Jennifer Alcantara-Castillo, Soon Hoo Lee, Amatya Pathak, Janika Peyasena, Nivedha Satyamoorthi
December 19, 2021
BMIN-GA 1001 Bioinformatics: Topic 7

Assignment 6: Investigating Transcriptional Heterogeneity of Cell Types in AML Using scRNA-Seq

Introduction:

The goal of this assignment is similar to those of Assignment 2 and 3, where datasets containing RNA sequencing (RNA-Seq) data were aligned to the transcriptome, visualized, and analyzed via differential gene expression (DGE). RNA-Seq is a straightforward technique that yields an average gene expression pattern across millions of differently specialized sample cells. However, this sequencing method is best when cell heterogeneity is not a variable, and when we want to differentiate between samples with and without certain biomarkers irrespective of the cell type.

The real-world application of the analysis is to identify the transcriptome differences between healthy bone marrow samples and those with acute myeloid leukemia (AML). It is necessary to differentiate between the primary types of blood cells (lymphocytes and myelocytes) and ideally, the types of myelocytes as well (like neutrophils, monocytes, and macrophages) since AML is caused by their mutated versions uncontrollably dividing and reproducing. Therefore, to capture transcriptome differences down to the individual cell type and even down to individual cell level, single cell RNA sequencing (scRNA-Seq) is the better option.

Now, scRNA-Seq is not without its disadvantages. Attempting to sequence RNA in every sample cell would generate a much bigger volume of data and thus a lengthier analysis process when running computational based analyses on the data. Furthermore, existing droplet-based methods only capture about 50% of the transcriptome per cell in the best case, meaning in analysis, if a gene expression count reads zero, that could either mean no expression or no detection to begin with. (Kirchner 2020). Of course, this makes scRNA-Seq more prone to biological and technical variability that traditional RNA-Seq partially mitigates with a composite gene expression profile. This is solvable by individually sequencing a larger sample of cells. Though this implies more computational space and time requirements, when analyzing minute transcriptomes transcriptome where blood cell heterogeneity matters, scRNA-Seq is the best way to understand AML's effects to develop better treatment plans.

Methods:

This assignment sources its raw scRNA-Seq bone marrow sample data from the 2019 Van Galen study, and so the first step is to download the necessary data from the Gene Expression Omnibus (GEO) database, with accession number GSE116256 (FTP protocol). The expression matrix and metadata, both .txt files, were downloaded for each sample, of which there are eight total. Four bone marrow samples are healthy (BM3, BM4, BM5-34p, and BM3-34p38n), and four samples are from AML patients (AML419A-D0, AML707B-D0, AML916, and AML921A-D0). The expression matrices and corresponding metadata were all

imported into an RStudio working environment, and a package called Seurat was used to perform data normalization, QC, merging, and clustering. This was done by creating eight individual Seurat objects in R by adding the annotation dataframe to the counts matrix for each of the eight samples. The count matrices were log-normalized, and features unique to each sample based on variance stabilizing transformation were identified. Using these features as anchors, we integrated the individual Seurat objects into a single, large Seurat object.

The results were visualized in three UMAP plots shown in Figure 1, depicting nFeature_RNA (number of genes detected in a cell), nCount_RNA (number of RNA molecules detected in a cell), and pct_mito (percentage of reads that map to a mitochondrial genome). High quality data should have high values in the first two metrics, and a low value in the third.

The next phase is to perform DGE analysis on the scRNA-Seq data. Using the findMarkers() function of Seurat, DGE matrices were generated to compare the clustered data from the previous phase. Matrices were also made to compare the cell types within healthy and AML samples, and their malignance statuses. Finally, the most expressed genes found in DGE analysis were visualized in a heatmap.

