Analyzing Lung Cancer Cell Subtypes with CELDA

Eric Sinton, Dr. Joshua Campbell
1Tufts University, 2Boston University Bioinformatics BRITE REU Program Summer 2019, 3Boston 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.

Methods

Raw Counts Matrix (Fig 4) Filtered Counts Matrix (Fig 6,7) PCA and tSNE clustering (Fig 1,2) Analysis of Modules (Fig 5) Decision Tree (Fig 3)

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 note that the lower tail of expression contains at least 25 distinct clusters, the various color groups do represent different cell types.

Figure 3: A decision tree showing the modules that best split the clusters of tumor cells for a given patient. The result from one module is then compared to the self-representational data, and to the right hierarchical. Bottom numbers are the resulting cell clusters.

Figure 4: An example of a scRNAseq clusters that would indicate the different TIM states that were mentioned for this particular lung cancer patient. 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.

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

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) is often limited by the number of dimensions (in this case 2) that we can view with our eyes.

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


Acknowledgments

This work was funded, in part, by NSF grant DBI-1559529, 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) 1U1CA224043-01.