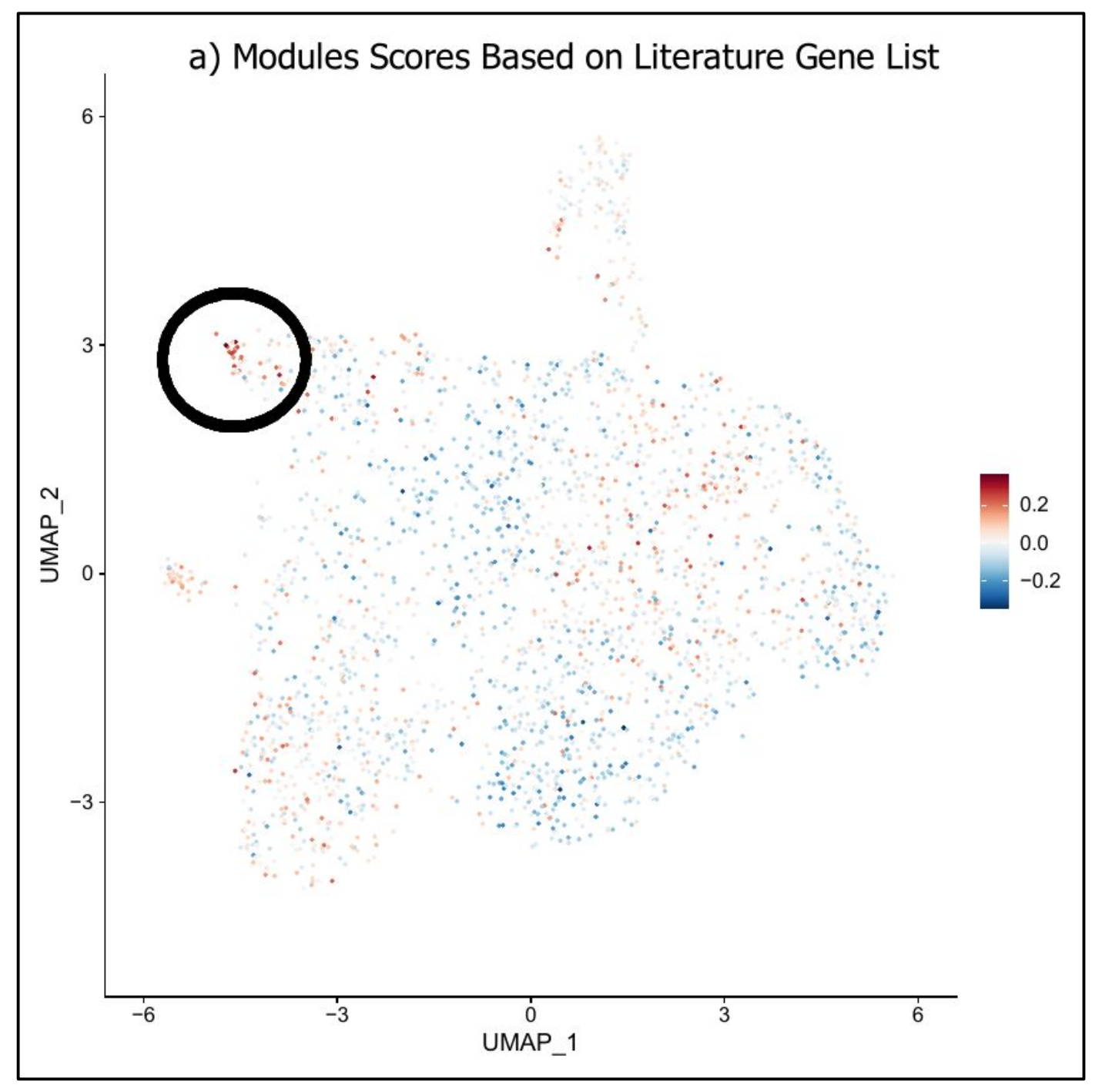


Figure 4: Clusters formed when clustering according to the top 3000 highly variable genes. Circled is cluster 6, which shows interesting results when analyzing the literature and the WGCNA gene lists. Cluster 6 consists of 124 out of the 2432 cells in the sample, which is approximately 5.1% of the sample.