Results

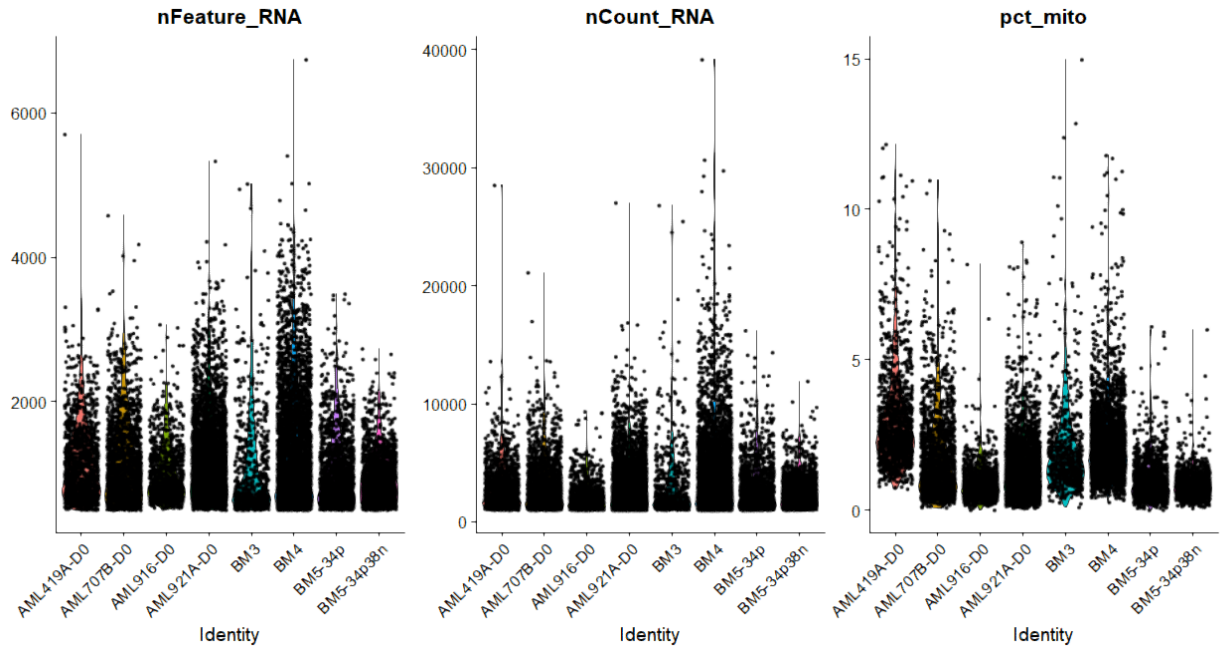


Figure 1. QC metrics for the RNA assay (unmodified counts matrices). *(Left)* number of unique genes. *(Center)* read count number. *(Right)* Percent mitochondrial transcript.

The standard log-normalized analysis of the integrated Seurat object, includes dimensional reduction and clustering analyses. We first scaled the integrated counts matrix

linearly, then performed linear dimensional reduction by PCA. Then, we calculated the K-nearest neighbor clusters based on the euclidean distance in the PCA space. Using these PCs as input, we performed the nonlinear UMAP dimensional reduction. We visualized the clustering of single cells by HSC-score, cell-type, patient (sample origin), and malignance status in the UMAP plot (Figure 2).

