UTILIZING TRANSFERRING LEARNING APPROACH IN SINGLE-CELL RNA SEQUENCING DATA ANALYSIS

by
Jinrui Liu

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
May 2020

© 2020 Jinrui Liu
All rights reserved

Abstract

The aim of this thesis is to develop theoretical understanding and enhance programming skills in computational biology research. Advanced high-throughput sequencing technologies have rendered data with high dimensionality such as single-cell data. Dimension reduction methods are widely applied to high-dimensional data, with additional analytical approaches we can interpret such data and discover novel biological phenomena during the cell differentiation process. Transferring learning is one of the approaches that can be used to discover associations and heterogeneity between single-cell datasets. This method is used to explore multiple datasets at the same time and facilitates a better understanding of the complicated cell differentiation process.

Acknowledgement

This research project was guided by Dr. Carlo Colantuoni, assistant Professor of Neurology and Neuroscience in Johns Hopkins University.

I would like to thank Carlo for his patience and generous amount of time and efforts that he contributed to lead into this bioinformatics world, which eventually led me towards my passion towards PhD study in computational biology.

I would like to thank Dr. Patrick Cahan for his time in advising me and helped me crafted a fruitful study plan for my master's study, which helped me smoothly switched to computational study with a zero-programming undergraduate background.

I would like to thank Dr. Brian Caffo, for introducing me to my mentor, Carlo; and for his data science class which helped me build my foundations in R programming so that I can work on this project.

Table of contents

List of Table

Table of Figures

Introduction:

Dimensionality reduction is a common approach to handle high-dimensional data created in computational biology research such as sequencing data. Single-cell RNA sequencing data is one typical type of multi-dimensional biological data. This project presents principal component analysis and transferring learning (TL) [1] method over a series of single-cell RNA sequencing datasets of neocortex development from different studies.

Problem:

The development of human's nervous system is a complicated biological process that involves interactions between a large diversity of cell types. Identifying unique cellular processes underlying human brain development has been challenging due to the fact that certain human brain tissues simply cannot be accessed experimentally [2]. Such study performed experimentally outside living organism is called "*in vitro*" study; while study performed in living organism, such as research involving human brain tissue, is referred as "*in vivo*" study. One of the neuroscience research focuses is on how neural progenitor cells (NPCs) develop and become postmitotic neurons, which eventually contributes to the formation of neocortex (fig.1). This process involves different types of NPCs and neurons that produce different neurotransmitters. Studies of this topic have been performed both *in vivo* and *in vitro* [3]. One approach to collect the data from both types of studies is through single-cell RNA sequencing technology, which yields the sequencing gene expression matrix of single cells that are at different stages during the development. The goal is to assess the resemblance and differences between *in vitro* and *in vivo* studies, through analyzing data coming from both systems.

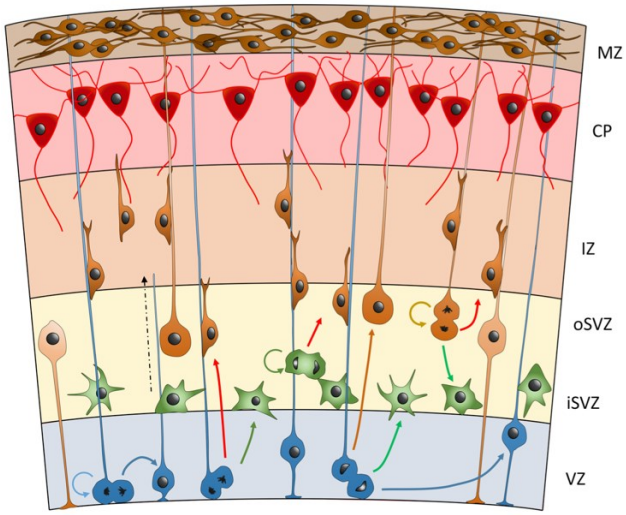

Figure 1 Schematic View of Six Layers of Neocortex [9].

