Investigating Inconsistencies in Single-cell RNA Sequencing Data

Pranay V. Singh (Bowdoin College — psingh@bowdoin.edu), Advisors: Juliet Emamaullee, Brittany Rocque, Cameron Goldbeck (University of Southern California)

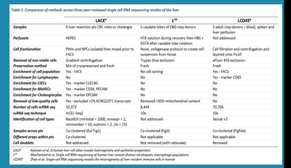
Bowdoin | Keck Medicine of USC

INTRODUCTION

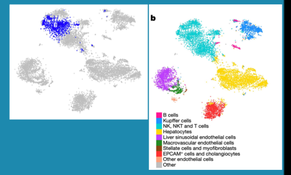
- Single-cell RNA sequencing (scRNA-seq) has quickly gained traction, emerging as a robust method for computational genomics. Researchers can use scRNA-seq to better understand how disease manifests in order to find cures, leading to significant and extensive impacts worldwide. In this study, we examined three normal 'human liver' scRNASeq datasets, LCD45^e, LACE^e, and L^{nb}, to see if different tissue processing techniques could alter scRNASeq outputs. Our analysis suggests a need for a standardized procedure to preserve and treat tissues before sequencing.

METHODS

- We used datasets from three sources: first, Liver atlas, cholangiocyte and endothelial enriched (LACE^e), second, Liver CD45+ enriched (LCD45^e), and third, Non-biased Liver (L^{nb}). Data from two of the three sources (LACE^e and L^{nb}) contained all liver cell types and data from the third source (LCD45^e) contained liver, spleen, and blood immune cells. Since we sought to compare a common set of cells across all three sources, we extracted immune liver cells from each dataset.

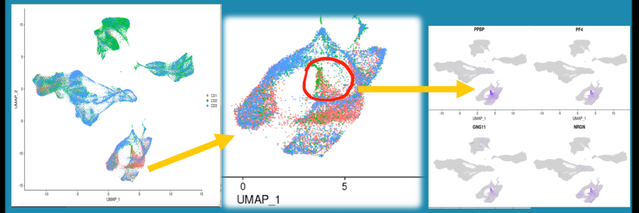


Comparison of treatment and preservation



Identifying immune cells from LACE^e

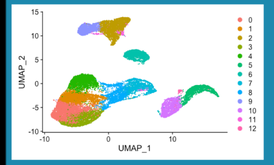
Extracting immune cells from L^{nb}



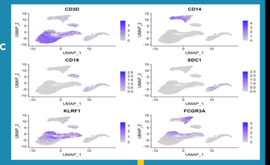
Extraction process for LCD45^e – split cells into 3 groups, plotted blood/spleen genes to identify liver immune cells (blue)

METHODS cont.

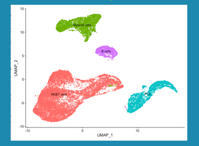
- We normalized, scaled, clustered, non-linearly dimensionally reduced (UMAP), and graphed each of the datasets and their genes to determine cluster cell types and identified four types: Myeloid, NK&T, B, and Plasma.
- Computed differential gene expression, comparing one cell type to all other cell types within and across datasets, and we found the absolute count gene expression for each cell type, both of which would be used for our analysis.
- Employed three types of evidence: cell proportions, gene expression, and visual integration.



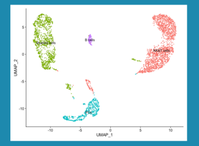
Cluster specific gene plotting to determine cell types



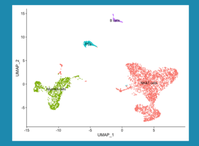
Cluster cell type identification



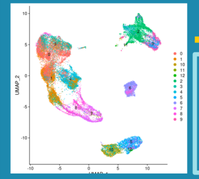
LCD45^e



L^{nb}



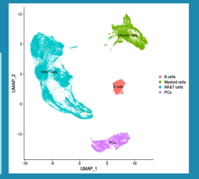
LACE^e



Seurat Clustering

Visual Integration of All Datasets

Used a multiprocessor system with 16 processors and 64GB memory for integration. The analysis of the integrated dataset (32,688 cells and 44,258 genes) took 46 minutes and ~50GB of memory.