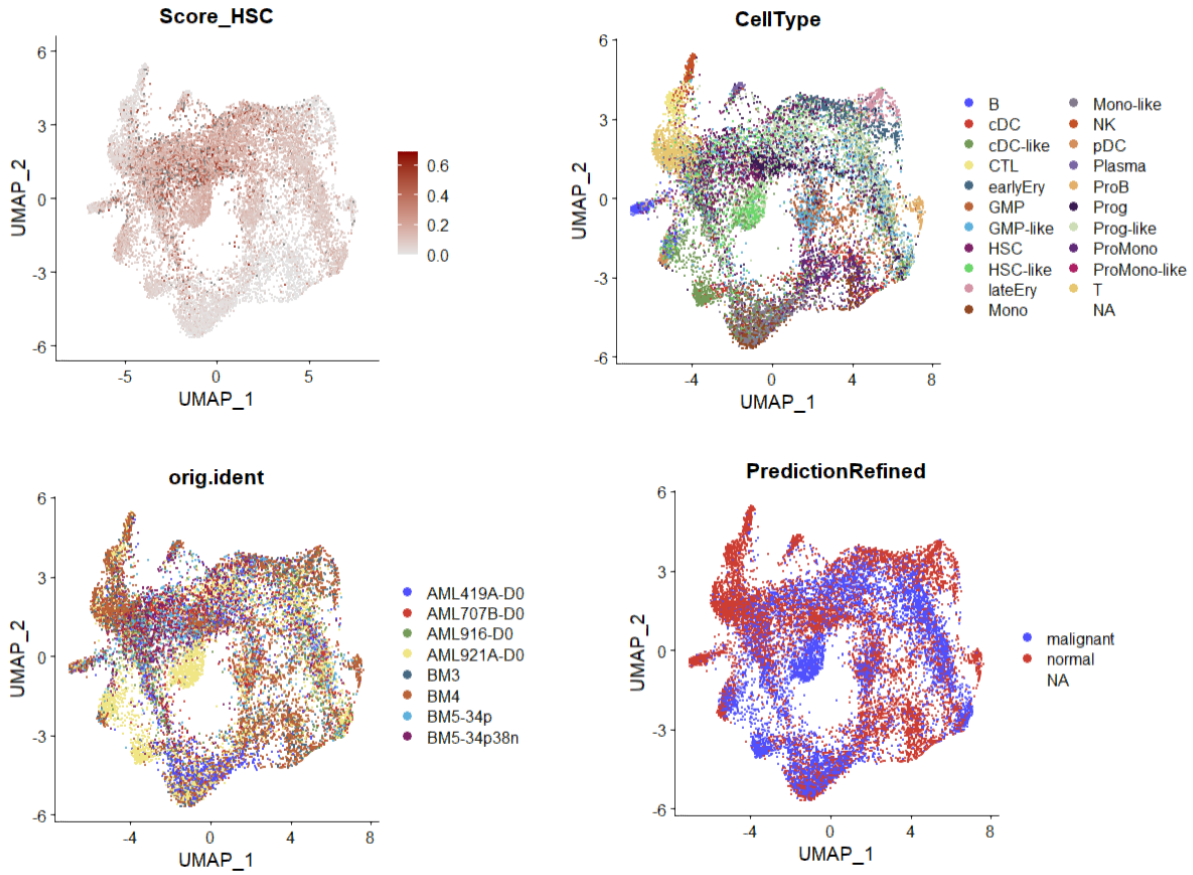


Figure 2. UMAP plot of log-normalized counts. Score_HSC = Hematopoietic Stem cell-like score. CellType = cell type. orig.ident = patient (sample origin). PredictionRefined = predicted malignance status.

Unexpectedly, log-normalization led to indistinguishable clusters for all four comparisons. We are still not sure why this occurred. Hafemeister and Satija describe possible pitfalls of log-normalization, including the addition of pseudocounts and log-transformation (Hafemeister and Satija, 2019). They propose an alternative modeling framework which addresses these issues, which they call SCTransform. Using SCTransform, we generated the same UMAP plots (Figure 3).