VZ, ventricular zone, iSVZ, inner subventricular zone; oSVZ, outer subventricular zone; IZ, intermediate zone; CP, cortical plate; MZ, marginal zone.

Data:

In this project we analyzed eight single-cell RNA sequencing datasets from five different studies: Yao *et al* (*in vitro*), Sestan *et al* (*in vitro* and *in vivo*), Treutlein *et al* (*in vitro* and *in vivo*), Lim *et al* (*in vivo*) and Quake *et al* (*in vivo*) [4-8]. The single-cell sequencing techniques usually generate data matrices of the level at which **gene i** (row) is expressed in **cell j** (column). The expression matrix is usually at gene level but in general, RNA sequencing data can be processed at different levels including isoforms, junction, exons, non-coding RNA and alternative poly adenylation sites.

In our project, we started with focusing on gene level data. The sequencing data was collected at different stage along the trajectory of cell differentiation. It contains information representing different cell types: NPC and neurons. NPCs' development into six cortical layers of neurons can be marked by genes. Through visualization of the gene expression data using marker genes, we can observe the heterogeneity in NPCs and neurons as different layers are produced.

Methods:

To assess the multi-dimensional single-cell RNA sequencing data, we chose Principal Component Analysis (PCA) as the dimensionality reduction tool. PCA is one of the commonly used dimensionality reduction tools in computational biology research. In single-cell RNA sequencing data, each gene serves as one dimension. As one cell will be captured expressing multiple genes at the same time, dimensionality reduction is required to discover technical artifacts and biological phenomena from the data. PCA performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. A resulting PCA plot generally displays the first two principle components (two dimensions), with the x dimension being PC1 which maps the maximized variance of data and y dimension being PC2 which maps data orthogonal to PC1. In the PCA plot, each cell from the dataset holds a unique position on the two-dimensional surface and is represented by a single dot, as each cell has its own PC values. Meanwhile, with selected marker genes, we can locate their gene expression values on the data matrix, which is to be used in coloring the cells (dots) on the PCA plot. This whole process generates a visualized PCA plot. Using appropriate marker genes such as NPC markers and neuronal markers coloring the cells, we can observe the distinguished clouds of NPC and neurons on the two-dimensional plot. PCA were first performed across out eight datasets.

Then using transferring learning method *projectR* developed by Sharma and Colantuoni *et.al* we can further interpret and validate the data coming from *in vivo* and *in vitro* studies. The transferring learning process via PCA was achieved through projections across the eight datasets. Projection uses PCA result from a source dataset (e.g. dataset1) as space; then fit a new target dataset's (e.g. dataset2) data into the space, which yield a new visualized PCA plot. In addition to the PCA results from the eight datasets, each of the eight datasets will have the other seven datasets to be projected onto its own PCA. Thus, our transferring learning PCA and projections result can be plotted as an eight by eight matrix. The transferring learning result can be used to

further interpret the cell differentiation process of neocortex development, and also provides validations of *in vitro* studies compared to *in vivo* ones by relating features from the former to the latter. Through neocortex studies and review papers [4-10], we have selected a list of marker genes, for NPCs and neurons, to help visualize the PCA and projection results.

Results:

We started with PCA result of *in vivo* data from Treutlein *et al.* as the source of projection, and Lim *et.al* (*in vivo*), Quake *et.al* (*in vivo* prenatal), Treutlein *et al* (*in vitro*), Yao *et al* (*in vitro*) as target datasets. Using neuronal and NPC markers, we can locate positions of neurons and NPC groups on the source PCA map and the following positions should reveal the same patterns of NPC and neuron clusters. Table 1 listed the marker genes selected to color the cells (dots) on PCA/projection plots: HES1, EOMES, MYT1L, GAD1, SLC17A7. These genes marks NPCs and neurons, including neurons with specific neurotransmitters.

Table 1 Selected Marker Genes List with Additional Comments.