Literature Based Markers are Overexpressed in a Cluster of Interest



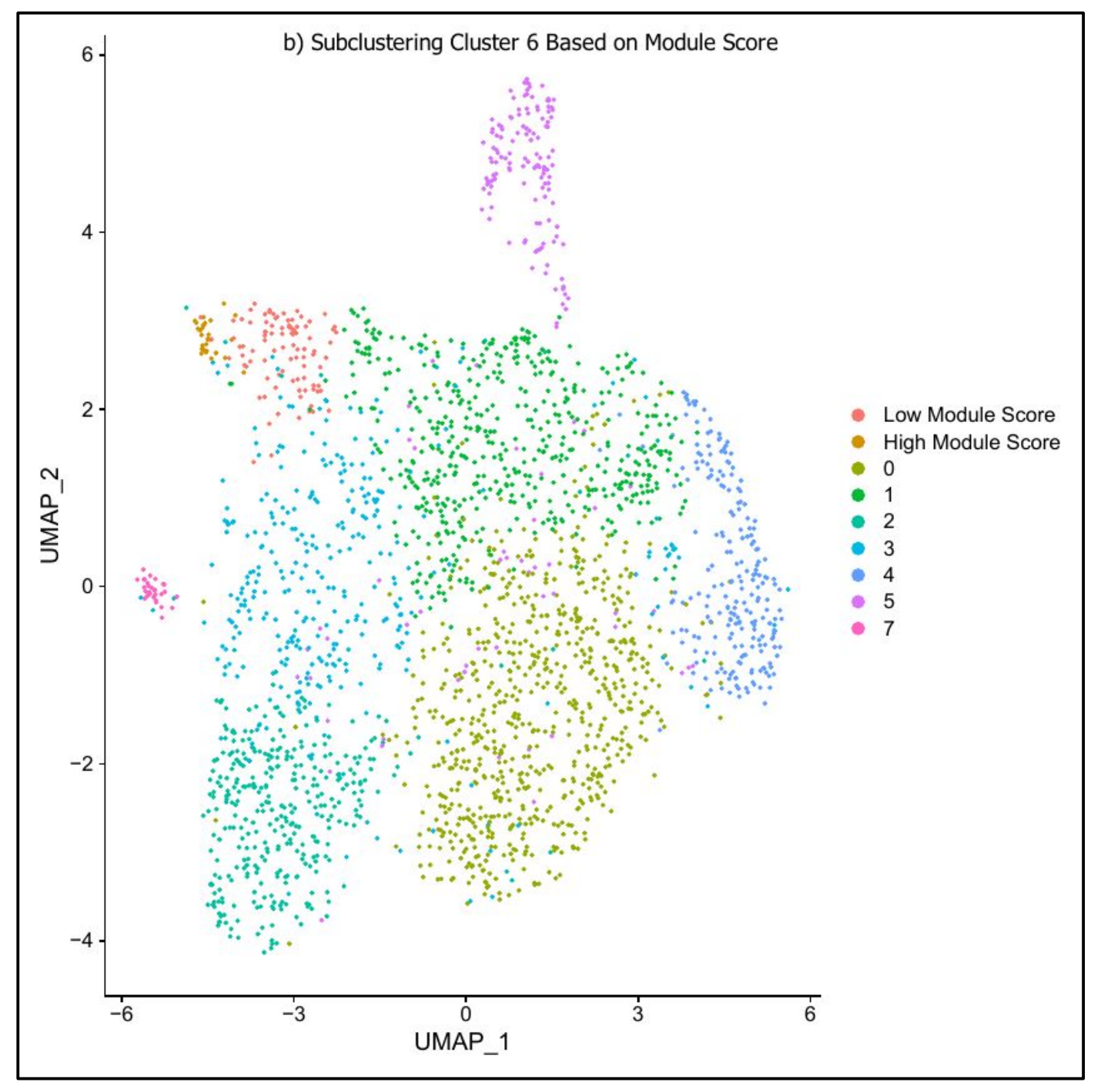


Figure 5: a) Module score visualization using the literature based genes. Circled is a region of cells that highly overexpressed the gene list compared to the rest of the population. **b)** Subclustering cluster 6 from the clusters made from the top variable genes identifies the cells in the circled region in (a). This region only had 31 cells, which is around 1.3% of the sample.