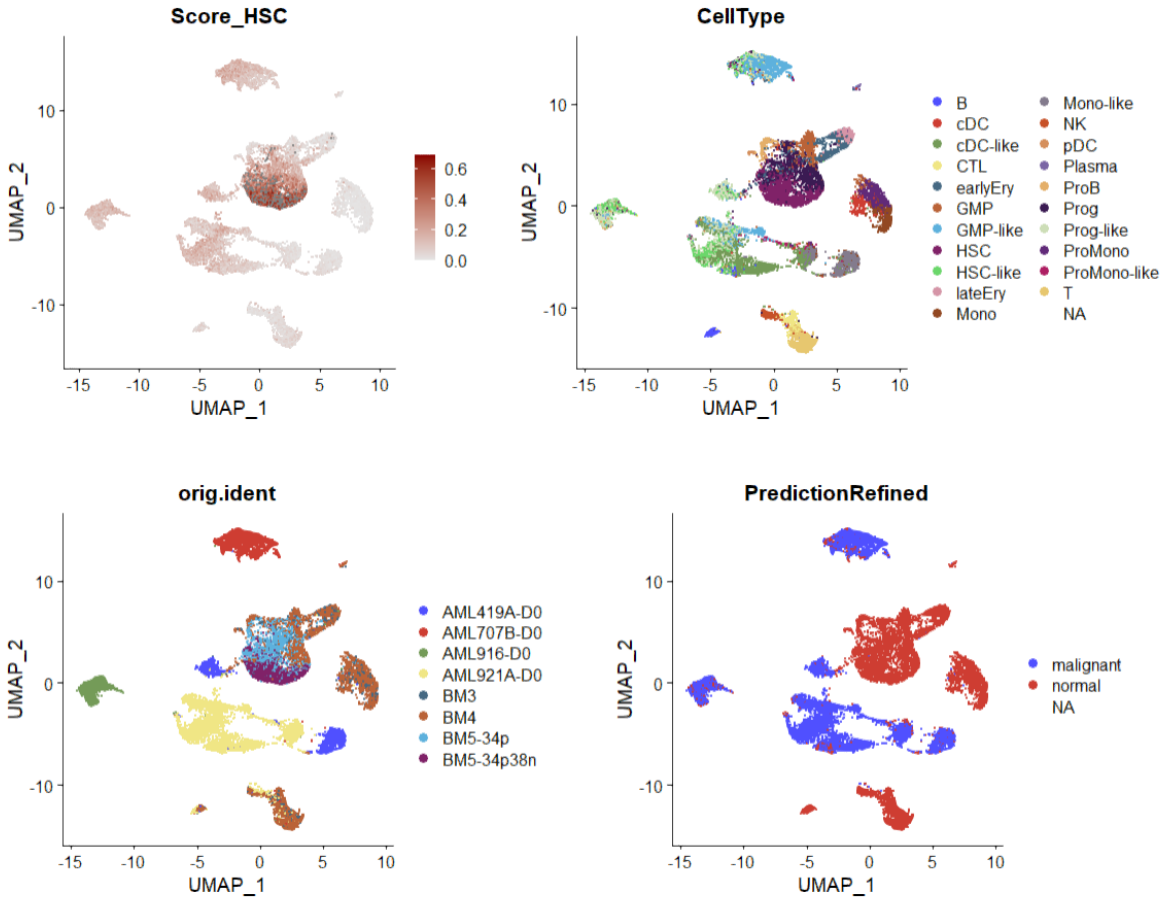