HES1 marks NPCs and neurons; EOMES makers IPC, which is one kind of NPC; DCX marks neurons. "~√" represents observable expression in certain type of cells.

Marker Gene	NPC	Neuron	note
HES1	√		
EOMES	√	~√	Intermediate Progenitor Cell (IPC)
MYT1L		√	
GAD1		√	GABAergic Neurons
SLC17A7		√	Glutamatergic Neurons

HES1 is a protein coding gene knowingly to be expressed by NPCs, which makes it a competent marker gene for all NPCs in our data. Similarly, MYT1L serves as a neuronal marker since its expression can be observed in most neurons. HES1 and MYTL1 as the general NPC and neuronal markers successfully helped us identify the two groups of NPC and neurons, as we can observe from the first and third rows from the matrix in fig.2. EOMES is mainly expressed by one type of NPC - intermediate progenitor cells (IPC), which produces neurons through neurogenesis. The existence of IPC validates the ongoing neocortex development. EOMES can also be expressed by newly born neurons. Thus, in some studies, it can also be used as a neuronal marker. Depending on the different neurotransmitters, the newly born neurons in the neocortex can also be classified into different types. The neocortex contains both excitatory and inhibitory neurons, named for their effect on other neurons, among which Glutamatergic and GABAergic neurons are the typical two types. Glutamatergic neurons produce glutamate, which is one of the most common excitatory neurotransmitters in the central nervous system; GABAergic neurons produce gamma-Aminobutyric acid (GABA) - the chief inhibitory neurotransmitter in the central nervous system. Thus, we selected GAD1 to mark GABAergic neurons and SLC17A7 to mark Glutamatergic neurons. The existence of these two types of neurons shows evidence of further formation of neocortex in addition to the confirmation of IPCs.

In row 1, marked by HES1, NPC population is clearly distinguished from neurons in the PCA plot of Treutlein *et.al in vivo* study. With PCA of Treutlein *et.al in vivo* study as source providing the "space", the projections with target datasets are supposed to match the pattern that right cloud of clustered cells to be NPCs and the left cloud to be neurons. Across the row of HES1 gene, every projection yields the expected pattern and boundary of NPCs and neurons, except that, in the projection of Treutlein *et.al in vitro* study, the ratio of NPCs (yellow dots) is much more significant comparing to other studies' projections. Similarly, in row 3, MYT1L successfully separated neurons and NPCs in both source PCA result and projections.