Cluster Identification

RESULTS

- Cell proportions differed significantly across sets. B cells were scarce but made up a higher proportion in the LCD45^e dataset (X^2 test, p<0.01). NK and T cells were the largest subset across all datasets but made up a smaller proportion in the L^{nb} cell population.
- Pairwise-correlation plots using gene expression counts separated by study demonstrated greater similarity between LCD45^e and L^{nb} (R^2 = 0.91), and comparison of LACE^e to L^{nb} and LCD45^e revealed decreased concordance (R^2 = 0.66, R^2 = 0.61). When separated by cell types, consistent R^2 values and results were found.
- Volcano plots of differential gene expression across studies revealed the number of differentially expressed genes with significant fold change values (p<0.01, 0.8<FC<1.25). Across cell types, we found an average of 59.75, 145, and 117.25 significantly differentially expressed genes for L^{nb} vs LCD45^e, LCD45^e vs LACE^e, and LACE^e vs L^{nb}, respectively, resulting in relatively low differential gene expression for L^{nb} and LCD45^e when compared to the other two combinations.
- We found the top 10 differentially expressed genes in each cell type for each data set and collected the unique genes. Heatmaps of the 12-13 unique genes perceptibly depict relative differences of gene expression across datasets. From left to right, the order of columns is LACE^e, LCD45^e, L^{nb}.

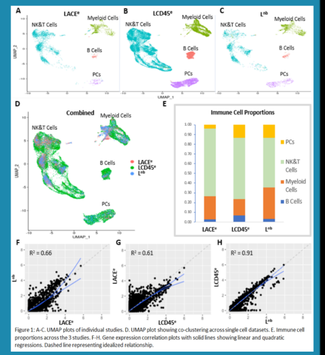
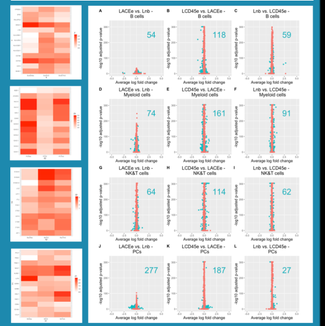


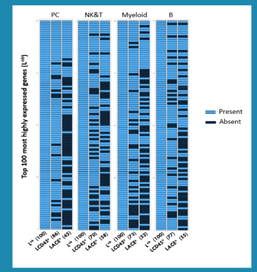
Figure 1: A-C. UMAP plots of individual studies. D. UMAP plot showing co-clustering across single cell datasets. E. Immune cell proportions across the 3 studies. F-H. Gene expression correlation plots with solid lines showing linear and quadratic regressions. Dashed line representing idealized relationship.

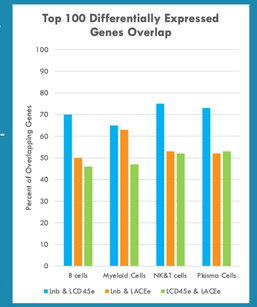


- Visual integration of the three datasets reveal that LACE^e and L^{nb} cluster with themselves within LCD45^e, indicating that each datasets' cells cluster together.

RESULTS cont.

- Using the gene expression results, we found the top 100 most expressed genes of L^{nb} and compared the other two datasets to L^{nb}, measuring gene overlap. On average, 76.5% LCD45^e genes and 52.5% LACE^e genes matched with L^{nb}, demonstrating greater similarity between LCD45^e and L^{nb}.
- Comparing the top 100 differentially expressed genes between the studies by cell type, L^{nb} and LCD45^e consistently matched the most—average was 70.75% of genes matching across all four cell types, while L^{nb} and LACE^e matched at an average of 54.5%, and LACE^e and LCD45^e matched the worst at 49.5%, further demonstrating similarity between LCD45^e and L^{nb}.





CONCLUSION

- Differences in cell proportions and gene expression could be due to variability in treating tissue. For scRNA-seq technique, LACE^e used mCEL-seq2 while L^{nb} and LCD45^e used 10x Genomics. Variability in preservation techniques, such as cryopreservation or using various solutions to put samples in, could dehydrate cells, contributing to an increase of cells adhering together. Although the livers across these studies are not biologically identical, biological differences between the studies are expected to be comparable across all three datasets.
- There is strong evidence for L^{nb} and LCD45^e resulting in more similarity than L^{nb} and LACE^e or LCD45^e and LACE^e; therefore, the notable differences we see between LACE^e and the other two sources are likely due to preanalytical variables. Due to the scarce and expensive nature of scRNA-seq, we thus must be careful in the process of preserving liver tissue and collecting their scRNA-seq data. Our research and analysis highlight the need for a standardized procedure for the treatment and preservation of liver cells before sequencing.

Acknowledgement: The authors are grateful to Bowdoin College for sponsoring Pranay's summer internship at USC's Transplant Research Lab.