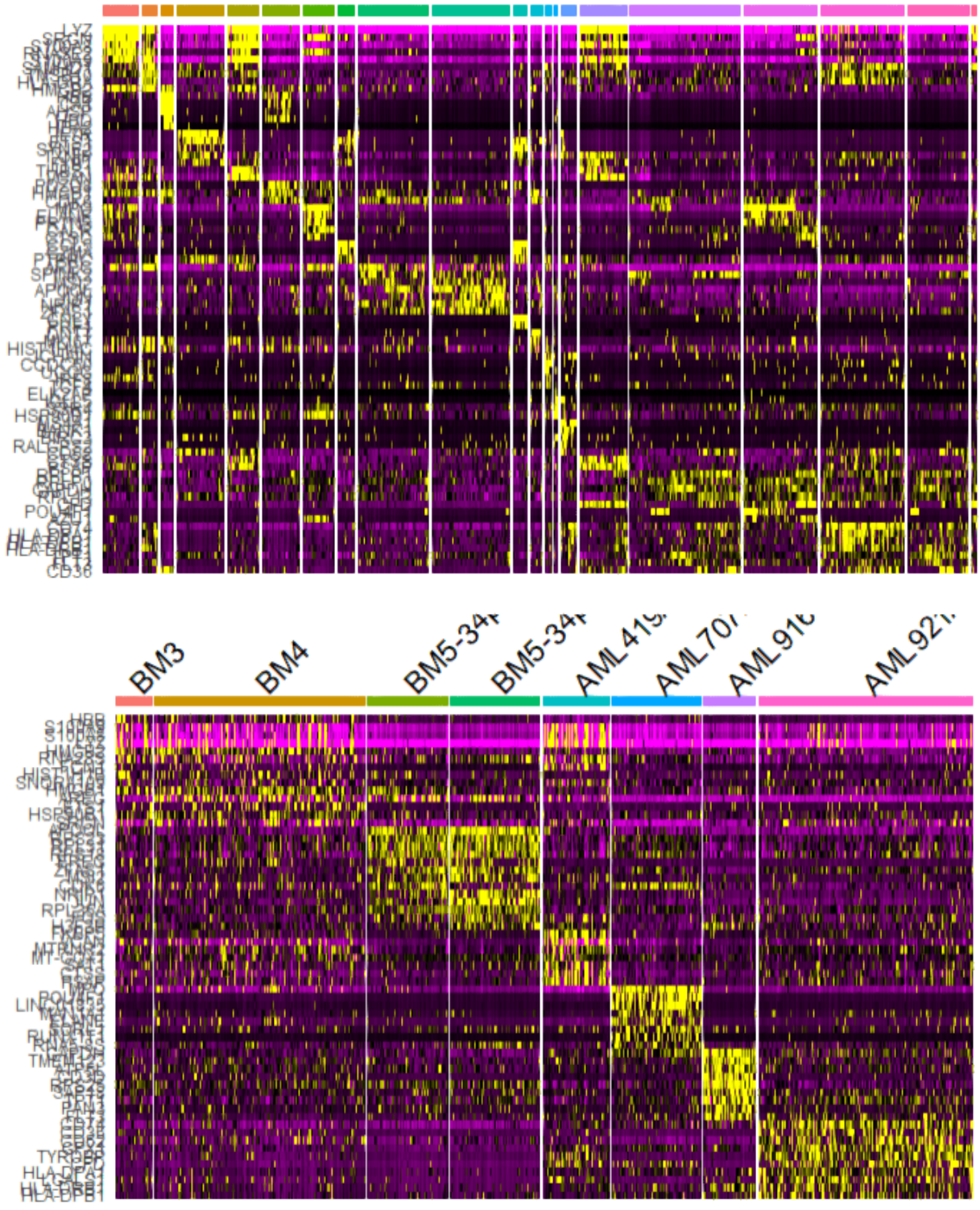