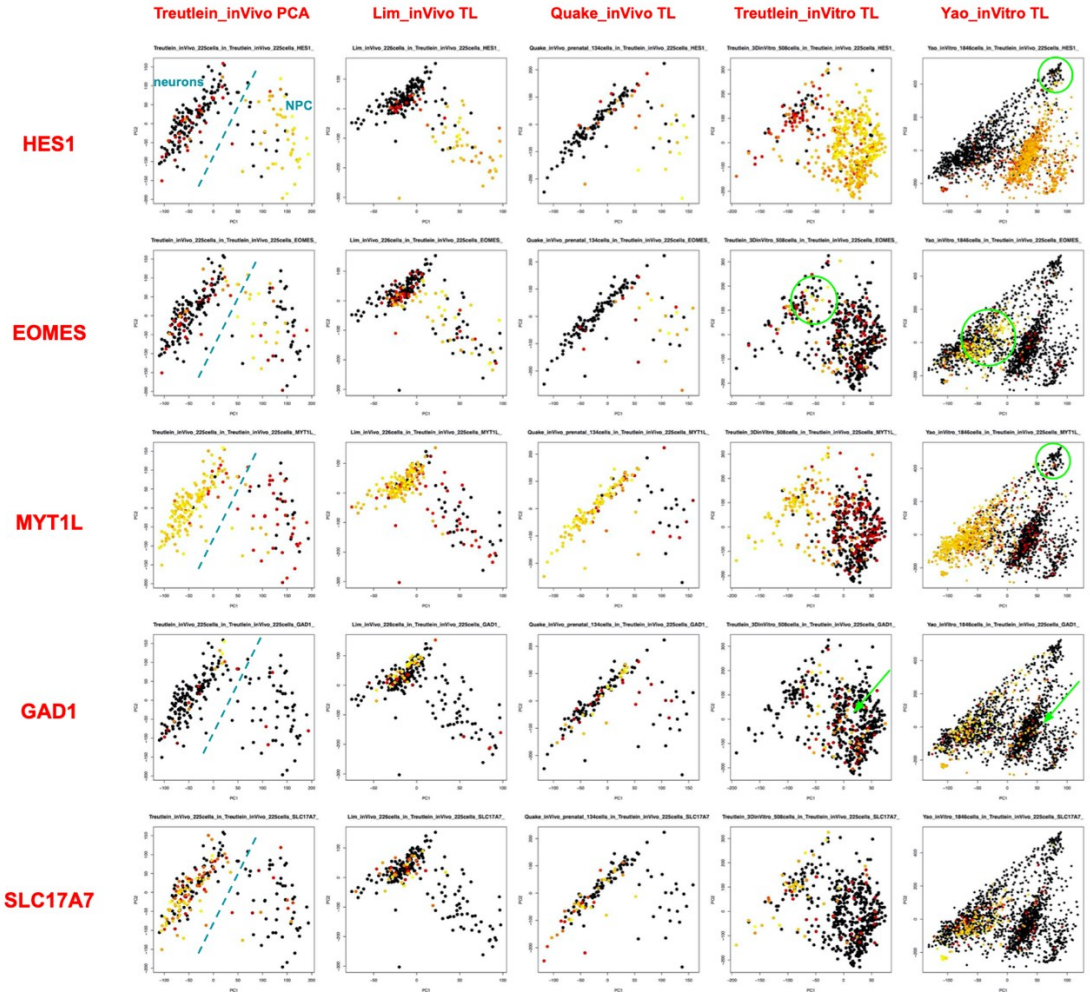

Figure 2 Transferring Learning plot using in vivo dataset as source.

Each row is different PCA/projection results marked by the same gene, each column is the same PCA/projection result. 1st column is the PCA of 225 cells from Treutlein *et.al in vivo* study (Treutlein-inVivo); 2nd column is the data of 226 cells from Lim *et.al in vivo* study projected onto Treutlein-inVivo; 3rd column is the data of 134 cells from Quake *et.al in vivo* study projected onto Treutlein-inVivo; 4th column is the data of 508 cells from Treutlein *et.al in vitro* study projected onto Treutlein-inVivo; 5th column is the data of 1846 cells from Yao *et.al in vitro* study projected onto Treutlein-inVivo. Based on expression level of the marker gene in each cell from high to low, the color of the cells changes from light to dark (yellow to black).

In row 2, the three *in vivo* datasets didn't record active EOMES expression in the neuron groups and it's mainly expressed by NPCs. Since IPC is only one type of the NPCs, NPCs clouds displayed

only partly active expression of EOMES as expected. The patterns of target *in vivo* studies (Lim *et.al* and Quake *et.al*) is in good consistency with the source (Treutlein *et.al in vivo*). Contrarily, the *in vitro* studies from Treutlein *et.al* and Yao *et.al*, showed active expression of EOMES only in their neuron clouds. This result indicates a divergence of EOMES expression in the neocortex when studies follow *in vivo* versus *in vitro* protocols.

In row 3, despite neurons are clearly distinguished from NPCs by MYTQL, in the study of Yao *et.al,* compared to its own plot marked by HES1, the supposed intersection of NPCs and neurons are silent with both marker genes. Moreover, the circled peak also didn't show expression of the other three genes. It is possible that those cells grew under *in vitro* protocol didn't successfully develop into NPCs or neurons. Additional marker genes can help confirm this result.

