Analyzing Lung Cancer Cell Subtypes with CELDA

Eric Sinton[1,2], Dr. Joshua Campbell[2,3]
[1]Tufts University, [2]Boston University Bioinformatics BRITE REU Program Summer 2019,
[3]Boston University School of Medicine, Department of Medicine, Division of Computational Biomedicine

Abstract

Single-cell RNA sequencing (scRNAseq) is a recently-developed technology that allows for enhanced determination for heterogeneity of gene expression between subpopulations. This is in contrast to previously-standard bulk RNA sequencing, which only gave information about average RNA expression across a sample. Since scRNAseq is still relatively new, there is no definitive method across labs for which raw sequence data is processed and displayed in an intuitive manner, especially for non-computational users. A new software package called Cellular Latent Dirichlet Allocation (CELDA) was developed in the Campbell Lab to address this problem, by taking downstream RNA transcription data that is already in the form of a counts matrix, and analyzing it via a suite of Bayesian hierarchical statistical methods. CELDA is now also implemented in a graphical user interface called the Single Cell Toolkit (SCTK), which allows the user to interact with RNA expression data without using a command line.

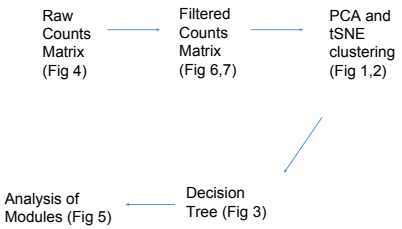
Cancer cells in particular exhibit great heterogeneity in gene expression, even between cells that are physically close. The purpose of this project was thus to test the functionality of CELDA on a set of RNA transcription data from human lung tumor samples. We compared the results that CELDA provided against the conclusions drawn by the paper containing the data. We found that our approach with the Campbell Lab software has similar analytical results when it came to identifying cell subtypes.

The project concluded that CELDA provides accurate analysis when tested on non-sample data, but further research is required to improve its functionality. The software should be tested on more human data. The user-interface can always be improved as well through user feedback. This will all contribute to a standard for which labs can process scRNAseq data for important tasks such as cancer research.

Background

Until recently RNA expression was profiled in "bulk", meaning the average expression data of many cells in a tissue was used to make conclusions about all of the cells in that tissue [1]. The use of scRNAseq, however, identifies the RNA profile of each cell in a sample. This difference is important because it is known that almost all human tissue exhibits heterogeneity between cell subtypes [2]. One example of this heterogeneity is in immune response. The immune system was already known to have various responses depending on an external threat, but scRNAseq helped to identify the unique cell subtypes that arise according to the body's response [1]. Humans exhibit heterogeneity in response to internal threats as well; Zilionis et al. [3] identified, with scRNAseq, 25 unique subtypes of Tumor-Infiltrating Myeloid cells (TIMs) in lung cancer patients. Thus the main goal of this project is to use CELDA to perform our own analysis on the data and distinguish cell sub-types as well. Sucessfully making these distinctions is important in this context because traditionally, few cancer patients respond well to targeted immunotherapies [3]. If we can accurately distinguish the types of TIMs and their relative prevalence when controlling tumor growth, it will give a better idea of how to construct patient treatment.

Methods



Results

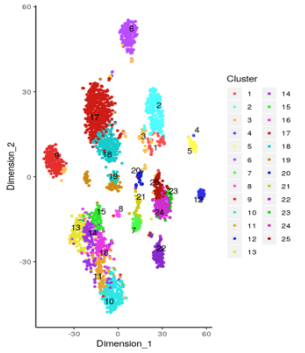


Figure 1: The labeled clusters as a result of tSNE for the patient's tumor sample cells. Each dot represents a cell, and the clusters are color-coded. Although we do not see at least 25 distinct clusters, the various color groups do represent different cell types.