Figure 3. UMAP plot of SCTransformed counts

In the SCTransformed UMAP plots, single cells formed distinct clusters by cell-type, patient origin, and malignance. One notable proof of principle is that the healthy bone marrow samples (BM3, BM4, BM5-34p, BM5-34p38n) together form a distinct cluster located at the central region of the UMAP plot. This cluster is characterized by having a high HSC score and is uniformly predicted to be non-malignant. Such results are expected because healthy bone marrows are largely composed of developing hematopoietic stem cells. Contrastingly, cancer cells had a tendency to locate far from this central cluster, and away from one another. This is most likely due to the fact that AML can result from a variety of genomic disruptions and are not limited to one carcinogenic mechanism.

We next performed the differential gene expression analysis of the scRNA-seq data based on the same three comparisons: cell types, patient origins, and malignance predictions. We plotted these results in the form of a heatmap.


BM3
BM4
BM5-34
BM5-34
AML41
AML70
AML91
AML92

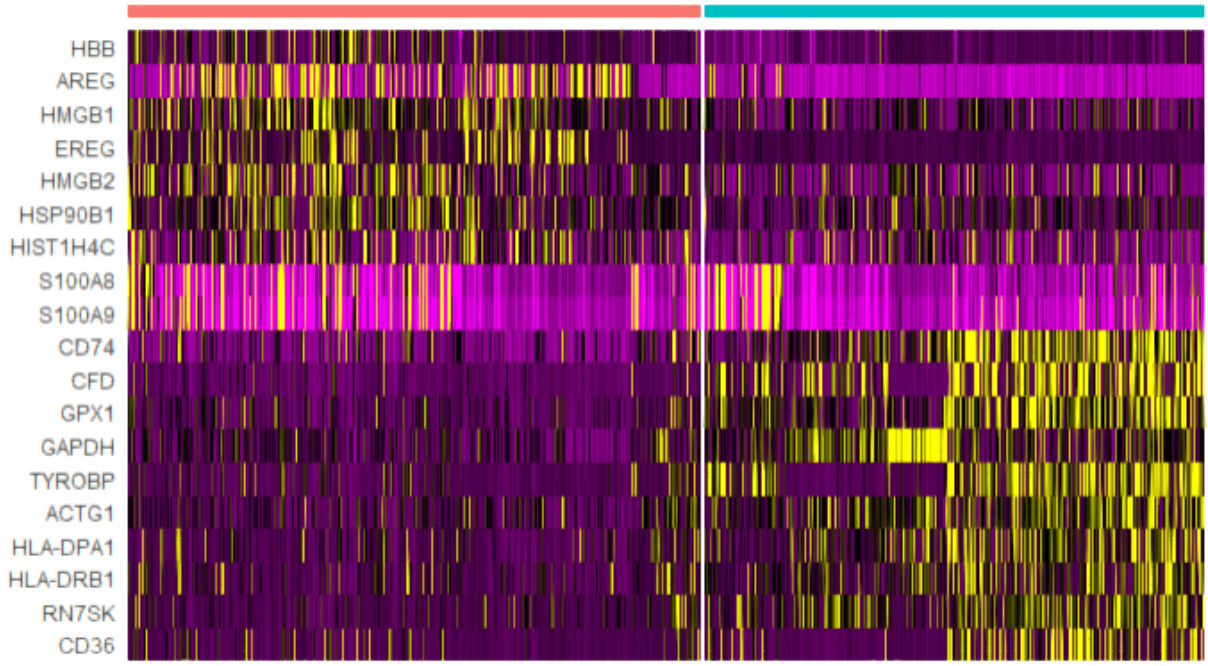


Figure 4. *(Top)* by cell type. *(Middle)* by patient. *(Bottom)* by malignance prediction. Note that a formatting issue caused the X-labels to be omitted in the *Top* and *Bottom* heatmaps. *Top* labels: ProMono, CDC, lateEry, T, Mono, earlyEry, GMP, CTL, Prog, HSC, NK, Prob, pDC, Plasma, B, Mono-like, Prog-like, GMP-like, CDC-like, HSC-like, ProMono-like. *Bottom* labels: Normal, Malignant.

Without looking at the identity of these DEGs, one noticeable feature of the heatmap comparing the cell types (Figure 4, Top) is that some cell types, notably ProMono, CDC, Mono, and Mono-like exhibit highly analogous DE signatures. This is because CDCs, or conventional dendrocytes, originate from monocytes. There were several other instances in which cell types sharing a common developmental origin exhibited highly similar DEGs.

The comparison by patient sample displayed distinct DEGs for BM5 and each of the AML samples, but not BM3 and BM4 (Figure 4, Middle). Single cells within these two healthy samples exhibited varying degrees of differential expression that were difficult to distinguish by eye. The cause of this variability is possibly that these genes are involved in a variety of cellular functions. For instance, HBB is the human beta hemoglobin protein which is expressed by erythrocytes. S100A8 and S100A9 are Ca2+ binding proteins belonging to the S100 family involved in inflammatory responses, typical of leukocytes. As healthy bone marrow samples are composed of HSCs at varying steps in the development and differentiation, it is expected that they exhibit less noticeable degrees of differential expression compared to AMLs which, by definition, arise from a common leukocyte ancestor and a carcinogenic mechanism.

Along similar lines, the GO terms associated with DEGs in healthy samples were highly variable, ranging from human hemoglobin protein and histone subunits to those involved in

inflammatory responses (i.e. tumor suppressor genes) (Figure 4, Bottom). On the other hand, those associated with malignant cells were those characteristic of proto-oncogenes as per our expectation.

Analysis and Conclusion

Using the Seurat package in RStudio, we were successfully able to perform QC, preliminary scRNA-Seq analysis, and more in-depth DGE analysis on our sample dataset to yield clear, separable results. The method of RNA sequencing, though more costly, was able to sort sample cells by type, predicted malignance, and patient origin, and identify the most upregulated and downregulated genes in these clusters. A large volume of invaluable results can be obtained from scRNA-Seq, applicable in further AML research and that of virtually all cancers and genetic disorders.

Citations:

Kirchner, Mary Piper, Lorena Pantano, Meeta Mistry, Radhika Khetani, Rory. "Introduction to Single-Cell RNA-Seq." *Introduction to Single-Cell RNA-Seq - ARCHIVED*, 24 Feb. 2020, hbctraining.github.io/scRNA-seq/lessons/01_intro_to_scRNA-seq.html. Accessed 14 Dec. 2021.

Van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, Pinkus GS, Pozdnyakova O, Galinsky I, Stone RM, Graubert TA, Shalek AK, Aster JC, Lane AA, Bernstein BE. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019 Mar 7;176(6):1265- 1281.e24. doi: 10.1016/j.cell.2019.01.031. Epub 2019 Feb 28. PMID: 30827681; PMCID: PMC6515904.

Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 20, 296 (2019).

https://doi.org/10.1186/s13059-019-1874-1

Contributions:

All members contributed to this project. The Seurat R script was compiled collectively by the group. Nivedha Satyamoorthi presented the initial UMAP plots, Soon Hoo Lee performed the Seurat object integration and DGE analysis, and Amatya Pathak also assisted in DGE analysis and compiled the written report.