In row 4 marked by GAD1, first PCA plot showed that neuron populations in Treutlein *et.al in vivo* dataset barely contains GABAergic Neurons. This can be attributes to cells in this study are collected in the different part of brain tissue during which GABAergic neurons haven't migrated there [11]. In the other two *in vivo* studies (Lim *et.al* Quake *et.al*), GABAergic neurons were correctly located in the neuron clouds. In the *in vitro* studies (Treutlein *et.al* and Yao *et.al*), slight expression of GAD1 are also recorded in some of the supposed NPC groups. In the projection of *in vitro* study of Treutlein *et.al*, GAD1 expressions were found in some cells in the NPC cloud. Referring to the same projection marked by MYTL1, lower neuronal expressions were also recorded at the same position. It is possible that *in vitro* experiment protocol Treutlein *et.al* applied may induced early neuronal expressions in supposed NPCs. This indicates a possible transitional state of NPCs developing into neurons in *in vitro* experiments.

In row 5, glutamatergic neurons were consistently located in neuron groups of both *in vivo* and *in vitro* studies. Hence, a preliminary conclusion can be drawn that the selected *in vitro* studies resembles the *in vivo* process correctly regarding glutamatergic neuronal development.

Discussion:

Transferring learning approach via PCA successfully related the features learned from source dataset to the target datasets. Fig.1 above results reveals a preliminary understanding in lineage relationship between *in vivo* and *in vitro* studies of neurogenesis. While *in vitro* studies by Treutlein *et.al* and Yao *et.al* were validated to have replicated the neurogenesis and is highly consistent with the *in vivo* studies of Treutlein *et.al,* Lim *et.al* and Quake *et.al*, EOMES and GAD1 still raised questions about the transitional process between NPCs and neurons, which requires further analysis with additional marker genes and projections using other datasets as PCA sources. In addition to transferring learning approach, we are planning on using *Seurat* pipeline developed by Satija *et al.* to identify and interpret sources of heterogeneity in the datasets.

References:

[1] Sharma, Gaurav, et al. "*projectR: An R/Bioconductor package for transfer learning via PCA, NMF, correlation, and clustering*." BioRxiv (2019): 726547.

[2] Arlotta, Paola, and Sergiu P. Paşca. "*Cell diversity in the human cerebral cortex: from the embryo to brain organoids.*" Current opinion in neurobiology 56 (2019): 194-198.

[3] Lui, Jan H., David V. Hansen, and Arnold R. Kriegstein. *"Development and evolution of the human neocortex."* Cell 146.1 (2011): 18-36.

[4] Yao, Zizhen, et al. "*A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development*." Cell stem cell 20.1 (2017): 120-134.

[5] Onorati, Marco, et al. "*Zika virus disrupts phospho-TBK1 localization and mitosis in human neuroepithelial stem cells and radial glia*." Cell reports 16.10 (2016): 2576-2592.

[6] Camp, J. Gray, et al. "*Human cerebral organoids recapitulate gene expression programs of fetal neocortex development*." Proceedings of the National Academy of Sciences 112.51 (2015): 15672-15677.

[7] Liu, Siyuan John, et al. "*Single-cell analysis of long non-coding RNAs in the developing human neocortex*." Genome biology 17.1 (2016): 67.

[8] Darmanis, Spyros, et al. "*A survey of human brain transcriptome diversity at the single cell level*." Proceedings of the National Academy of Sciences 112.23 (2015): 7285-7290.

[9] Penisson, Maxime, et al. *"Genes and mechanisms involved in the generation and amplification of basal radial glial cells."* Frontiers in cellular neuroscience 13 (2019): 381.

[10] https://www.labome.com/method/Neuronal-Cell-Markers.html

[11] Letinic, Kresimir, Roberto Zoncu, and Pasko Rakic. "Origin of GABAergic neurons in the human neocortex." *Nature* 417.6889 (2002): 645-649.

[12] Satija, Rahul, et al. "Spatial reconstruction of single-cell gene expression data." *Nature biotechnology* 33.5 (2015): 495-502.

