**Methods**

Packages:

* R (v. 3.6.2)
* DESeq2 (v. 1.26.0)
* Seurat (v. 3.1.4)
* SingleR (v. 1.0.5)
* ggplot2 (v. 3.3.0)

Processing & Clustering

The filtered features barcoded matrix from the sequencing core was used and read into R via Seurat for downstream analysis. A file containing cell barcodes and HTO’s was also supplied and cells without an HTO were excluded from future analysis (see non-tagged analysis, FUTURE). Standard processing (https://satijalab.org/seurat/v3.1/pbmc3k\_tutorial.html) of the data was followed, with differences in values and tuning parameters along the way.

Cells with fewer than 500 and greater than 5,000 unique features were excluded from analysis, along with cells that contained greater than 10% of the reads coming from mitochondrial RNA (see Figures 1 and 2, FUTURE). The 2,000 most variable features were used to determine clusters. A shared nearest neighbor graph was constructed and then clusters were created based on shared nearest neighbors, with a resolution of 0.45. Clusters were visualized using a UMAP projection.

Cell Labeling

SingleR, an R package, was used for a first pass of labeling the clusters. SingleR comes with two mouse datasets: Immunologic Genome Project and mouse bulk expression data of sorted cell populations. These datasets consist of 830 and 358 experiments respectively, each covering over 20,000 features (genes). Transcriptomic profiles of the clusters were then compared to all the reference experiments and then each cluster was labeled with the closest reference experiment. However, megakaryocytes, erythroid, and their progenitors were not among the reference panel, and are believed to be in the experiment.

The next step was to look at known marker genes and their expression in the clusters. Marker genes were provided by collaborators, found through a literature review and using PanglaoDB (a single cell database). Clusters were further identified and labels were placed on the cluster.

?still working on macrophage and stem cell labels (sent csv’s containing marker genes)

Differential Expression

Next differential expression within clusters between controls and mutants was performed. (Not sure what the gold standard is here, NOTES). The tests were performed using the FindMarkers function in Seurat using four different tests: DESeq2, wilcox (Wilcoxon Rank Sum test), bimod (likelihood-ratio test for single cell gene expression), and roc (ROC analysis). Concordance on significance ( < 0.05 adjusted p-value or > 80% power) was used to identify genes that were differentially expressed between cases and controls.