Investigating Inconsistencies in Single-cell RNA Sequencing Data

PRANAY V. SINGH

Bowdoin College

JULIET EMAMAULLEE (ADVISOR)

University of Southern California

BRITTANY ROCQUE (ADVISOR)

University of Southern California

CAMERON GOLDBECK (ADVISOR)

University of Southern California

CCS CONCEPTS • Applied Computing → Life and Medical Sciences → Genomics → Computational genomics

Additional Keywords and Phrases: Single-cell RNA sequencing (scRNA-seq), preanalytical variables, data analytics, Seurat, R, clustering, gene expression, immune cells, computational genomics.

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) has emerged as a robust method for computational genomics [1]. It combines cell biology with computer science in order to analyze thousands of cells at a single-cell level, allowing us to observe gene expression levels and how they differ across diverse samples. This translates to a knowledge about the exact functions of these cells which can be used to better understand disease processes and discover exciting new potential therapies. However, very few studies [2] analyze the effects of preanalytical variables such as tissue handling and scRNA-seq technique on sequencing results. In this study, we examined three normal 'human liver' scRNA-seq datasets from high-quality peer-reviewed publications to see if they yielded comparable results. We hypothesized the pre-analytical variables in scRNA-seq studies would have an impact on characterization of the human liver immune cell transcriptome.

2 METHODS

2.1 Data Acquisition

We used datasets from three sources: first, Liver atlas, cholangiocyte and endothelial enriched ($LACE^{e}$) [3], second, Liver CD45+ enriched ($LCD45^{e}$) [4], and third, Non-biased Liver (L^{nb}) [5]. Data from two of the three sources ($LACE^{e}$ and L^{nb}) contains all liver cell types and data from the third source ($LCD45^{e}$) contains liver,

spleen, and blood immune cells. To compare common cells across all three datasets, we extracted immune liver cells which expressed the marker CD45.

2.2 Data Analysis and Visualization

Using Seurat [6], we normalized, scaled, clustered, and non-linearly dimensionally reduced (UMAP) the extracted immune liver cells. We visualized the datasets and their genes and identified four cluster cell types: Myeloid, NK&T, B, and Plasma. We employed three types of evidence for our analysis: cell proportion, gene expression, and visual integration. First, we integrated the datasets into one large dataset and executed the same code sequence on the integrated object to find cell proportions. Second, we performed differential gene expression, comparing one cell type's genes to all other cell types within and across datasets. We also found absolute count gene expression for each cell type. We used ggplot2 in R [7] to generate volcano plots, heatmaps, pairwise-correlation plots, and top 100 gene plots. Finally, we visualized the integrated dataset using UMAP, noticing how the clusters moved apart or assimilated.

For computation and data analysis, we used a multiprocessor system with 16 processors and 64 GB memory. Analyzing the integrated dataset, containing 32,688 cells and 44,258 unique genes, took 46 mins and ~50 GB of memory.

3 RESULTS

Cell proportions differed significantly across datasets. Notably, B cells were scarce, with a higher proportion in $LCD45^e$ (X^2 test, $p<0.01$). NK&T cells formed the largest cell group but were a smaller proportion of the L^{nb} cell population. Gene expression analysis 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 ($R^2 = 0.66$, $R^2 = 0.61$).

Volcano plots of differential gene expression revealed genes with significant fold change values ($p<0.01$, $0.8<FC<1.25$). Across cell types, on average, 59.75, 145, and 117.25 genes were significantly differentially expressed between L^{nb} and $LCD45^e$, $LCD45^e$ and $LACE^e$, and $LACE^e$ and L^{nb}, respectively. Next, we found the top 10 differentially expressed genes in each cell type for each dataset. Due to strong overlap among the top 10 genes, 12-13 unique genes were used for heatmaps. The heatmaps perceptibly depict relative differences of gene expression between sources. Further, L^{nb}'s top 100 most expressed genes, on average, exhibited 76.5% and 52.5% overlap in $LCD45^e$ and $LACE^e$'s top 100, respectively. Comparison of the top 100 genes differentially expressed within sources for each cell type, on average, found 70.75% matching genes between L^{nb} and $LCD45^e$, 54.5% between L^{nb} and $LACE^e$, and 49.5% between $LACE^e$ and $LCD45^e$. Finally, visual integration of the three datasets reveals that $LACE^e$ cells co-cluster and L^{nb} cells co-cluster within $LCD45^e$, indicating that despite some differences in gene expression, the cells do have a general trend of expression overlap.

4 DISCUSSION

Differences in cell proportions and gene expression could be due to differences in tissue processing. While 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 when compared to $LACE^e$; therefore, the notable differences we see between $LACE^e$ and the other two studies are likely due to differences in technique (i.e., pre-analytical variables). Due to the scarce and

expensive nature of scRNA-seq, and the need to combine large datasets in order to increase sample size, we caution that care must be taken 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. After standardization, we as a community can trust scRNA-seq data in order to explore how diseases manifest and gain a better understanding of cellular phenotypes.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Tang, F., Barbacioru, C., Wang, Y. *et al.* 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6** (2009)**,** 377–382. https://doi.org/10.1038/nmeth.1315

[2] Lori L Bonnycastle, Derek E Gildea, Tingfen Yan, Narisu Narisu, Amy J Swift, Tyra G Wolfsberg, Michael R Erdos, and Francis S Collins. 2019. Single-cell transcriptomics from human pancreatic islets: sample preparation matters, Biology Methods and Protocols, Volume 4, Issue 1, 2019, bpz019, https://doi.org/10.1093/biomethods/bpz019

[3] Aizarani, N., Saviano, A., Sagar *et al.* 2019. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572** (2019)**,** 199–204. https://doi.org/10.1038/s41586-019-1373-2

[4] Zhao, J., Zhang, S., Liu, Y. *et al.* 2020. Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discov* **6,** 22 (2020). https://doi.org/10.1038/s41421-020-0157-z

[5] MacParland, S.A., Liu, J.C., Ma, X. *et al.* 2018. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9** (2018)**,** 4383. https://doi.org/10.1038/s41467-018-06318-7

[6] Butler, A., Hoffman, P., Smibert, P. *et al.* 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36** (2018)**,** 411–420. https://doi.org/10.1038/nbt.4096

[7] R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/