Curriculum Vitae

Jinrui Liu

1 E University Parkway, Unit 411, Baltimore, MD (573) 825-1282
jliu179@jhu.edu https://louisjrliu.github.io/

EDUCATION

Johns Hopkins University, *Baltimore, MD* Expected: *May 2020*
Master of Science in Biomedical Engineering
- Relevant Courses:
  Applied Comparative Genomics, Foundations in Bioinformatics, Data Structure, Biostatistics, Machine Learning, Biomedical Data Science

University of Missouri, *Columbia, MO* *2016 - 2018*
Bachelor of Science in Bioengineering (Dual Bachelor) - Cum Laude
- Awards and Grants:
  Bioengineering Honors Scholar: 15/300 students with outstanding performance in research, *2018*
  Speaker of Faculty Excellence in Research Award: 1/12 Bioengineering Honors scholars, *2018*
  Curator's Grant-in-Aid Scholarship: 50/3000 international students, top price, *2017*
  Undergraduate Research Fellowship: 20/100 undergraduate research assistants, top price, *2017*
  Dean's List, *2016 – 2018*

East China University of Science and Technology, *Shanghai, CHINA* *2013 - 2016*
Bachelor of Science in Bioengineering (Dual Bachelor)

SKILLS

Programming	R, Python, Java, Unix Scripting, HTML
Tools	Bioconductor, bedtools, PHATE, etc.
Technologies	PyCharm, IntelliJ, Anaconda, Google Colab, RStudio, GitHub

RESEARCH EXPERIENCE

Graduate Research, Prof. Carlo Colantuoni
Johns Hopkins University *May 2019 – Present*
- Master's thesis in dimensionality reduction analysis of single-cell RNA sequencing data
- Programed in R language on JHPCE cluster to perform PCA and corresponding projections
- Wrote master's thesis report
- Use Seurat Pipeline for further analysis

Undergraduate Research, Prof. Heather Hunt
University of Missouri *Sep 2016 – May 2018*
- Bench work of chemical syntheses for 24 series of silica materials
- Conducted data acquisition and analysis for totally 48 series of syntheses
- Wrote the honor thesis manuscript and received the honors scholar title

Undergraduate Research, Prof. Dong Xu
University of Missouri *Jun 2017 – Aug 2017*
- Programmed in HTML and JavaScript for the maintenance of the university's websites
- Performed genomic transcription and translation practice with MATLAB

RELEVANT PROJECTS

Monocle Performance Assessment, Prof. Michael Schatz
Johns Hopkins University *Apr 2020 – Present*
- Explore the utility of Monocle toolkit and assess its performance with low-coverage sequencing data

Graduate Research, Prof. Adam Sapirstein, Prof. Raimond Winslow
Johns Hopkins University *Sep 2019 – Present*
- Machine learning project for disease risk stratification based on large-scale patient datasets
- Data mining in Python and machine learning model construction in R

Bioinformatics Project with HIV Genomics Data, Prof. Rachel Karchin
Johns Hopkins University *Feb 2019 – May 2019*

- Analyzed genomics data using statistical approaches to stratify mutation fidelity

R Shiny App – Coffee, Coffee, Prof. Brian Caffo
Johns Hopkins University *Apr 2019 – May 2019*
- Built an interactive shiny application featuring search trend analysis and stock prediction

TEACHING EXPERIENCE

Teaching Assistant, Advanced Data Science for Biomedical Engineering, Prof. Brian Caffo
Johns Hopkins University *Jan 2020 – May 2020*
- Guide students on GitHub utilization, Unix scripting and R programing
- Hold office hours and grade homework

PROFESSIONAL EXPERIENCE

Business Analyst Intern
Mybiogate Inc. *Houston, TX* *Jun 2018 – Aug 2018*
- Translated 25 biomedical projects and corresponding FDA regulations for Chinese biotechnology investors
- Attended *2018 China (Suzhou) Cross-border Technology Transfer Convention* conference, working as the coordinator and interpreter for 13 American groups and 20 Chinese investment institutions