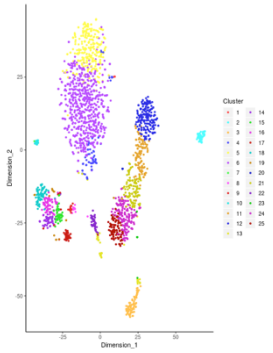
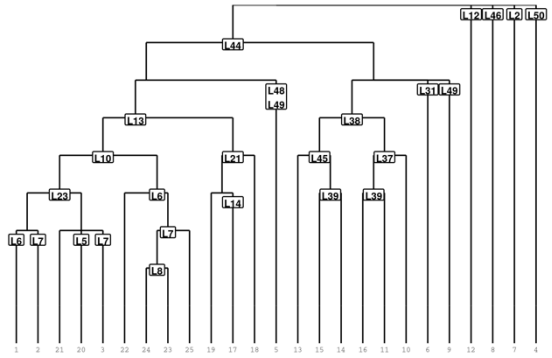


Figure 2: The unlabeled clusters as a result of tSNE for the same patient's blood sample. We see noticeably fewer clusters than in the tumor cells to represent cell sub-types.

Figure 3: A decision tree showing the feature modules that best split the clusters of tumor cells for a given patient. Lines leading from a module to the left represent downregulation, and to the right upregulation. Bottom numbers are the resulting cell clusters.



	Cell 1	Cell 2	Cell 3
Gene A	1	0	3
Gene B	0	6	0
Gene C	17	0	12

Figure 4: A small example of a counts matrix. When CELDA runs it will likely group cells 1 and 3 into a cluster, and genes A and C into a module

L7
"MT.ATP6"
"MT.ND1"
"MT.ND4"
"MT.CO3"
"MT.ND3"
"MT.CO1"
"MT.CYB"
"MT.CO2"
"MT.ATP8"
"MT.ND2"
"MT.ND5"
"MT.ND4L"
"MT.ND6"

Figure 5: The RNA features that make up module 7, which can be seen as the basis for several of the decision tree splits between cell clusters.

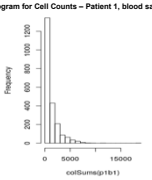


Figure 6: The distribution of counts across the cells in the blood sample of one of the lung cancer patients. Notice that the lower tail has been removed by filtering out cells that have fewer than 250 counts total.

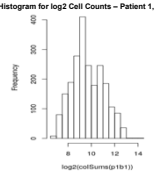


Figure 7: The distribution of the log (base 2) of the counts across the cells in the same blood sample. Since these are normally-distributed, we can use standard deviations to determine outliers.

Conclusions

When comparing the blood sample tSNE plot versus that from the tumor sample, we observe differences both in the number of clusters and in the shape of each cluster. The linear shape of the clusters in the blood sample indicate that cell subtypes are organized along some sort of continuum. What we do not observe in the tumor plot, however, is the presence of at least 25 *visual* clusters that would indicate the different TIM states that were mentioned for this particular data. Instead we count roughly 9 or 10 main clusters with a few smaller clusters in between the larger ones. This may still represent more cell subtypes, but further analysis is required. We should note that the visual analysis for tSNE clustering, and k-means clustering in general is often limited by the number of dimensions (in this case 2) that we can view with our eyes.

The decision tree does show the breakdown of cell clusters based on which modules are differentially-expressed between them. As an example, we list the genes in module 7, which bifurcates cluster 1, 2 and 3 from their respective branches. More analysis is required to view how module 7 is tied to these three clusters, and which specific cell subtype it may refer to.

Discussion and Future Research

Since the main goal of distinguishing 25 distinct TIM states was not accomplished, the use of different statistical methods should be used to investigate the data further and identify those states. More of the patient data can be analyzed, since there were seven lung cancer patients in total and the samples from only one was used. The interface for the tSNE plots could be improved. We note this because when the number of clusters jumps above 10, the colors are too similar for the human eye to distinguish easily. This could be solved by creating a graphic that makes a given cluster "pop out" when the mouse hovers over it.

References

[1] Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018;18(1):35-45.
[2] Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. Nat Rev Nephrol. 2018;14(8):479-92.
[3] Zilionis R et. al. Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. Immunity. 2019;50(5):1317-34.

Acknowledgments

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program, by LUNGevity Career Development Award, National Library of Medicine (NLM) R01LM013154-01, and by Informatics Technology for Cancer Research (ITCR) 1U01CA220413-01