Coexpression Based Markers are Underexpressed in Cluster 6

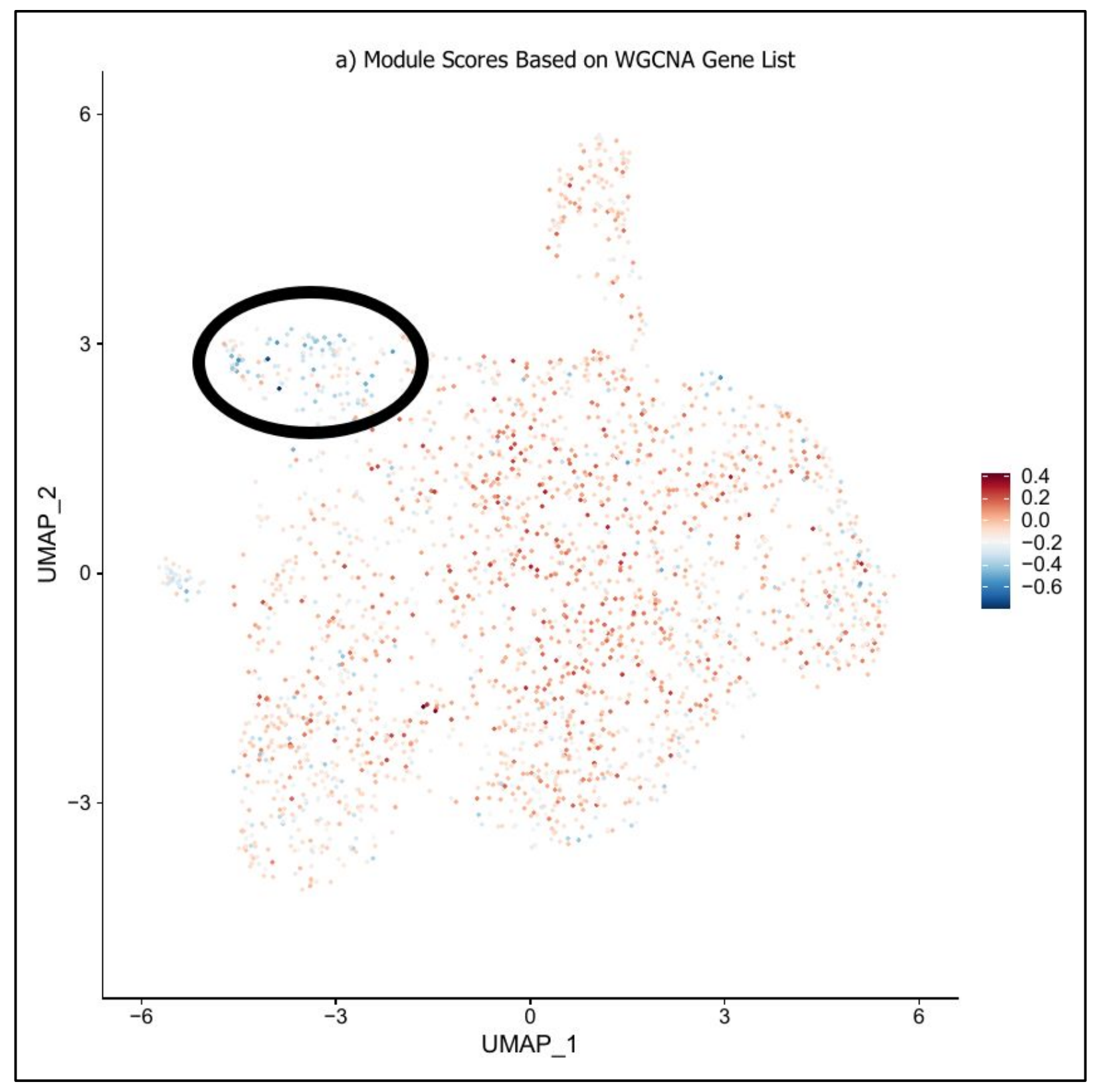
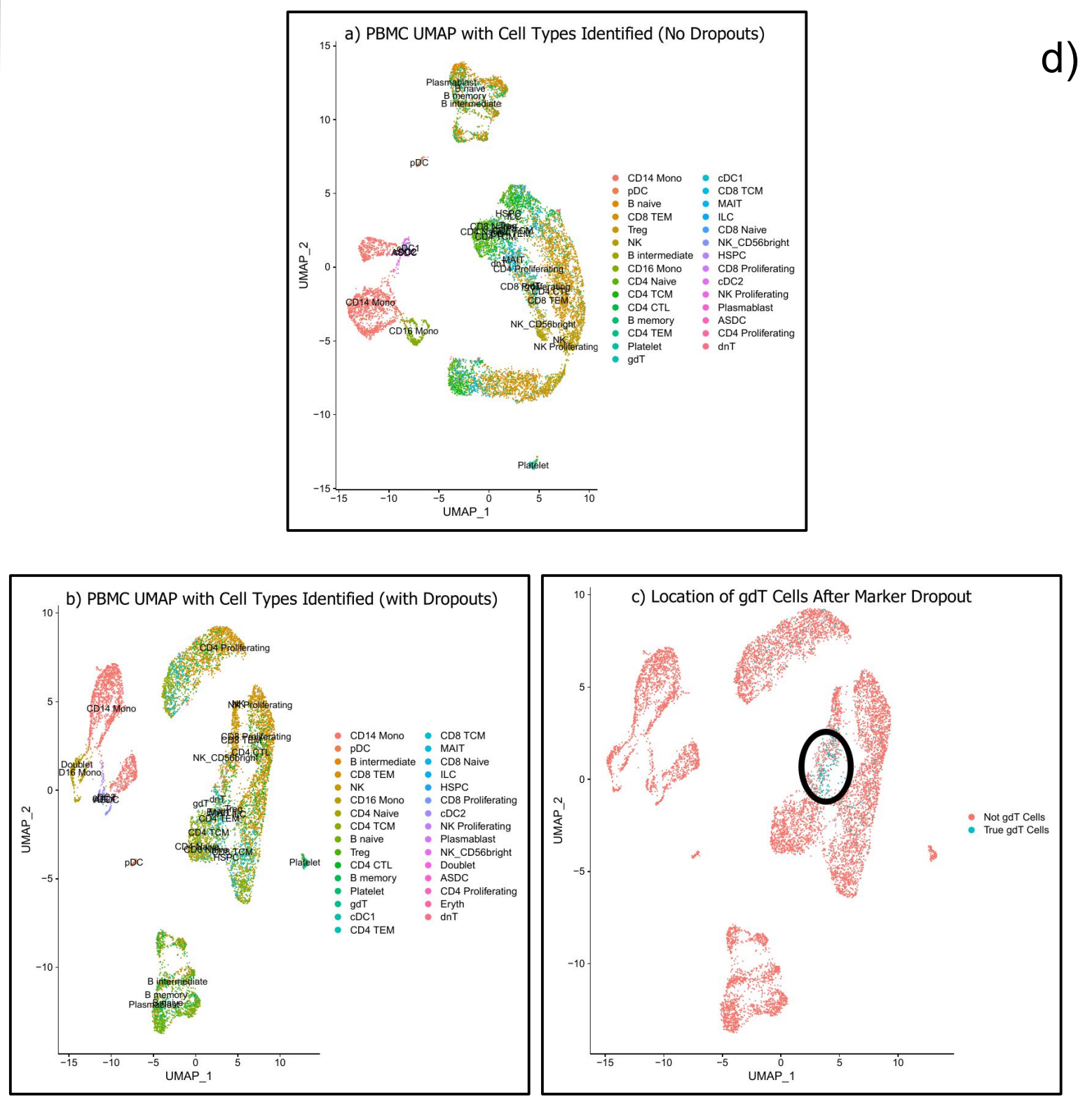


Figure 6: a) Module score visualization identifies cluster 6 (circled region) from the clusters according to highly variable genes to be relatively underexpressed in the coexpression gene list compared to the rest of the sample.

Validation: Establishing Ground Truth in PBMC scRNA-seq data by Targeting gdT Cells



d)

Marker Gene	Module Color
TRDC	NULL
TRGC1	NULL
TRGC2	NULL
CTSW	darkorange
CD3D	darkgrey
KLRG1	cyan
KLRB1	green
KLRD1	cyan
KLRC1	green
IL32	coral4
CCL5	coral3
HCST	sienna3
NKG7	cyan
GNLY	blue
CST7	cyan
LYAR	brown
HOPX	coral4
GZMA	cyan

Figure 6: a) True cell type clusterings in PBMC data without dropping out gdT marker genes. 373 gdT cells were identified. **b)** Cell type clusterings after dropping top 18 gdT markers. 25 gdT cells were identified. **c)** Circled is where the true gdT cells are after dropping out markers. **d)** Modules each dropped gdT marker was assigned to. Cyan had 5 of the markers. 9 modules were identified total. 3 genes weren't assigned a module.

Cyan Module Genes are Overexpressed in Region Including gdT cells

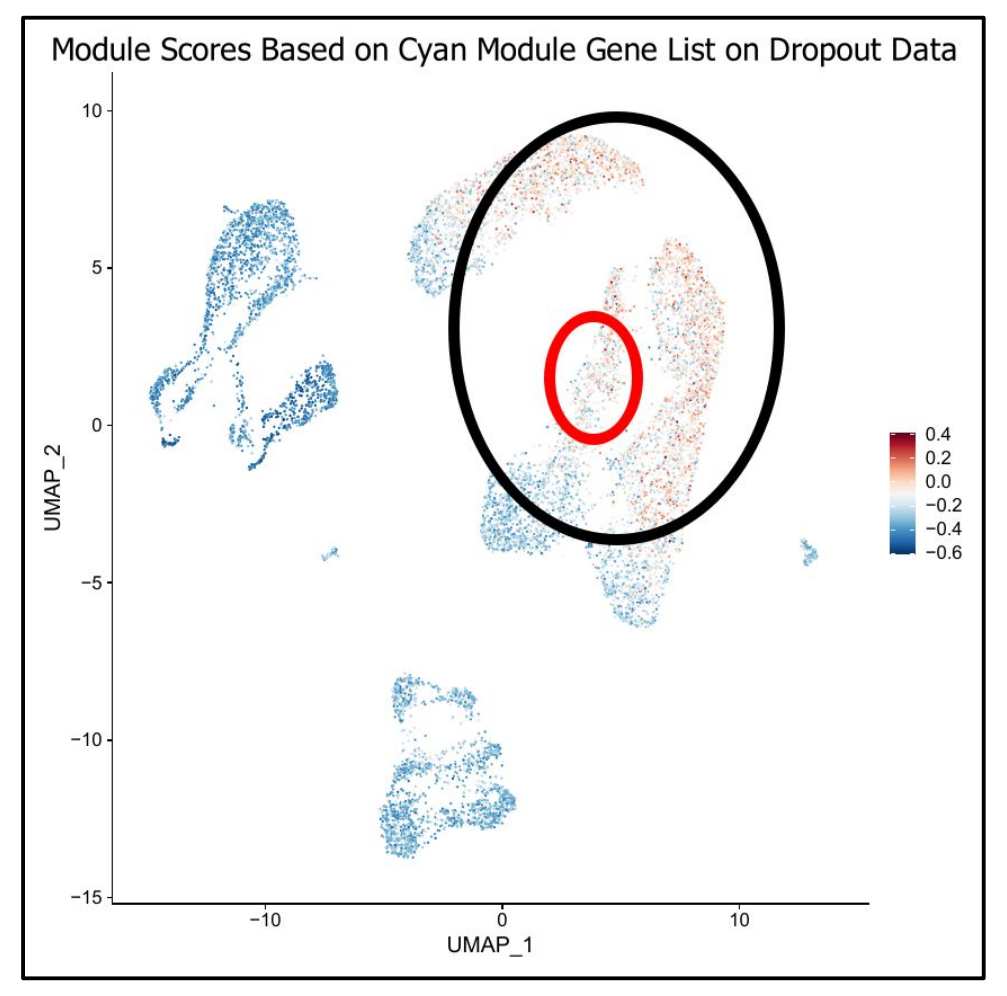


Figure 8: Module score figure shows a large region of cells (black circle) that overexpress the gene list. The region includes but is not limited to the true gdT cells (red circle).

Darkorange Module Genes are not Differentially Expressed in gdT Cell Region

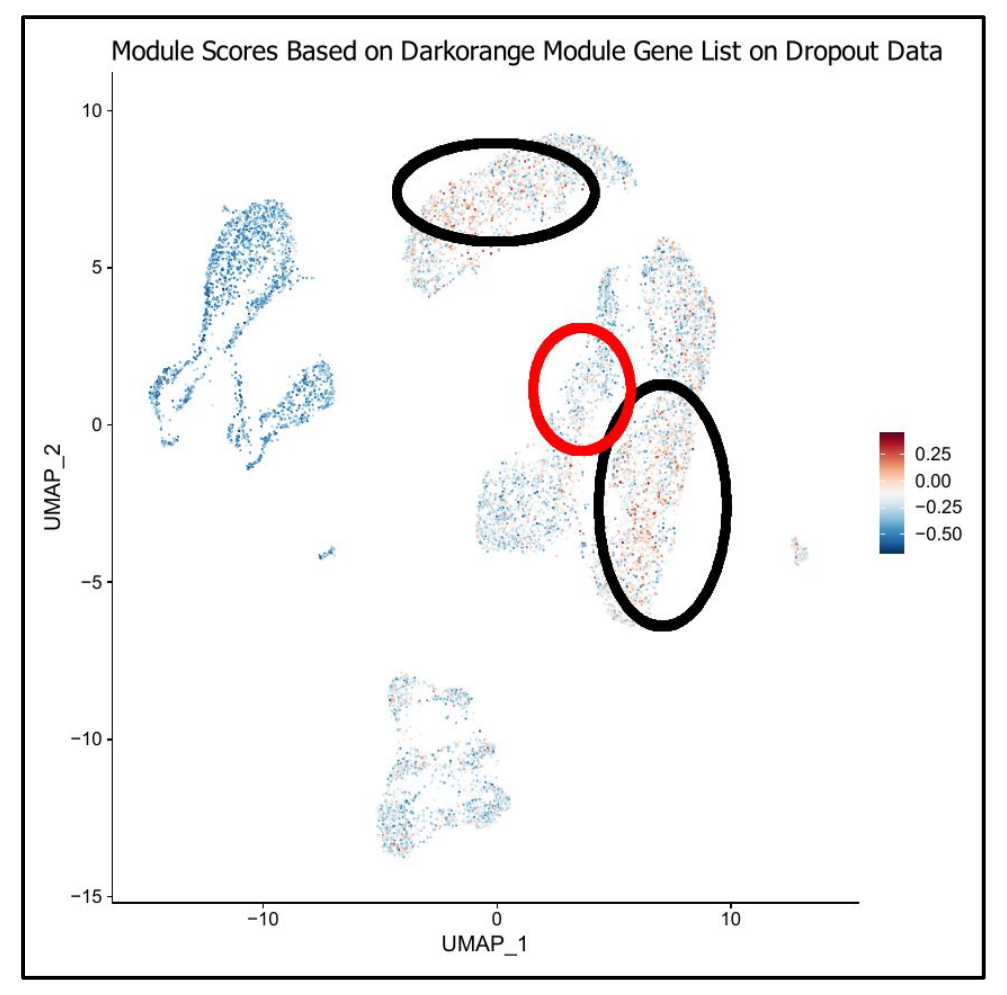


Figure 9: Module score visualization shows a region (black circle) that over expresses the gene list. The region does not include the true gdT cells (red circle).

Discussion:

- WGCNA Module Score analysis identified a roughly 5% cell population in the PDAC sample that could be ALDH1A1+
- However, preliminary attempts to validate using this method suggest that this method may lack the specificity required to identify rare subpopulations.

Future Steps:

- Run validation on the other identified modules
- Gather more data - Validation, WGCNA
- Consider methods of combining multiple gene lists/modules

Acknowledgements:

We would like to thank the Big Data in Biology FRI Stream for their guidance and support. This work was supported by the TIDES FRI Summer Research Fellowship.

References: