

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**MATHEMATICAL MODELLING OF MOLECULAR
TRAJECTORIES IN STEM CELL TRANSFORMATION PROCESSES
VIA BIOINFORMATIC ANALYSIS OF SINGLE CELL OMICS DATA**

BATUHAN ÇAKIR
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

GEBZE
2021

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**MATHEMATICAL MODELLING
OF MOLECULAR TRAJECTORIES IN
STEM CELL TRANSFORMATION PROCESSES
VIA BIOINFORMATIC ANALYSIS
OF SINGLE CELL OMICS DATA**

BATUHAN ÇAKIR
**A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE**
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

THESIS SUPERVISOR
ASSIST. PROF. DR. PINAR PİR

GEBZE

2021

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KÖK HÜCRE
BAŞKALAŞIM SÜREÇLERİNDEKİ
MOLEKÜLER YÖRÜNGELERİN
TEK HÜCRE OMİK VERİSİNİN
BİYOİNFORMATİK ANALİZİYLE
MATEMATİKSEL MODELLENMESİ

BATUHAN ÇAKIR
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ PROGRAMI

DANIŞMANI
DR. ÖĞR. ÜYESİ PINAR PİR

GEBZE
2021



YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 15/01/2021 tarih ve 2021/03 sayılı kararıyla oluşturulan jüri tarafından 19/01/2021 tarihinde tez savunma sınavı yapılan Batuhan ÇAKIR'ın tez çalışması Biyomühendislik Anabilim Dalı Biyoinformatik ve Sistem Biyolojisi Programı YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI): Dr. Öğr. Üyesi Pınar PİR

ÜYE

: Doç. Dr. Tunahan ÇAKIR

ÜYE

: Dr. Öğr. Üyesi Tolga SÜTLÜ

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
...../...../..... tarih ve/..... sayılı kararı.

SUMMARY

Stem cell is the cell type that has the capabilities of self-renewal and differentiation into specialized cell types. Ability to proliferate throughout their lifespan and transform themselves are the main advantages of the stem cells. With the latest technologies, stem cell studies have been accelerated in the last decade and stem cell applications were improved. They can be used in the medical treatments such as stem cell therapy and tissue engineering applications, or in vitro models can be created to model diseases for drug testing. Among the stem cells, induced pluripotent stem cells (iPSCs) are the most preferred type as they are similar to embryonic stem cells (ESCs) with the ability of differentiation into any cell type and they can be generated from somatic cells easily. Improving stem cell differentiation processes is critical to eliminate current drawbacks of stem cells such as low efficiency and unwanted cell formation.

To learn more about these processes, transcriptomics technologies (mostly RNA-seq) can be used to quantify the change in expression levels of each transcript in different processes/timepoints and compare the cell states. Gene expression profiles of the individual cells can be measured by using single-cell RNA-seq (scRNA-seq). This technique provides novel analysis methods for investigating cellular heterogeneity, which are not possible with the conventional techniques. Trajectory inference, which is one of these analysis methods, finds transformation routes the cells can follow.

The aim of this thesis project is the analysis of time-series single-cell omics data from differentiation and reprogramming processes by using clustering and trajectory inference to examine these processes in detail and find the genes that play an important role during these processes. Targeting these genes may help to develop new approaches to make the processes more efficient and more successful, and allowing the stem cells to be more useful.

Key Words: single-cell RNA-seq, scRNAseq, clustering, trajectory inference, pseudotime analysis, stem cell differentiation, reprogramming.

ÖZET

Kök hücre, kendini yenileme ve diğer özel hücre türlerine farklılaşma yeteneklerine sahip hücre türüdür. Yaşam süreleri boyunca çoğalma ve kendilerini dönüştürme yeteneği, kök hücrelerin temel avantajlarıdır. Son teknolojilerin ortaya çıkmasıyla birlikte son on yılda kök hücre çalışmaları hızlanmış ve kök hücre uygulamaları iyileştirilmiştir. Kök hücreler, kök hücre tedavisi ve doku mühendisliği uygulamalarıyla tıbbi tedavide kullanılabilirler veya hastalıkları modellemek için ilaç testi ve tasarımıyla yönelik in vitro modeller oluşturulabilir. Kök hücre tipleri arasında indüklenmiş pluripotent kök hücreler (iPSC'ler), herhangi bir hücre tipine farklılaşma kabiliyeti ile embriyonik kök hücrelere (ESC) benzemesi ve somatik hücrelerden kolayca üretilebilmesi nedeniyle en çok tercih edilen türdür. Kök hücre farklılaştırma süreçlerinin iyileştirilmesi, düşük verimlilik ve istenmeyen hücre oluşumu gibi kök hücre kullanımının mevcut dezavantajlarını ortadan kaldırmak için kritik öneme sahiptir.

Bu süreçler hakkında daha fazla bilgi edinmek için, farklı süreç/zaman noktalarında her bir transkriptin değişen ifade seviyelerini ölçmek ve hücrelerin durumlarını yorumlamak için transkriptomik teknolojileri (çoğunlukla RNA dizileme) kullanılabilir. Her bir hücrenin gen ifade profilleri tek hücreli RNA dizileme (scRNA-seq) teknolojisi ile ölçülebilir. Bu teknik, geleneksel tekniklerle mümkün olmayan yeni analiz yöntemleri sağlayarak hücre heterojenliğinin incelenebilmesini mümkün kılar. Bu analiz yöntemlerinden biri olan yörünge çıkarımı, hücrelerin izleyebileceği farklılaşma veya yeniden programlama yollarını bulmayı sağlar.

Bu tez projesinin amacı, zamana bağlı tek hücreli omik verilerinin analizi ile farklılaşma veya yeniden programlama süreçlerini kümeleme ve yörünge çıkarımı yöntemleri ile ayrıntılı olarak incelemek ve bu başkalaşım süreçleri boyunca önemli rol oynayan genleri bulmaktır. Farklılaştırma ve yeniden programlama süreçlerindeki bu genler, süreçleri daha verimli ve daha başarılı hale getirmek için yeni yaklaşımlar geliştirmeye yardımcı olabilir ve bu da kök hücrelerin daha kullanışlı olmasını sağlar.

Anahtar Kelimeler: tek hücre RNA dizileme, kümeleme, yörünge çıkarımı, zaman-benzeri analizi, kök hücre farklılaşması, yeniden programlama.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude and appreciation to my supervisor Asst. Prof. Pınar Pir for her amazing support during my master's studies, this thesis would not be possible without her. I would like to thank Dr. Vladimir Y. Kiselev for his supervision during my internship in Sanger Institute (Cambridge, UK), which was an amazing opportunity to improve my skills on scRNAseq data analysis and visualization. I would like to thank Wellcome Trust, Erasmus+ Internship Mobility Program and TÜBİTAK for their generous funding.

I would like to thank my other professors in the department, their criticisms in the courses and presentations helped me a lot in improving my research skills. And I would like to thank my colleagues with whom I shared my offices. I felt motivated by their help, joy and energy.

Lastly, I would like to thank someone who encouraged and believed me to study master's degree and always made me feel better by being outstandingly cheerful and sending the bubbles of calmness and happiness.

TABLE of CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
LIST of ABBREVIATIONS	x
LIST of FIGURES	xii
LIST of TABLES	xiv
1. INTRODUCTION	1
2. BACKGROUND ASPECTS	3
2.1. Stem Cells	3
2.2. Cellular Differentiation Processes	6
2.3. Fibroblast Cells	7
2.4. Neuron Cells	7
2.5. Systems Biology and Omic Technologies	8
2.5.1. Transcriptomics	9
2.5.2. Single-cell Transcriptomics	11
2.6. Analysis of scRNA-seq data	13
3. MATERIAL AND METHODS	15
3.1. Datasets	15
3.1.1. Experimental Design	16
3.2. Toolboxes and Software	18
3.3. scRNA-seq Data Analysis	18
3.3.1. Clustering	19
3.3.2. Trajectory Inference	19
3.3.3. Gene Ontology (GO) Enrichment Analysis	22
4. RESULTS	24
4.1. Differentiation of mESC to MN	24
4.2. Differentiation of MEF to MN	27
4.3. Reprogramming of MEF to iPSC	30
5. DISCUSSION	37

5.1. Trajectory Inference	37
5.2. Stochasticity and Determinism of the Processes	38
5.2.1. Profiles of the Marker Genes	39
5.3. Analysis of the Driver Genes	45
5.4. Comparison of Previous Analyses to Current Analyses and Technical Challenges	51
6. CONCLUSION	54
REFERENCES	56
BIOGRAPHY	62
APPENDIX	63

LIST of ABBREVIATIONS

<u>Abbreviations</u>	<u>Descriptions</u>
3D	: Three Dimensional
A2S	: Ascorbic acid, 2i, SGC (inhibitor)
BAM	: Brn2, Ascl1, Myt1l
BH	: Benjamini-Hochberg
BP	: Biological Process
cDNAs	: Complementary DNAs
CSCs	: Cancer Stem Cells
DEGs	: Differentially Expressed Genes
DNA	: Deoxyribonucleic Acid
ECM	: Extracellular Matrix
ESCs	: Emrbyonic Stem Cells
FACS	: Fluorescence-activated Cell Sorting
FBS	: Fetal Bovine Serum
GEO	: Gene Expression Omnibus
GO	: Gene Ontology
GSK	: Glycogen Synthesis Kinase
IVT	: In-vitro Transcription
iPSCs	: Induced Pluripotent Stem Cells
LCM	: Laser Capture Microdissection
MAPK	: Mitogen-activated Protein Kinase
MEF	: Mouse Embryonic Fibroblasts
mESC	: Mouse Embryonic Stem Cells
MN	: Motor Neurons
mRNA	: Messenger RNA
NGS	: Next Generation Sequencing
OSKM	: Oct4, Sox2, Klf4, c-Myc
PCA	: Principal Component Analysis
PCR	: Polymerase Chain Reaction
RAM	: Random Access Memory

RA	: Retinoic Acid
RNA	: Ribonucleic Acid
RNA-seq	: RNA Sequencing
scRNA-seq	: Single-cell RNA Sequencing
Shh	: Sonic hedgehog gene
SP	: Standard Protocol
TFs	: Transcription Factors
TI	: Trajectory Inference
tSNE	: t-distributed Stochastic Neighbor Embedding
UMAP	: Uniform Manifold Approximation and Projection
UMIs	: Unique Molecular Identifiers

LIST of FIGURES

<u>Figure No:</u>	<u>Page</u>
2.1: Summary of stem cells in different potency levels.	5
2.2: An illustration of uses of iPSCs in research and medicine.	6
2.3: Representation of neuron and its parts.	8
2.4: Summary of the experimental steps in RNA-seq.	11
2.5: Overview of the development of scRNA-seq experiments.	12
2.6: Overview of scRNA-seq downstream analysis methods.	14
3.1: A representation of differentiation processes.	15
3.2: Summary of differentiation details of datasets.	16
3.3: Example of a trajectory inference result colored by real time periods.	21
3.4: Example of a trajectory inference result colored by pseudotime.	21
4.1: Clustering and trajectory inference results of mESC -> MN process.	25
4.2: Significant GO terms for cluster 0-10 of cells in mESC -> MN process.	26
4.3: Clustering and trajectory inference results of MEF -> MN process.	28
4.4: Significant GO terms for cluster 0-8 of cells in MEF -> MN process.	29
4.5: Clustering and trajectory inference results of cells from A2S-based protocol in MEF -> iPSC process.	31
4.6: Clustering and trajectory inference results of cells from FBS-based protocol in MEF -> iPSC process.	32
4.7: Significant GO terms for cluster 0-8 of cells from A2S-based protocol in MEF -> iPSC process.	34
4.8: Significant GO terms for cluster 0-8 of cells from FBS-based protocol in MEF -> iPSC process.	35
5.1: Violin plots of (A) pluripotency and (B) neuron marker genes from mESC -> MN process (Briggs et al, 3029 cells from three timepoints).	39
5.2: Violin plots of (A) MEF, (B) neuron and (C) pluripotency marker genes from MEF -> MN process (Treutlein et al, 405 cells from five timepoints, Ascl1 only and BAM experiments included).	40

5.3:	The percentage of cells that expressed marker genes at high levels in (A, B) mESC -> MN and (C, D) MEF -> MN processes.	41
5.4:	Expression levels of Gsk3b gene in mESC -> MN and MEF -> MN processes.	42
5.5:	Violin plots according to the expression levels of (A) MEF, (B) pluripotency markers in MEF -> iPSC process (Tran et al, 2642 cells from five timepoints, A2S-based protocol).	43
5.6:	Violin plots according to the expression levels of (A) MEF, (B) pluripotency markers in MEF -> iPSC process (Tran et al, 3127 cells from five timepoints, FBS-based protocol).	44
5.7:	The percentage of cells that expressed marker genes at high levels from (A, B) A2S-based protocol and (C, D) FBS-based protocol in MEF -> iPSC process.	45

LIST of TABLES

<u>Figure No:</u>	<u>Page</u>
3.1: Table of the datasets used in this thesis.	15
5.1: Correlation coefficients of the gene expression in pluripotency pathways with pseudotime values.	47
5.2: Correlation coefficients of the expression of driver genes in terms of pseudotime values in Ascl1 and BAM-triggered experiments.	50

1. INTRODUCTION

Stem cell research has gained a new perspective after milestone studies on cell type transformation processes were reported. Especially, the discovery of induced pluripotent stem cells (iPSCs) provided new opportunities to study differentiation to any cell type, such studies were not possible with embryonic stem cells (ESCs) due to the ethical issues [K Takahashi, 2006]. iPSCs have similar characteristics to ESCs and have the ability to differentiate into any cell type due to their pluripotency, hence use of iPSC can lead to development of new treatment therapies and discovery of new drug targets.

Transcriptomics technologies are the techniques used to study all RNA transcripts expressed in a biological sample. Transcriptomics can be used in a variety of research areas, for example, comparison of gene expression levels between disease and healthy states shows which genes change significantly in disease state and these genes can be studied in further detail. Currently, most common technique for transcriptomics is RNA sequencing (RNA-seq). Standard RNA-seq protocol is applied to a population of cells, and average gene expression values are retrieved for each gene, whereas single-cell RNA-seq (scRNA-seq) is applied to the individual cells and gene expression values for each gene can be retrieved for each cell. Therefore, single-cell transcriptomics provides gene expression levels in single-cell resolution and allows novel analysis methods which are not possible with standard RNA-seq [Lowe et al., 2017].

One of the analysis methods for scRNA-seq is trajectory inference, which is also known as ‘pseudotime analysis’. In trajectory inference methods, scRNA-seq data is sorted computationally according to the gradual transition of their transcriptomes in the progress route of interest. Cells evolve at different speeds in a process such as cell differentiation or development. Cells collected at any particular time point during these kind of processes contain cells representing different stages of these processes. Using single-cell RNA-seq data, an ordered sequence of cells can be constructed to describe the gradual transition of the single-cell transcriptome. scRNA-seq changes the understanding of diseases or biological processes, and it reveals cell-to-cell heterogeneity within a tissue at very high resolution [Poirion et al., 2016, Saelens et al., 2019].

Reprogramming and differentiation of the iPSCs takes place stochastically, hence there are no specific events that are determined to occur, instead a variety of events with different probabilities may occur during the transformation [Stumpf et al., 2017]. Therefore, to observe the stochastic changes, integration of different types of data at single-cell resolution via bioinformatic analysis tools are necessary. Analyses will reveal how those cells with high heterogeneity transform their phenotypes, which events take place, and which transformation path or trajectory is selected during the transformation.

The aim of this thesis project is the analysis of time-series single-cell omics data collected from differentiation and reprogramming processes by using clustering and trajectory inference methods to investigate these processes and find the genes that play an important role during the transformations.

2. BACKGROUND ASPECTS

2.1. Stem Cells

Stem cells are defined as immature cells with the capabilities of self-renewal, and differentiation to other mature and specialized cell types [Chagastelles and Nardi, 2011]. Stem cells can proliferate throughout their lifespan and transform themselves to other cell types by changing their molecular profiles. Stem cells have the potential to be used as a source of cells for medical treatment, including the applications such as stem cell therapy and tissue engineering. Additionally, stem cells are used to create *in vitro* models to test hypotheses such as use of disease models for drug design and testing [Elbuluk et al., 2017, Rowe and Daley, 2019]. These unique properties of stem cells and the emergence of advanced techniques for stem cell studies make them evolve into one of the most remarkable research areas to be focused on.

Stem cells can be classified according to their ability to differentiate into specialized cells (potency) and their origin. According to their potency, stem cells can be classified into five groups as shown in Figure 2.1 adapted from Zhang et al [Zhang et al., 2009]:

- Totipotent stem cells
- Pluripotent stem cells
- Multipotent stem cells
- Oligopotent stem cells
- Unipotent stem cells

Totipotent stem cells are the cells that have ability to differentiate into any kind of cells, which means the ability of a cell to create an entire organism. Zygotes and spores can be the examples of totipotent stem cells [Mitalipov and Wolf, 2009]. Pluripotent stem cells are the cells that have the capacity to differentiate into three germ layers which are ectoderm, mesoderm, and endoderm [Bang and Carpenter, 2009]. From these layers, cells can differentiate into any cell type. Embryonic stem cells and induced pluripotent stem cells are the examples of pluripotent stem cells. Multipotent stem cells are the cells which are able to differentiate into fewer cell types.

Hematopoietic stem cells and mesenchymal stem cells are the examples of multipotent stem cells. Oligopotent stem cells are the cells that have the ability to differentiate into even fewer cell types. Examples for this type of stem cells are myeloid stem cells and lymphoids. Unipotent stem cells are the cells that can produce only one type of cells. Muscle stem cells are the examples for the unipotent stem cells.

According to the origin of the cells, stem cells can be classified into four main groups [Barky et al., 2017]:

- Embryonic stem cells
- Induced pluripotent stem cells
- Adult (somatic) stem cells
- Cancer stem cells

Embryonic stem cells (ESCs) are the type of cells that are formed from inner cell mass of blastocysts and show pluripotent properties, meaning that they can differentiate into any cell type. They are also able to proliferate without any limitations [Zakrzewski et al., 2019]. With these properties, embryonic stem cells can be seen as the most useful stem cell type. Induced pluripotent stem cells (iPSCs) have properties very similar to the embryonic stem cells. The difference between them is that ESCs are derived from blastocysts whereas iPSCs are derived from adult cells by overexpression of four transcription factors, Oct4, Sox2, Klf4 and Myc, which are called Yamanaka factors [K Takahashi, 2006, Shi et al., 2017]. Adult (somatic) stem cells are multipotent cells that have the ability to self-renew in particular tissues and organs. They are found in many organs such as bone marrow, brain, teeth [Caglayan et al., 2019]. Cancer stem cells (CSCs) are sparse subsets of cancer cells sharing similar properties with normal stem cells.

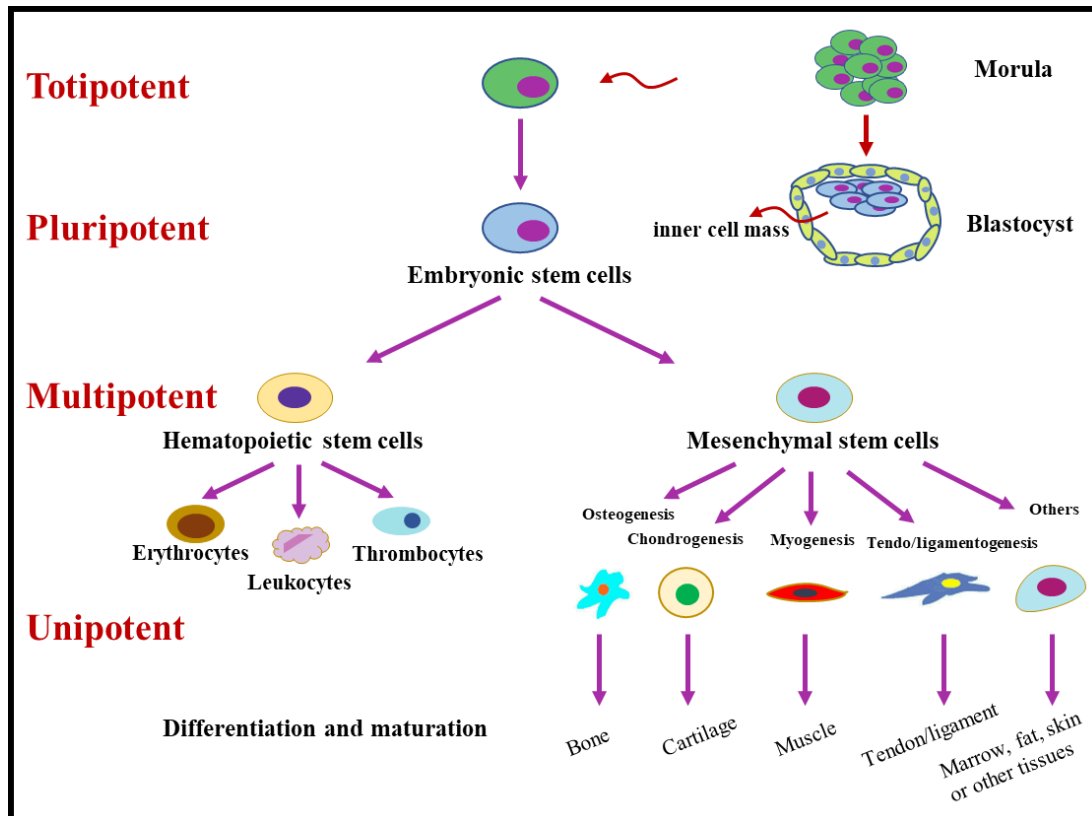


Figure 2.1: Summary of stem cells in different potency levels.

Between those four types of stem cells, iPSCs are the most preferable type of stem cells due to having very similar characteristics to ESCs such as the capability of differentiating into any cell type and the easy generation from any cell type with a cocktail of transcription factors without using any embryos. For instance, iPSCs produced from the somatic cells of the patient can be used by producing healthy specialized cells for autologous therapy, or they can be used to create 3D cell cultures which are called organoids for disease modeling to test novel therapeutic approaches as summarized in Figure 2.2 adapted from Rowe and Daley et al. [Rowe and Daley, 2019]. Additionally, iPSCs can be generated for treatment of injured and degenerated tissues. Although there are some drawbacks of the iPSCs usage such as the uncontrolled integration of retroviral DNA into the host genome and the association of OSKM factors with the development of the multiple tumors in oncogenetics, iPSCs are still advantageous for being used in stem cell therapy [Medvedev et al., 2010].

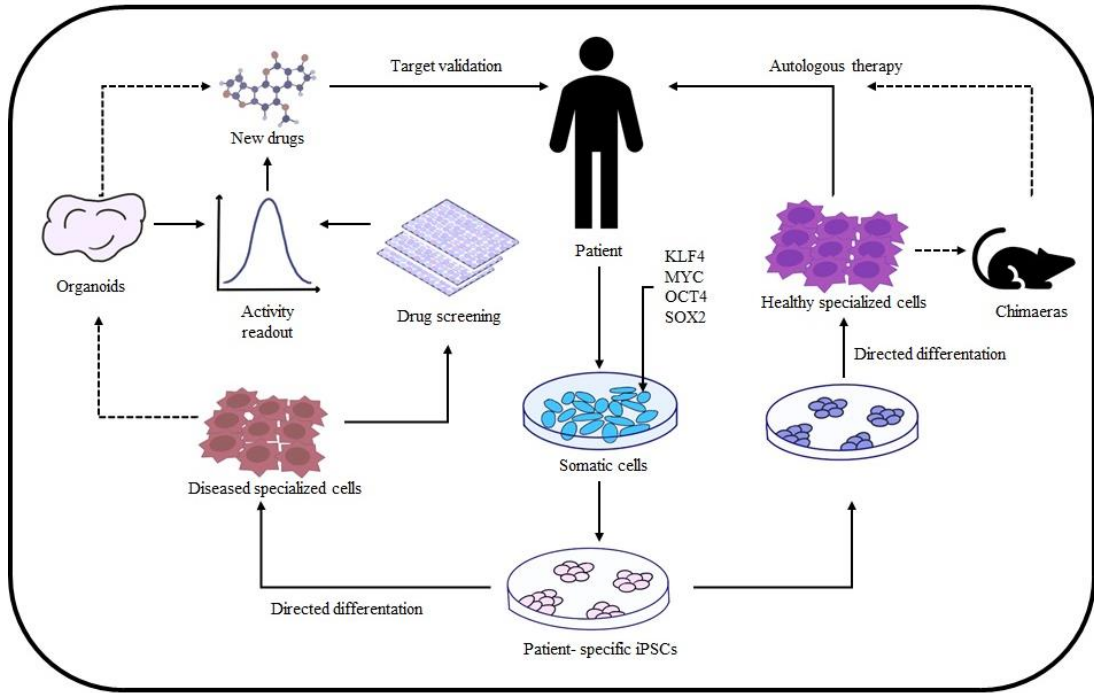


Figure 2.2: An illustration of uses of iPSCs in research and medicine.

2.2. Cellular Differentiation Processes

Cell differentiation is a general term for the processes through which the cells change their function, phenotype and the other characteristics to a specialized cell type by a series of signaling events. Since each cell type expresses a different set of genes, most obvious change during cell differentiation is the regulation in gene expression of the cells. Transcription factors, which are proteins that can affect gene expression, are the main elements for this regulation. Although cell differentiation is an essential process for embryological development, it is also vital for processes involving cell renewal in the body [Zhang and Wolynes, 2014]. With the innovations in stem cell technologies, it has become possible to differentiate cells *in vitro* [Evans and Kaufman, 1981, Reubinoff et al., 2000].

As mentioned above and as in the formation of iPSCs, stem cells differentiate by being exposed to overexpression of some transcription factors. With exposure to high levels of reprogramming-mediating transcription factors, cells are transformed into the stem cells; whereas stem cells are transformed into the other cell types via differentiation-mediating transcription factors [Mah et al., 2011]. Transformations can also take place without stem cell phenotype formation at any stage of the process.

These processes are called ‘direct differentiation’ and they are carried out by inducing a set of process-specific transcription factors. Differentiation of mouse embryonic fibroblasts (MEFs) to neurons is an example of direct differentiation [Vierbuchen et al., 2010].

2.3. Fibroblast Cells

Fibroblast is the type of cell that produces collagen and extracellular matrix, as well as being the primary element for connective tissue in animals [Murray et al., 2009]. Fibroblasts were mainly discovered in between 1906-1921 by observing their characteristic manner and the response, and its fundamental properties could be compiled in 1926 [Carrel and Ebeling, 1926]. The use of fibroblasts specifically in stem cell studies was started by the discovery of OSKM factors (as known as Yamanaka factors) [Takahashi and Yamanaka, 2006]. In addition to physiological and pathological tissue fibrosis, fibroblasts have crucial functions in ECM maintenance, reabsorption, cancer progression, inflammation, angiogenesis and wound healing [Kendall and Feghali-Bostwick, 2014]. CD90, DDR2, FSP1, Sca1, fibronectin, vimentin, collagen types I/III, FAP and α SMA are some of the fibroblast markers [Ivey and Tallquist, 2016].

In the last decade, it was discovered that fibroblasts can directly differentiate into neurons by using a set of defined factors. This discovery led fibroblasts to be used extensively in neuroscience [Vierbuchen et al., 2010]. Especially mouse embryonic fibroblasts (MEFs) gained popularity because of having robust systems, fast growth rate, easy approachability and multiplexing, which is large number of simultaneous experiments [Qiu et al., 2016].

2.4. Neuron Cells

Neuron is a specialized cell type that can transmit electrical impulses. It is a primary component of the nervous system and consists of three parts, which are dendrites, soma (cell body) and axon. Dendrite receives chemical signals from axon terminal of the neighboring neurons, converts the signals to electrical impulses (action potential) and transfers them to the axon. Axons then transmit the electrical impulse to

other neurons [Darnell et al., 2000]. Neurons generally have multiple dendrites. Cell body is the part where all membranes and neuronal proteins are synthesized. Basic representation of a neuron is shown in Figure 2.3 adapted from Thanapitak [Thanapitak, 2012].

Due to the current ineffectiveness in treatment of neurological diseases, there is a need for better understanding of the processes taking place inside neuron cells. In order to achieve this, in vitro differentiation of cells can be used to generate neurons for purposes such as revealing the differentiation mechanisms, drug development and disease modeling. Production of any subtypes of neurons is possible thanks to the pluripotency of the stem cells [Hou et al., 2017].

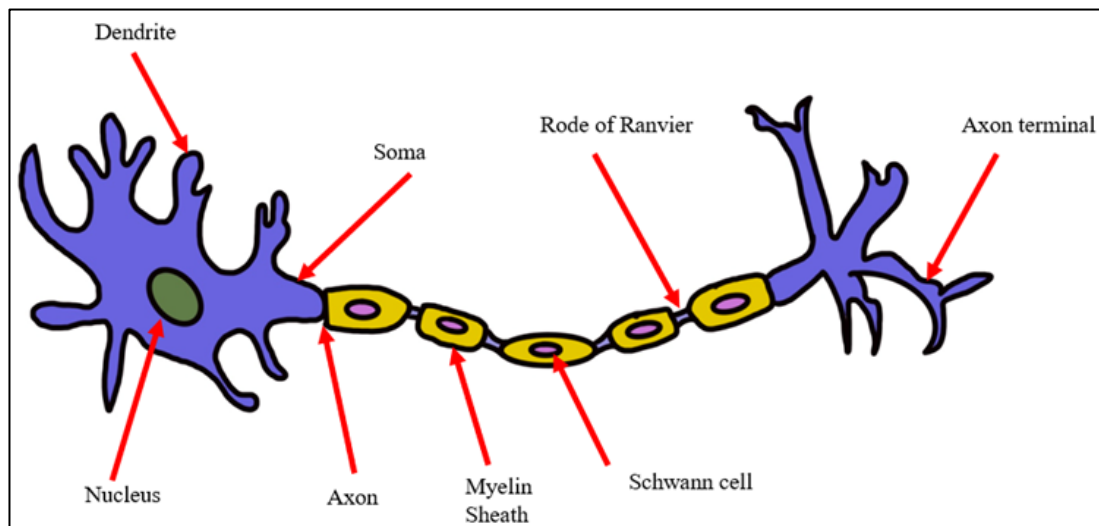


Figure 2.3: Representation of neuron and its parts.

2.5. Systems Biology and Omic Technologies

Living organisms are quite complex creatures. They include a large number of pathways, reactions and processes and all of them take place in every millisecond inside the living body. The tools to observe them comprehensively are insufficient for the moment. In order to understand the living systems better and to get an overview of the transformations taking place inside them, “systems biology” approaches are used, which allows a holistic view of the components of the systems and interactions among them.

“Omics” term is a comprehensive quantification and characterization of a group of biomolecules. “-omics” suffix is added to the name of the target of interest. Genomics which focus on structure, function and characterization of the genomes; transcriptomics which focus on the concentration of the total of RNA molecules; proteomics which focus on the amount, biochemistry, function and the structure of the proteins, epigenomics which focus on epigenetic modifications such as DNA methylation and histone modifications are some of the examples for omics. Using omics technologies, high-throughput data is generated from the cases to be examined, analyzed in detail and the results from the analyses are interpreted considering the case-specific hypotheses [Vailati-Riboni et al., 2017]. For each omic technology, data is produced by experimental and quantification steps in upstream processes (low level analyses); and biological results are derived in downstream processes (high level analyses). The omics data allows systems biology approach to be applied to the system of interest.

2.5.1. Transcriptomics

Transcriptome is total of RNA molecules transcribed in a single cell or a population of cells. It is a snapshot of total transcripts in a particular timepoint, only 2% of which are mRNA molecules. Transcriptome mostly refers to mRNAs due to the fact that transcript concentrations (mRNA levels) can mainly determine the protein levels and quantification of mRNA levels are often used for the inference of protein levels in the cells. Transcriptome may also refer to all RNAs, hence its description is based on the design of the experiment. By using transcriptomics technologies, it is possible to measure the transcriptome of any RNA-containing medium. The main purpose of the transcriptomics is to detect the genes whose expression levels are significantly different between conditions (which will be referred to as differentially expressed genes (DEGs) for the rest of this thesis) to learn more about activities of the pathways of interest. In addition to that, transcriptomics is used to classify and quantify other types of transcripts (non-coding RNAs, small RNAs etc) in the sample and find the transcriptional landscape of the genes (e.g. to study alternative splicing or post-transcriptional modifications). Most common techniques used for transcriptomics are DNA microarrays and RNA sequencing (RNA-seq) [Lowe et al., 2017].

2.5.1.1. RNA Sequencing

RNA sequencing (RNA-seq) is a technique to measure the quantity of RNA molecules (mostly mRNA molecules as mentioned in Transcriptomics section) in a sample of interest by using next generation sequencing (NGS) technology. 454 sequencing, a massively parallel pyrosequencing technique, was the first effective NGS strategy that received broad acceptance by the community [Margulies et al., 2005]. 454 sequencing was superseded by Illumina sequencing, and currently it is the most used RNA-seq technology [Shendure et al., 2005, Weber, 2015]. Having many advantages compared to the microarray technology, RNA-seq is considered as a revolutionary technology for transcriptomics. Microarray is limited to the probe sequences on the chip, whereas RNA-seq does not have such limitations and allows both quantification of known transcripts and discovery of novel transcripts. RNA-seq has a wider dynamic range, higher accuracy and higher sensitivity, can detect higher number of DEGs and transcripts including the ones with low abundance [Rai et al., 2018]. In RNA-seq experiments, sequencing depth or coverage (number of unique sequenced reads in reconstructed sequence) is a critical point to consider in detail [Sims et al., 2014].

For RNA-seq, the cells are sampled from the tissue of interest, and RNA molecules of the cells are extracted. And these molecules are fragmented, reverse transcribed to make the cDNA molecules available for sequencing. Then, adapters are ligated to one or both sides of the fragments and the fragments are amplified. The fragments are loaded to sequencing machine and the sequence information of the fragments are determined. Afterwards, the sequence information is processed to determine which genes the fragments belong to by using some specific softwares. And at the end, gene expression values are quantified for each sample. The steps of RNA-seq is summarized in Figure 2.4 adapted from Berge et al [Berge et al., 2019].

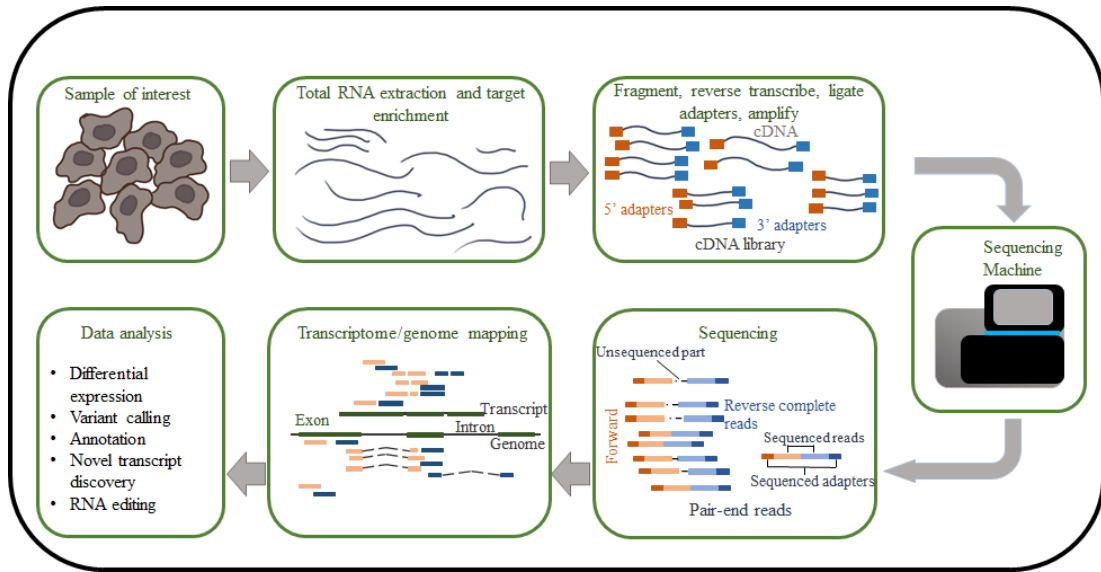


Figure 2.4: Summary of the experimental steps in RNA-seq.

2.5.2. Single-cell Transcriptomics

In majority of the studies, transcriptomics analysis is based on a population of cells. Even if a cell seems to be identical to the neighboring cells, gene expression profiles of those cells are not the same due to the nature of stochasticity in gene expression. This phenomenon is called cellular heterogeneity [Altschuler and Wu, 2010]. Transcriptomic technologies applied on a population of cells (which is called bulk RNA-seq) produce average transcriptome of those cells. Bulk RNA-seq technique is not sufficient to observe the intercellular differences in gene expression. By using single-cell transcriptomics, it is possible to study the transcriptional profile of individual cells at single-cell resolution. First single-cell RNA sequencing (scRNA-seq) protocol was published in 2009 by Fuchou Tang and his colleagues. Since then, more advanced single-cell sequencing methods were developed and the amount of sequenced single cells in a single experiment have increased dramatically from 1 cell to more than 1 million cells [Cao et al., 2019, Tang et al., 2009]. Isolation of single cells can be classified into two groups which are well-based and droplet-based methods. Whereas well-based methods are more controlled, more flexible, lower-throughput and more expensive; droplet-based methods are more efficient, high-throughput and cheaper [Ziegenhain et al., 2018]. By using these isolation methods as a basis, a large number of single-cell sequencing methods are developed [Chen et al., 2019]. An overview of the single-cell isolation methods and number of cells that can

be sequenced by each sequencing method was given in Figure 2.5 adapted from Svensson et al [Svensson et al., 2018].

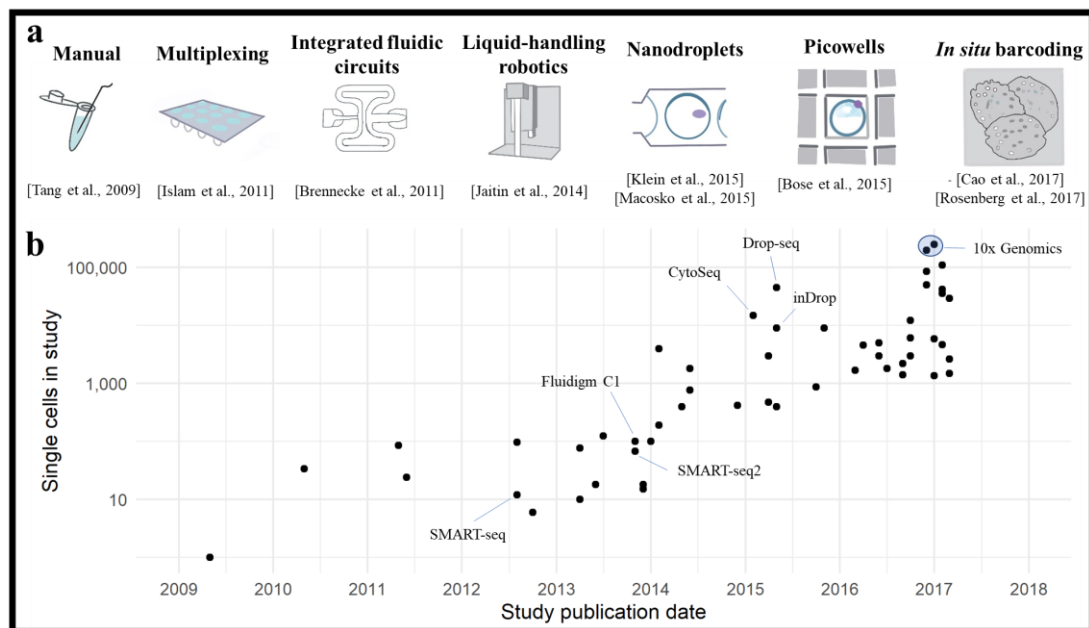


Figure 2.5: Overview of the development of scRNA-seq experiments.

To generate scRNA-seq libraries, first step is to isolate single cells or single nuclei from a bulk of cells. Each cell needs to be dissociated from each other and from the extracellular matrix (ECM). Cell isolation can be carried out by well-based or droplet-based methods. In well-based methods, single cells are dissociated and placed into the wells by using fluorescence-activated cell sorting (FACS) or microfluidic systems. In droplet-based methods, cells are encapsulated inside oil droplets. Isolation method should be chosen carefully to prevent unintended changes in expression of some genes during the isolation process. In addition to those methods, laser capture microdissection or patch clamp can be used if identification of the connection between spatial coordinates and transcriptome profile is aimed. After RNA isolation, complementary DNAs (cDNAs) are generated with the reverse transcription of mRNAs by polymerase chain reaction (PCR) or in vitro transcription (IVT), amplification is done and the sequencing libraries are created. In the step of generation of cDNAs, it is critical to use cell-specific barcodes in the primers. These cell-specific barcodes are the random nucleotide sequences and called unique molecular identifiers (UMI). By using UMIs, reads can be annotated with the cell they originate from, and the bias and noise introduced by PCR or IVT can be removed. Final steps are

sequencing the libraries and retrieving the reads from the fragments in the sample [Ziegenhain et al., 2018]. Since the first protocol for scRNA-seq, a variety of different protocols were developed, each with their own advantages [Chen et al., 2019].

After the experimental steps are completed, reads are processed with a set of software tools to create the count matrix which includes expression levels for each gene and each individual cell. In order to quantify the expression values, reads are aligned to the reference genome of the organism being studied. For this purpose, pseudoalignment algorithms which prioritizes speed over full-length alignment of the reads can be used. Finally, a count matrix is created for downstream analyses. Difference of the count matrix in single-cell sequencing to that of bulk RNAseq is that each sample is an individual cell instead of average values of a bulk sample from the cell culture or patient.

2.6. Analysis of scRNA-seq data

As mentioned in “Transcriptomics” section, where mostly bulk RNA-seq was described, the main purpose of transcriptomics experiments is to find DEGs between two conditions. But, with the transcriptomics data in single-cell resolution, it is possible to analyze the cells in addition to genes. In cell level analyses, cellular heterogeneity is the key property. Cell level analyses can be separated into two groups, which are clustering and trajectory inference (pseudotime analysis). Single-cell analysis methods are summarized in Figure 2.6 [Luecken and Theis, 2019].

By using clustering algorithms, groups of cells with similar gene expression levels and cell types can be identified based on the transcriptome profiles of the cells in an unsupervised way. K-means clustering, graph-based (community-detection-based) clustering, hierarchical clustering and density-based clustering are the main methods for clustering [Zhang et al., 2020]. Clustering problem is not an easy task to solve and all of the methods have their own advantages, hereby there is no superior method to the others [Kiselev et al., 2019].

Trajectory inference is another analysis method for scRNA-seq data. In trajectory inference (TI) methods, a “pseudotime” value is assigned to each cell to indicate their degree of progress in the process of interest by using cellular heterogeneity [Chen et al., 2019]. Afterwards, cells are ordered according to their

pseudotime values in the reduced-dimension space and based on the coordinates of the cells in the space, trajectories are created. Developmental or differentiation processes can be defined and analyzed via these trajectory results. This type of analysis is applied for studying developmental or other time-course processes.

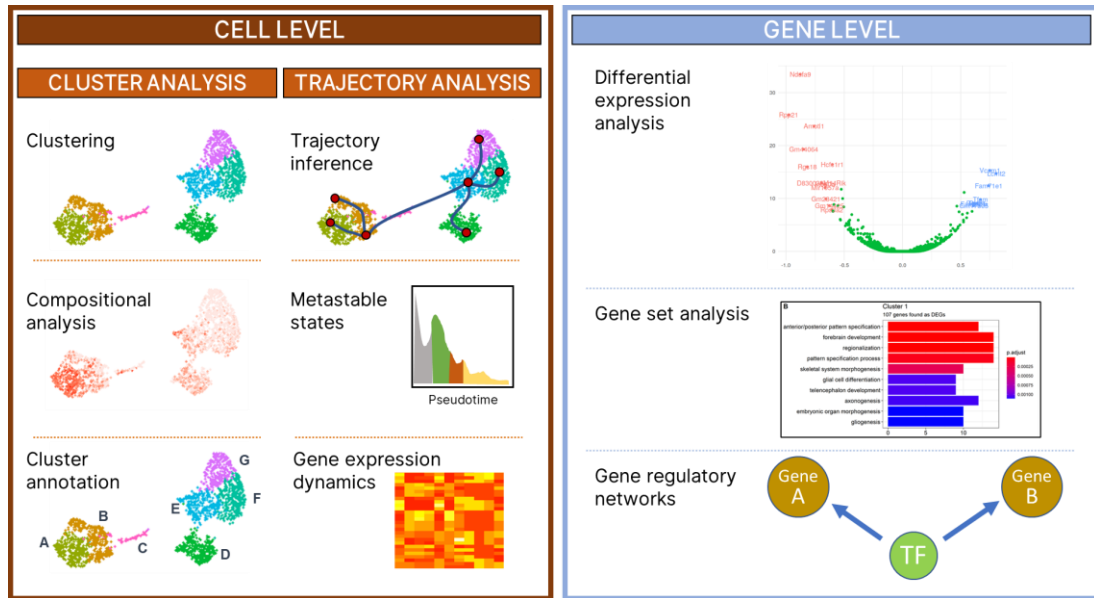


Figure 2.6: Overview of scRNA-seq downstream analysis methods.

3. MATERIAL AND METHODS

3.1. Datasets

In the scope of this thesis, three scRNA-seq datasets were analysed. Overview of the datasets is shown in Table 3.1 and representation of differentiation processes of interest is shown in Figure 3.1.

Table 3.1: Table of the datasets used in this thesis.

	Accession Number	Organism	Source of Differentiation	Target of Differentiation	Number of Cells	Reference
1	GSE97391	Mouse	mESC	Neuron	3023 cells	[Briggs et al., 2017]
2	GSE67310	Mouse	MEF	Neuron	405 cells	[Treutlein et al., 2016]
3	GSE108222	Mouse	MEF	iPSC	2642 cells (A2S) 3127 cells (FBS)	[Tran et al., 2019]

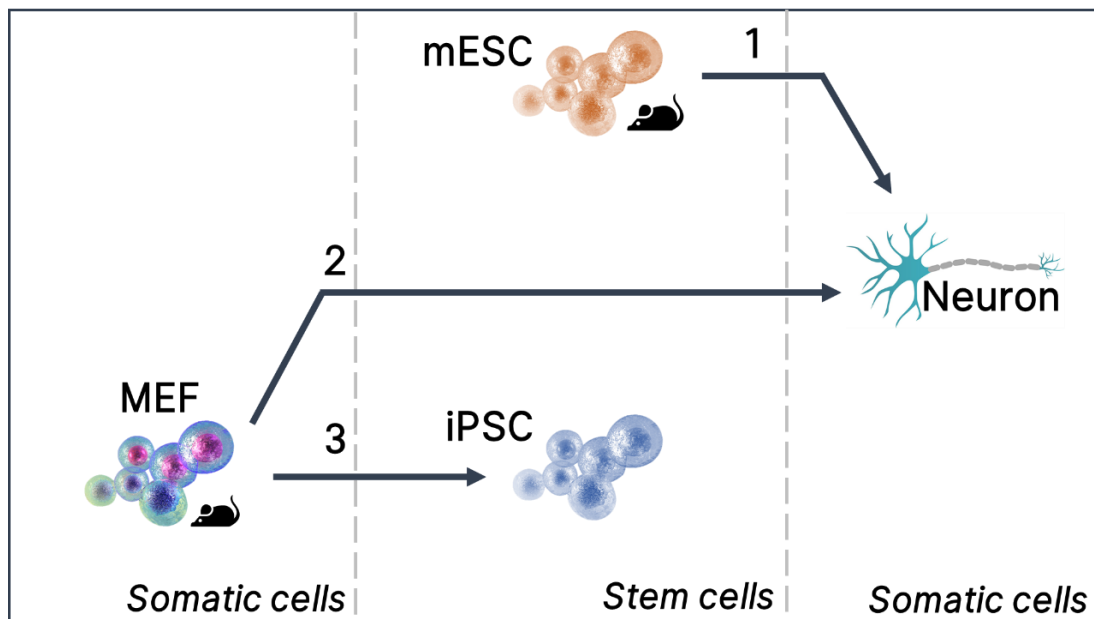


Figure 3.1: A representation of differentiation processes.

The datasets were downloaded in ‘count matrix’ format from Gene Expression Omnibus (GEO) platform provided by National Center for Biotechnology Information (NCBI). The first dataset was published by James Alexander Briggs and his colleagues with the accession number of GSE97391 in 2017. The dataset is composed of scRNA data from the differentiation process of mouse embryonic stem cells (mESCs) into (spinal) motor neurons (MNs) by applying standard and direct protocols. In this thesis study, data retrieved from direct protocol was used due to having much higher efficiency of differentiation. Data from previously published study of the same research group was used for day 0 reference [Klein et al., 2015]. The second dataset was published by Barbara Treutlein and her colleagues with the accession number of GSE67310 in 2016. It includes scRNA data from direct differentiation of MEFs to (motor) neurons. The third dataset was published by Khoa A. Tran and his colleagues with the accession number of GSE108222 in 2019. This dataset includes scRNA data from reprogramming process of mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs). Cells were differentiated by using the experimental design explained in Chapter 3.1.1.

3.1.1. Experimental Design

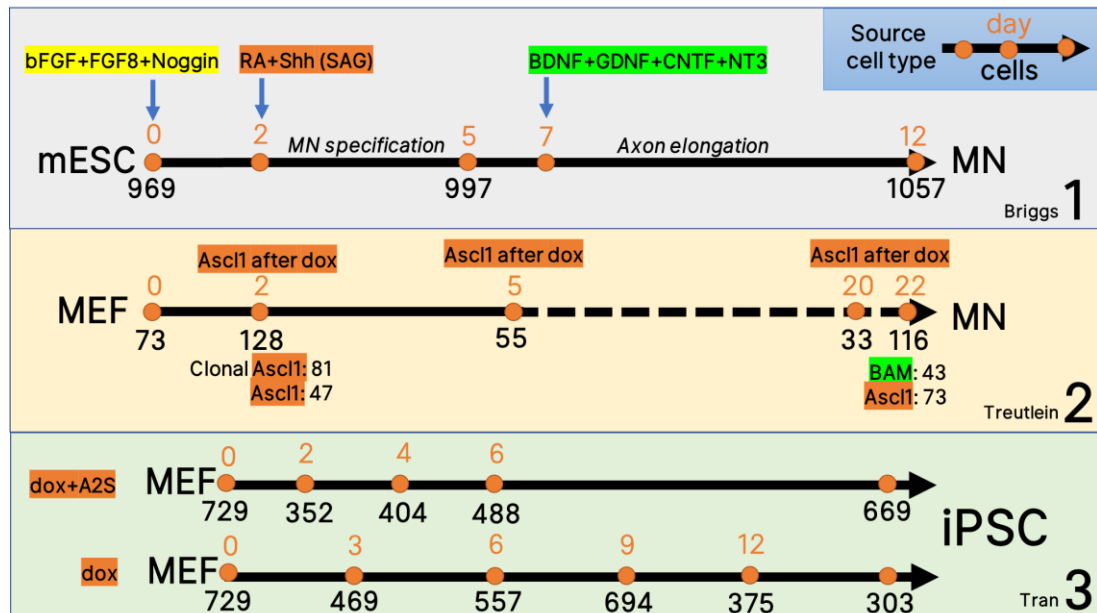


Figure 3.2: Summary of differentiation details of datasets.

In the study reported by Briggs et al, mESC to MN differentiation was carried out, cells were transformed using both standard and also direct protocols, but standard protocol was studied in the scope of this thesis. In standard protocol, mESCs were seeded for twenty-four hours before the differentiation begins. After twenty-four hours, mESCs were primed with bFGF, FGF8 and Noggin. In day 2, the cells were infected with retinoic acid and SAG (Shh agonist) for motor neuron specification process step. In day 7, BDNF, GDNF, NTF and NT3 were added for axon elongation step. During this experiment, 969 cells were collected at day 0, 997 cells were collected at day 5, 1057 cells were collected at day 12.

In the study of direct differentiation from MEF to MN by Treutlein et al, MEFs were differentiated separately targeting designated time points. Day 0 cells did not undergo any differentiation processes. Day 2, day 5 and day 20 cells were infected with Ascl1 after dox induction and collected at the specified time points. A subset of day 22 cells were infected with Ascl1 only after dox induction and collected after 22 days, the rest of the day 22 cells were infected with Brn2, Ascl1 and Myt1l after dox induction and again collected after 22 days. Details about these datasets were summarized in Figure 3.2.

In the study of MEF to iPSC reprogramming by Tran et al, MEFs were treated with doxycycline to induce OSKM (Oct4, Sox2, Klf4, Myc) genes and irradiated MEFs were added into the plate. While the cells that were reprogrammed with FBS-based protocol did not have any additional treatment, cells that were reprogrammed with A2S-based protocol were first treated with ascorbic acid and histone methyltransferase inhibitor named SGC0946 in the starting time point. Then, GSK3 (glycogen synthesis kinase) inhibitor named CHIR-9901 and mitogen-activated protein kinase (MAPK) inhibitor named PD-0325901 (2i) were added to the plate after 12 hours. While the cells that were reprogrammed with FBS-based protocol were sampled in days 3, 6, 9 and 12; cells that were reprogrammed with A2S-based protocol were sampled in days 2, 4 and 6. Mouse embryonic stem cells (mESCs) were untreated in the FBS-based protocol, however treated with inhibitors and 2i for 2 days in the A2S-based protocol.

3.2. Toolboxes and Software

All the analyses were run in R v4.0.2 via RStudio v1.2 using the packages that are described below. Dimension reduction and clustering were done using Seurat v3.2.2. Trajectory inference was done using Monocle3 v0.2.0. GO enrichment analysis was done using clusterProfiler v3.16.1. Spearman correlation for each dataset was calculated using base R function “cor”. Visual outputs (clustering, trajectory inference and violin plots) were produced in Seurat and Monocle3 packages. All of the processes explained above were done in an all-in-one computer with Windows 10 operating system, 8GB of RAM and Intel Core i5-6200U processor funded by the TÜBİTAK ARDEB 1001 project 116S388.

3.3. scRNA-seq Data Analysis

In the studies that the datasets were produced, they analyzed their data according to their questions of interest. Briggs et al analyzed their dataset to compare standard and direct differentiation protocols. They used local density gradient clustering and they visualized their dataset by using tSNE [Briggs et al., 2017] . They inferred trajectory on tSNE plots by using SPRING package [Weinreb et al., 2017]. Treutlein et al analyzed their data to focus on intermediate states of the direct differentiation. They applied hierarchical clustering to the genes and they used older version of Monocle to infer the trajectory [Treutlein et al., 2016]. Tran et al focused to modify the reprogramming protocol by using a combination of epigenetic modifiers. They used older version of Monocle for tSNE clustering and pseudotime trajectory [Tran et al., 2019].

In this thesis, a streamlined workflow was applied to all datasets for scRNA-seq data analysis. The analyses were carried out using two main downstream analysis methods, which are trajectory inference and gene ontology (GO) enrichment analysis. Prior to these analyses cells were clustered according to their gene expression profiles.

3.3.1. Clustering

Clustering is the process of finding subgroups of the cells based on the given data. In the scope of single-cell RNAseq analysis, clustering is done by using gene expression values. Clustering is a crucial step for single-cell RNAseq analysis. Since scRNA-seq datasets include enormous number of samples (single cells), the most convenient approach that can be applied to get an overview of the datasets is to check different (distinct) populations that the single cells originate from.

In the scope of this thesis, Seurat package was used to obtain basic information about single-cell datasets. With Seurat, cells can be examined according to their gene expression levels and their distribution in a low-dimensional space, and with the clustering algorithms, it can be determined how the cells in the dataset are grouped and how many groups they are divided into. The distribution of cells in the datasets in low-dimensional space was done using the UMAP (Uniform Manifold Approximation and Projection) method, and the clustering process was done using the Leiden algorithm [Traag et al., 2019]. When compared to other dimension reduction methods, UMAP produces meaningful representations of scRNA-seq data in reduced dimensions in comparison to tSNE, and UMAP results are more reproducible [Becht et al., 2019]. These advantages make UMAP more useful in visualization of transcriptome profiles of single cells in reduced dimensions. In UMAP representation, cells located close to each other show high resemblance and cells located away from each other show high difference according to the gene expression profiles. After the clustering process, marker genes that identify the clusters were detected and GO analysis was performed to show the biological processes that the marker genes of clusters were related to. Thus, the cells in each cluster were assigned a cell type based on the transcriptome characteristics of the clusters.

The genes which were specifically examined for the cellular processes of interest were shown with violin plots according to the time points and clusters in the datasets.

3.3.2. Trajectory Inference

Trajectory inference is one of the techniques used for inspecting single-cell data retrieved from time-series processes such as development and differentiation. This

technique shows the different molecular routes of cell transformation that took place and significance of these routes for the process of interest. Because each cell has a different molecular profile, processes follow a different route in each cell and this causes asynchrony, which may lead to biases in the analysis of regulatory changes. To solve the asynchrony problem, cells are sorted according to “pseudotime” values which are assigned to each cell by using cellular heterogeneity. Pseudotime indicates where the profile of a single cell fits in timeline of the process of interest. Trajectory inference analysis finds the expression changes as a function of pseudotime (occurred along the trajectory).

First step of trajectory inference analysis is retrieval of dataset of interest and its metadata, followed by loading the data to R environment and then data rearrangement is done to organize the data in required format by the software tools. Quality control is applied to remove the low-quality cells with very few detected genes, and also the genes that are expressed only in a few cells, and also the cells that have high percentage of mitochondrial gene expression are removed, finally *cell_data_set* object that Monocle3 uses to store single-cell expression data is created.

After creating *cell_data_set* object, data is preprocessed by normalizing it via log transformation and multiplication with a size factor and finally applying principal component analysis (PCA) which creates a lower dimensional space for the next step. Then, dimension reduction is applied by using t-stochastic neighborhood embedding (tSNE) or uniform manifold approximation and projection (UMAP), the latter is more preferable in single-cell analysis as explain in the previous section. Unsupervised clustering is applied after dimension reduction step. Afterwards, construction of trajectory is initiated: Principal graph is learned from reduced dimension space (by using reversed graph embedding), and cells are ordered (pseudotime values are calculated) after selecting the starting cells. Order of the cells is determined according to their location through the developmental program, which are called pseudotime values.

Once all the steps are completed, inferred trajectory is plotted as shown in Figure 3.3. Black line corresponds to the structure of the principal graph. Circles correspond to the branch points in the trajectory. Gray circles show the different cell fates, and black circles show the branch nodes which are the points where branching starts. Trajectory can also be plotted by coloring according to the pseudotime values as shown in Figure 3.4.

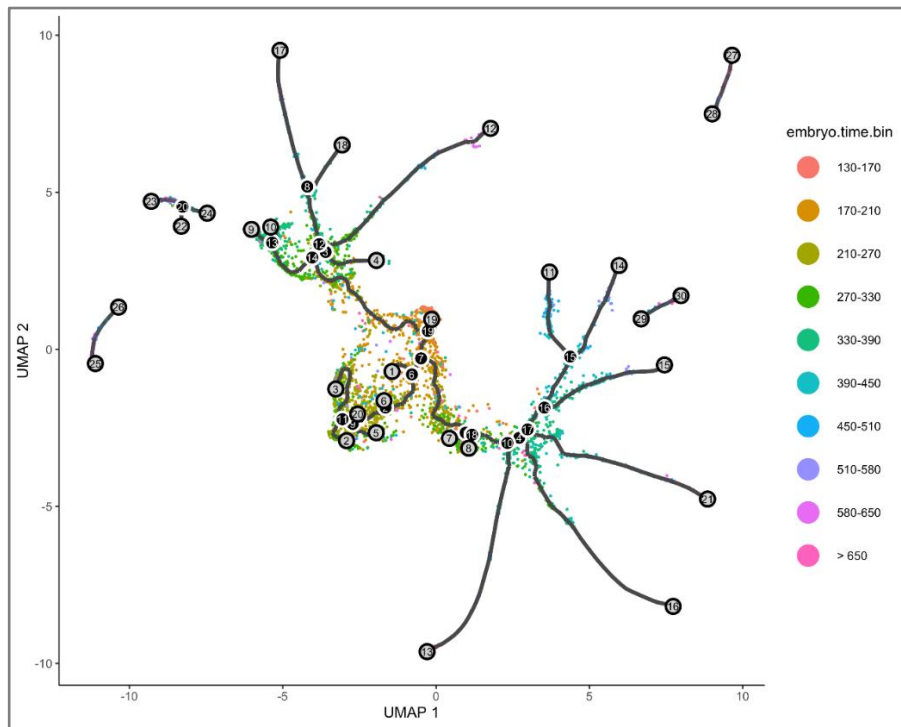


Figure 3.3: Example of a trajectory inference result colored by real time periods.

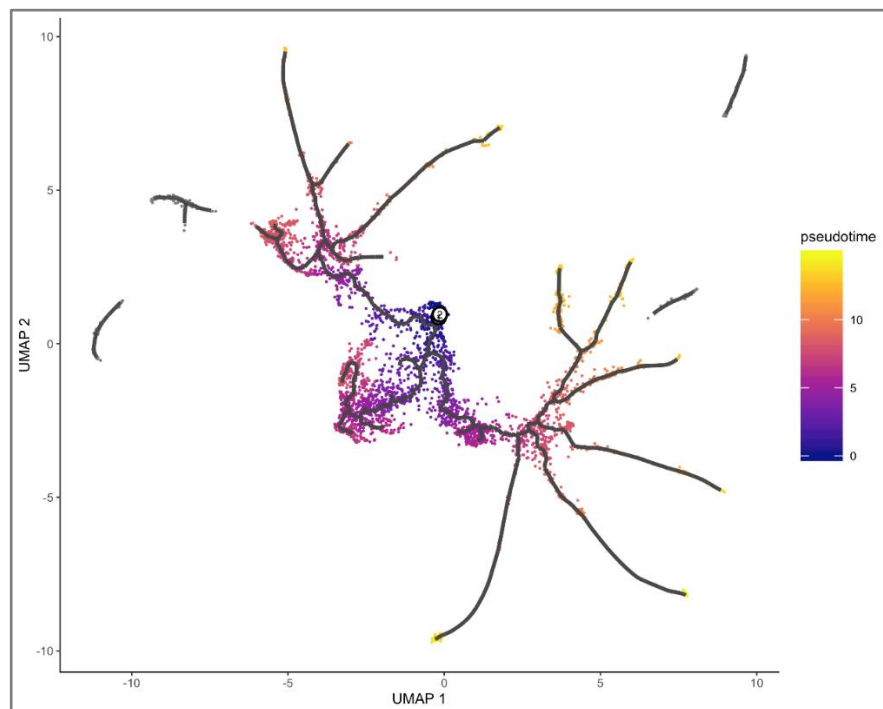


Figure 3.4: Example of a trajectory inference result colored by pseudotime.

3.3.3. Gene Ontology (GO) Enrichment Analysis

Gene ontology (GO) provides a comprehensive database of knowledge about gene and gene product functions. GO project aims to build structured vocabularies (ontologies) for crucial areas of molecular biology and connecting terms with biological sequences and their products [Harris et al., 2004]. Gene ontology is organized in three classes, which are molecular function (activity of a gene product in the molecular resolution), biological process (wide-ranging biological perspective that the molecular function of the gene is organized) and cellular component (location of the activity of the gene product). Each class includes GO terms which are logically connected to each other and these terms have linkages to the gene products. These connections are also called “annotations” [The Gene Ontology Consortium, 2018].

GO is used to interpret the outputs from high-throughput analyses. These outputs mostly include a set of genes. It is possible to analyse the significantly under- or overrepresented biological locations, functions or processes in the gene set of interest. This type of analysis is called GO enrichment analysis [Rhee et al., 2008]. There are a number of tools available for such analysis. Mostly used tools are BiNGO [Maere et al., 2005], GOrilla [Eden et al., 2009], PANTHER [Mi et al., 2013] and g:profiler [Raudvere et al., 2019].

In this thesis, R package of clusterProfiler v3.16.1 was used for GO enrichment analysis [Yu et al., 2012]. The package clusterProfiler uses OrgDbs which are the organism-specific annotation packages that include the detailed information on pathways and genes which are members of these pathways [Carlson, 2019]. Retrieved gene lists in clustering were given as an input to clusterProfiler’s “enrichGO” function for functional profiling. This function was used with some modified parameters as explained below:

- Genes that have average log fold change lower than 0.5 are filtered out and the given list of genes is sorted according to their adjusted p values. P-values were calculated in clustering step by comparing the expression level of the gene in each cluster with respect to expression in all other clusters.
- Complete list of genes in each dataset is considered as the ‘background gene list’ for each dataset.
- OrgDb is selected as “org.Mm.eg.db” since all the datasets are mouse-based.

- Key type of input gene is selected as “symbol”.
- Data source to be used is ‘biological processes’ (BP) database of GO.
- P-value adjustment method is selected as “BH” (Benjamini-Hochberg).
- P-value cutoff is chosen as 0.01.
- Q-value cutoff is chosen as 0.05.

4. RESULTS

4.1. Differentiation of mESC to MN

To examine the differentiation from mESCs to motor neurons, the count matrix was downloaded from GEO database with an accession number of GSE97391. Firstly, cells were clustered, plotted with UMAP algorithm and the genes that define these clusters were found with Seurat package. Then, the same count matrix and already-calculated UMAP coordinates were used to infer trajectories according to calculated pseudotime values of the cells with Monocle package. Summary of all results can be seen in Figure 4.1.

This dataset consists of 3023 cells in total, which includes 969 cells from day 0, 997 cells from day 5 and 1057 cells from day 12. By using Seurat package with default parameters, it was found that 3023 cells were separated into 11 clusters. These clusters indicate the groups of cells with similar gene expression profiles. Most of the clusters consist of the cells collected from the same timepoint, but also the cells from the same timepoint can be seen in different clusters. As seen in Figure 4.1, two of the clusters (cluster 0, 3) consist of day 0 cells (mESC), 3 of them (cluster 1, 4, 5) consist of day 5 cells and 6 of them (cluster 2, 6, 7, 8, 9, 10) consist of day 12 cells.

To get more information about the clusters, the cells in different timepoints were examined according to their clusters and number of cells in the clusters were found for each timepoint as seen in Figure 4.1B. Additionally, marker genes that represent each cluster were detected and GO analysis was applied to significant marker genes to characterize the clusters. With the analysis of the markers, it was found that day 12 cells include different cell types such as astrocytes and myocytes. Markers of each cluster and GO enrichment analysis results of these markers are in the Supplementary File 1 and 2 respectively.

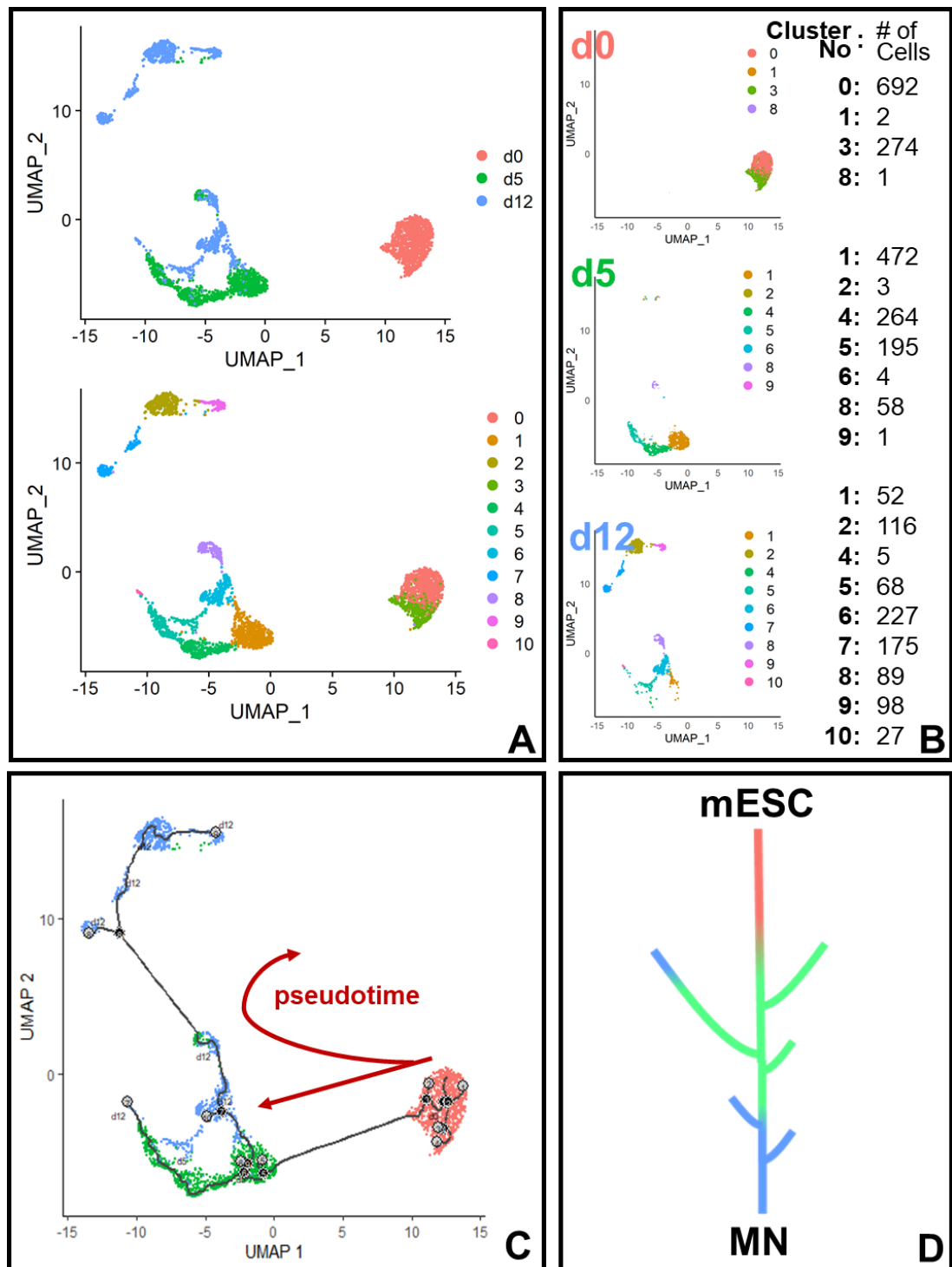


Figure 4.1: Clustering and trajectory inference results of mESC -> MN process. A) UMAP representation of 3023 cells colored according to the timepoints and clusters. B) UMAP representations of the cells at each timepoint colored according to the clusters and the number of cells in the clusters. C) Differentiation trajectory of the cells and the arrow that shows the direction of differentiation. D) Simplified representation of the trajectory.

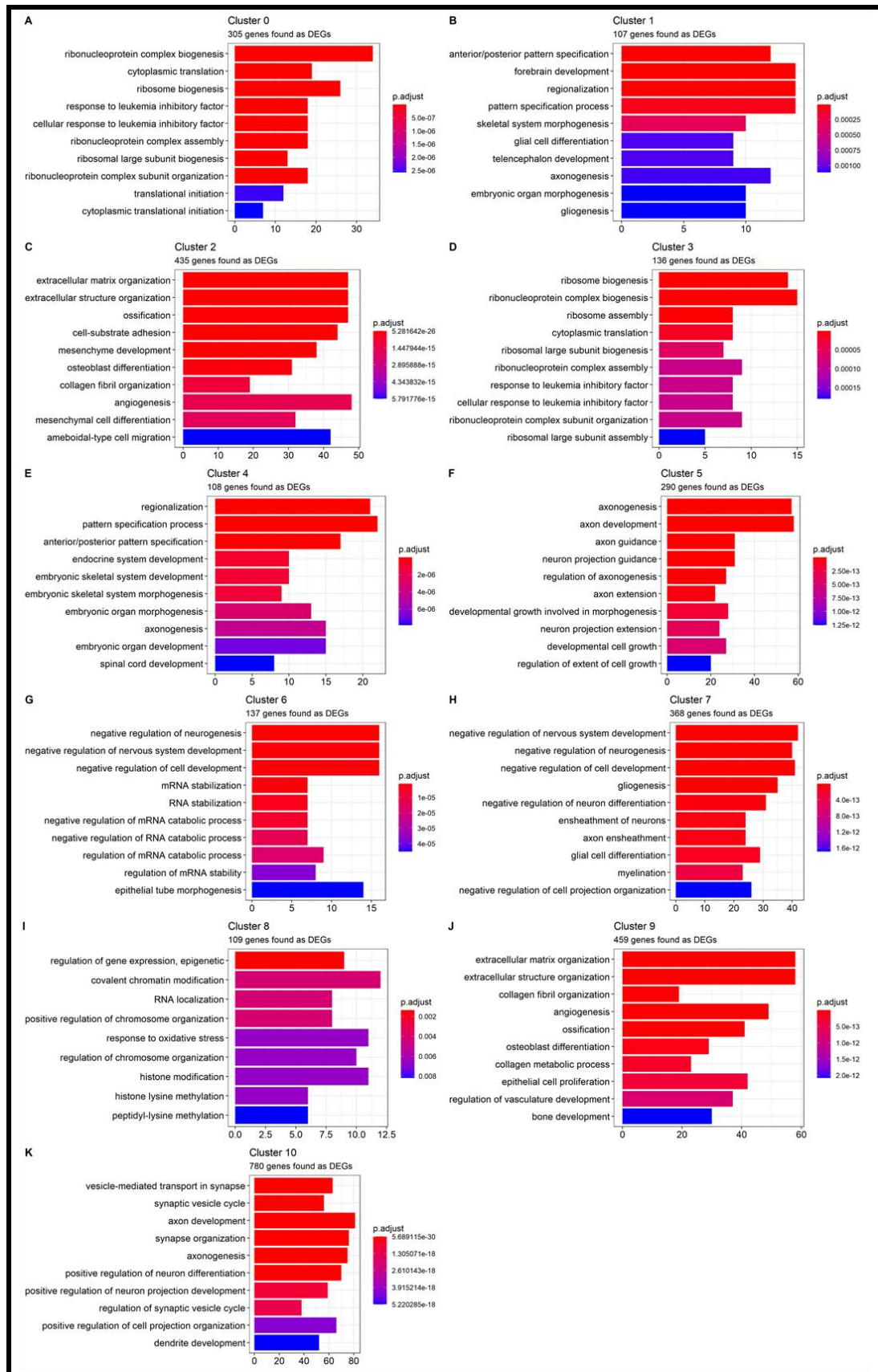


Figure 4.2: Significant GO terms for cluster 0-10 of cells in mESC -> MN process.

Figure 4.2 shows GO enrichment terms for each cluster in mESC -> MN process. In these plots, y axis shows the GO terms and x axis shows the number of the genes that is related with the corresponding term. As expected, some clusters mostly include neuron-related terms. Although the terms can allow us to interpret these clusters, they cannot exactly point which cell type these clusters actually represent. Therefore, GO enrichment analysis finds general terms for each cluster.

After clustering, count matrix and UMAP coordinates calculated in Seurat were used to infer trajectories in Monocle package for the step of fitting a principal graph and ordering the cells according to the pseudotime. During differentiation, cells form two short and one long branch in day 5 and two short branches in day 12 as seen in Figure 4.1C-D.

4.2. Differentiation of MEF to MN

In this dataset, Treutlein et al. studied the differentiation from mouse embryonic fibroblasts to motor neurons by using Brn1+Ascl1+Myt1l (BAM) factors and Ascl1 factor separately, and cells collected in specified timepoints were stored in the same dataset. The count matrix was downloaded from GEO database with an accession number of GSE67310. As done in mESC -> MN process, cells were clustered, plotted with UMAP algorithm and the clusters were found with Seurat package. Later, the count matrix and pre-calculated UMAP coordinates were used to infer trajectories with Monocle package. Summary of all results can be seen in Figure 4.3.

This dataset consists of 405 cells in total which includes 73 cells from day 0, 128 cells from day 2, 55 cells from day 5, 33 cells from day 20, 73 cells from day 22 upon Ascl1 induction; and 43 cells from day 22 upon induction of Ascl1, Brn2 and Myt1l. By using Seurat package with default parameters, it was found that the cells were separated into 9 clusters. Distribution of the cells collected at the same timepoint into multiple clusters and the inclusion of cells from multiple timepoints in a single cluster indicate that there is a high variation of the gene expression profiles of the cells. As seen in Figure 4.3, two of the clusters (cluster 1, 2) consist of day 0 cells (MEF), two of them (cluster 0, 1) consist of day 2 cells, two of them (cluster 3, 8) consist of day 5 cells, mainly one of them (cluster 5) consist of day 20 cells and three of them (cluster 4, 6, 7) consist of day 22 cells.

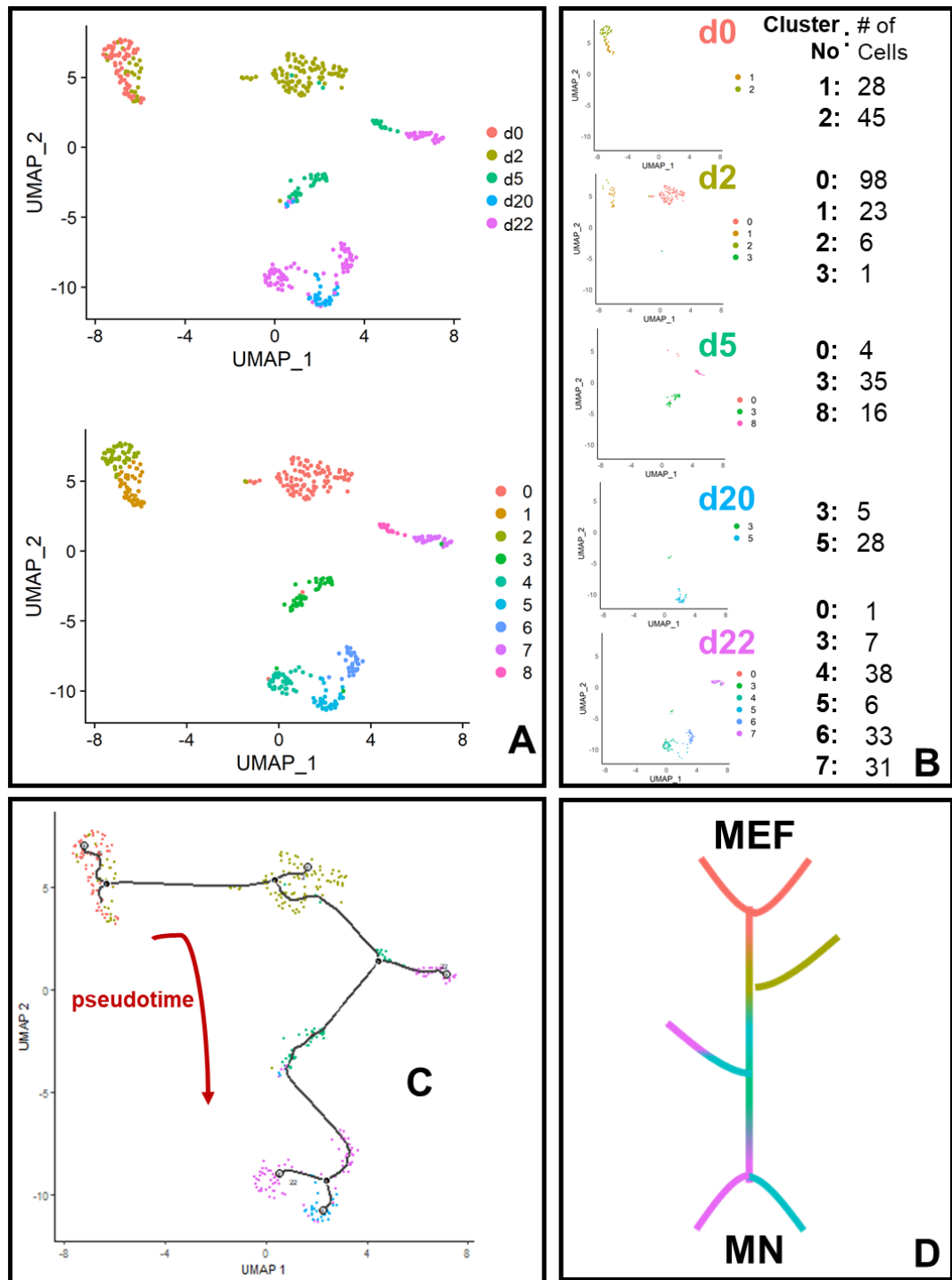


Figure 4.3: Clustering and trajectory inference results of MEF -> MN process. A) UMAP representation of 405 cells colored according to the timepoints and clusters. B) UMAP representations of the cells at each timepoint colored according to the clusters and the number of cells in the clusters. C) Differentiation trajectory of the cells and the arrow that shows the direction of differentiation. D) Simplified representation of trajectory inference.

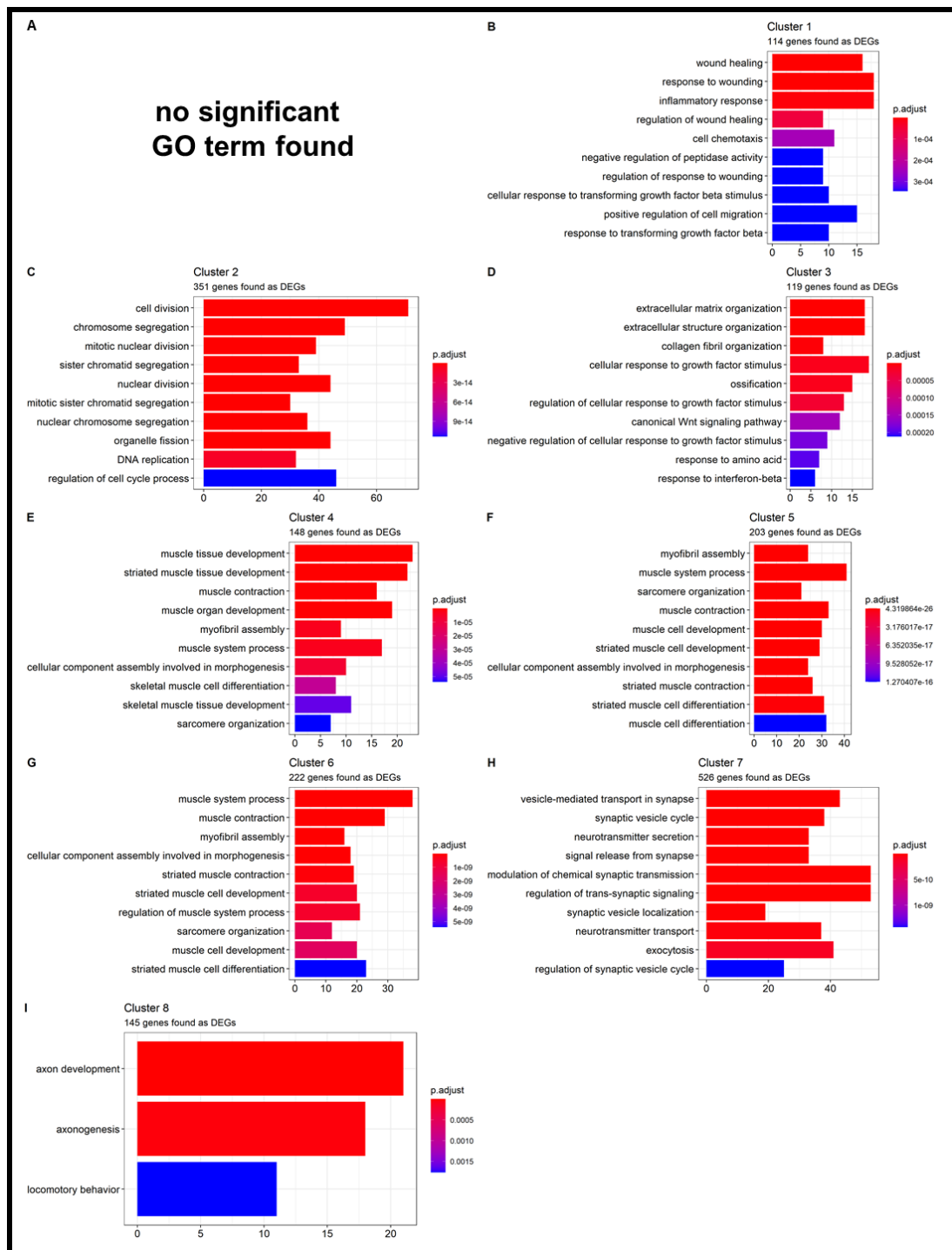


Figure 4.4: Significant GO terms for cluster 0-8 of cells in MEF -> MN process.

To get more information about the clusters, the cells from different timepoints were examined according to their clusters and number of cells in the clusters were found for each timepoint as seen in Figure 4.3B. Additionally, marker genes that represent each cluster were detected and GO analysis was applied to significant marker

genes to characterize the clusters. Markers of each cluster and GO results of these markers are in the Supplementary File 3 and 4 respectively.

As previously seen in mESC -> MN process, GO enrichment terms found for clusters in MEF -> MN process show general terms in Figure 4.4. In these plots, y axis shows the GO terms and x axis shows the number of the genes that is related with the corresponding term. During the differentiation, some cells were transformed into muscles instead of neurons. As seen here, these terms are not enough by themselves to interpret cellular composition of these clusters, therefore more analyses were conducted in the scope of this thesis, and these unexpected cells were explained in “Discussion”.

As done in the mESC -> MN process, Monocle package was used with the count matrix and UMAP coordinates from Seurat to infer trajectories. During the direct differentiation, cells form a small branch in day 2 and the differentiation takes place in two distinct differentiation routes due to the different treatments.

4.3. Reprogramming of MEF to iPSC

Tran et. al examined the differentiation from MEF to iPSCs in two different protocols as explained in Experimental Design section. Authors stated in the paper that reprogramming of the cells to iPSC is more accurate in A2S-based protocol than FBS-based protocol. The count matrix was downloaded from GEO database with an accession number of GSE108222. Cells from A2S-based protocol and cells from FBS-based protocol were separated from the dataset and the following steps were applied to each group separately. As done in the previous datasets, cells were clustered, plotted with UMAP algorithm and the genes that define these clusters were found with Seurat package. Then, the same count matrix and pre-calculated UMAP coordinates were used to infer trajectories with Monocle package. Summary of all results can be seen in Figure 4.5 and Figure 4.6.

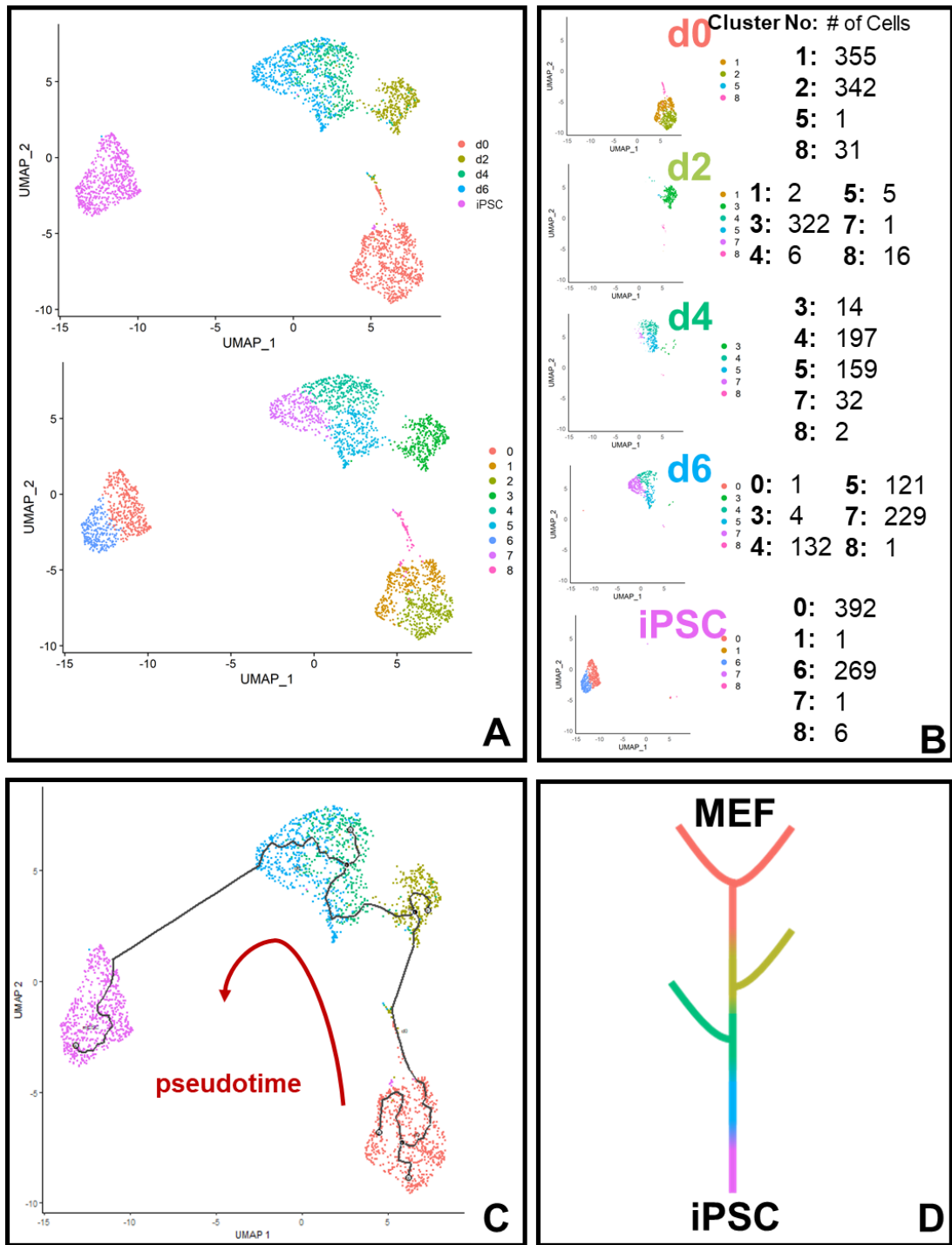


Figure 4.5: Clustering and trajectory inference results of cells from A2S-based protocol in MEF -> iPSC process. A) UMAP representation of 2642 cells colored according to the timepoints and clusters. B) UMAP representations of the cells at each timepoint colored according to the clusters and the number of cells in the clusters. C) Reprogramming trajectory of the cells and the arrow that shows the direction of reprogramming. D) Simplified representation of trajectory inference.

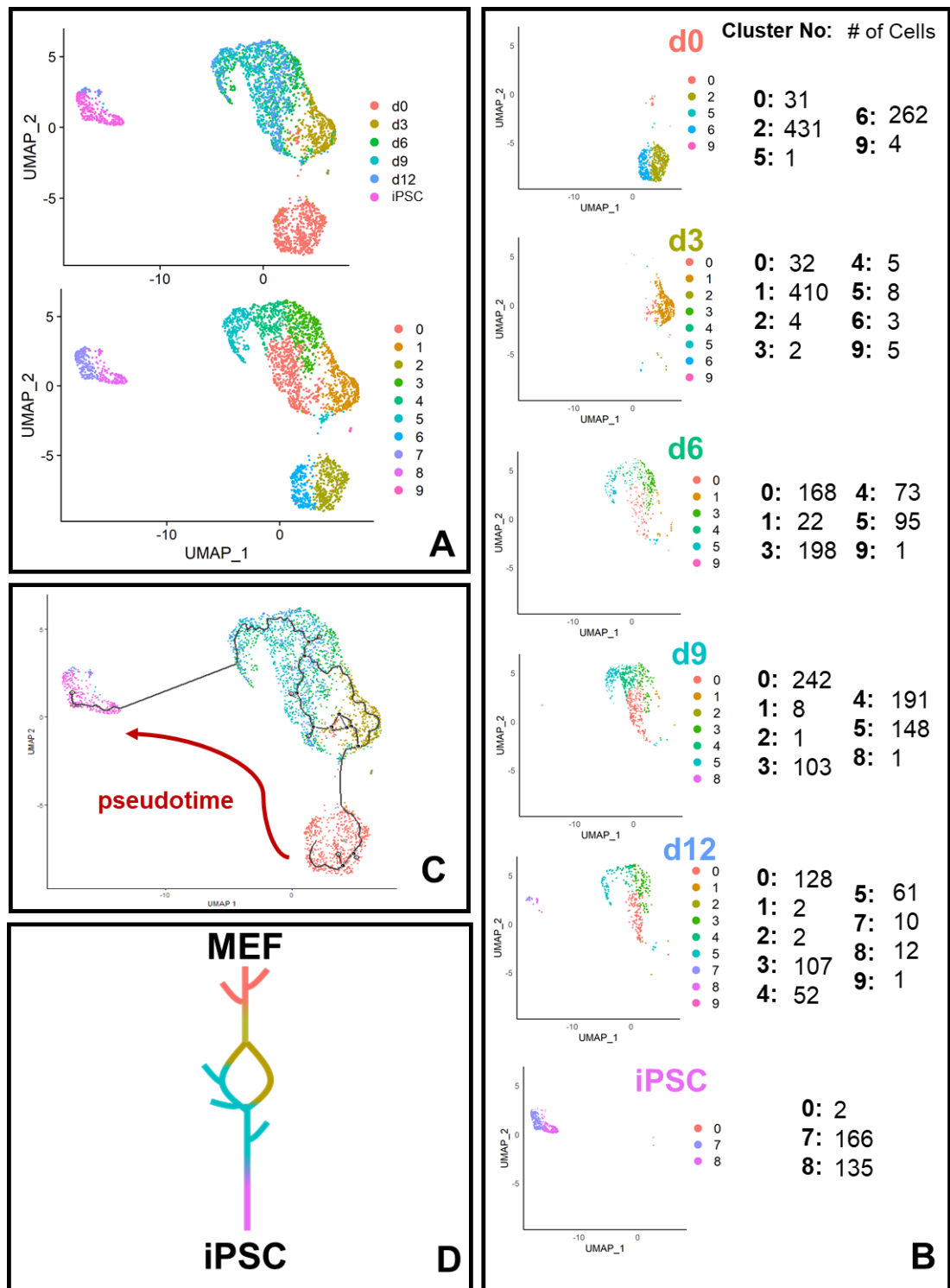


Figure 4.6: Clustering and trajectory inference results of cells from FBS-based protocol in MEF -> iPSC process. A) UMAP representation of 3127 cells colored according to the timepoints and clusters. B) UMAP representations of the cells at each timepoint colored according to the clusters and the number of cells in the clusters. C) Reprogramming trajectory of the cells and the arrow that shows the direction of reprogramming. D) Simplified representation of trajectory inference.

This dataset consists of 2642 cells from A2S-based protocol, which includes 729 cells from day 0, 352 cells from day 2, 404 cells from day 4, 488 cells from day 6 and 669 cells from the last timepoint; and 3127 cells from FBS-based protocol, which includes 729 cells from day 0, 469 cells from day 3, 557 cells from day 6, 694 cells from day 9, 375 cells from day 12 and 303 cells from the last timepoint. By using Seurat package with default parameters, it was found that the cells from A2S-based protocol were separated into 9 clusters and cells from FBS-based protocol were separated into 10 clusters. Similar to the other datasets, there is a high variation of the gene expression profiles in both of the cell sets as can be seen in the distribution of the sampled cells from the same timepoint into the multiple clusters and the inclusion of cells from multiple timepoints in a single cluster. In the cells at middle timepoints (d6/d9/d12 in FBS-based protocol, d4/d6 in A2S-based protocol), there is no distinct separation, indicating that there is no significant difference between them. As seen in Figure 4.5 and Figure 4.6, cells in each timepoint are represented in two or three distinct clusters.

To get more information about the clusters, the cells in different timepoints were examined according to their clusters and number of cells in the clusters were found for each timepoint as seen in Figure 4.5B and Figure 4.6B. Additionally, marker genes that represent each cluster were detected and GO analysis was applied to significant marker genes to characterize the clusters. Markers of each cluster and GO results of these markers are in the Supplementary File 5 and 6 for MEF -> iPSC process with A2S-based protocol; and in the Supplementary File 7 and 8 for MEF -> iPSC process with FBS-based protocol respectively.

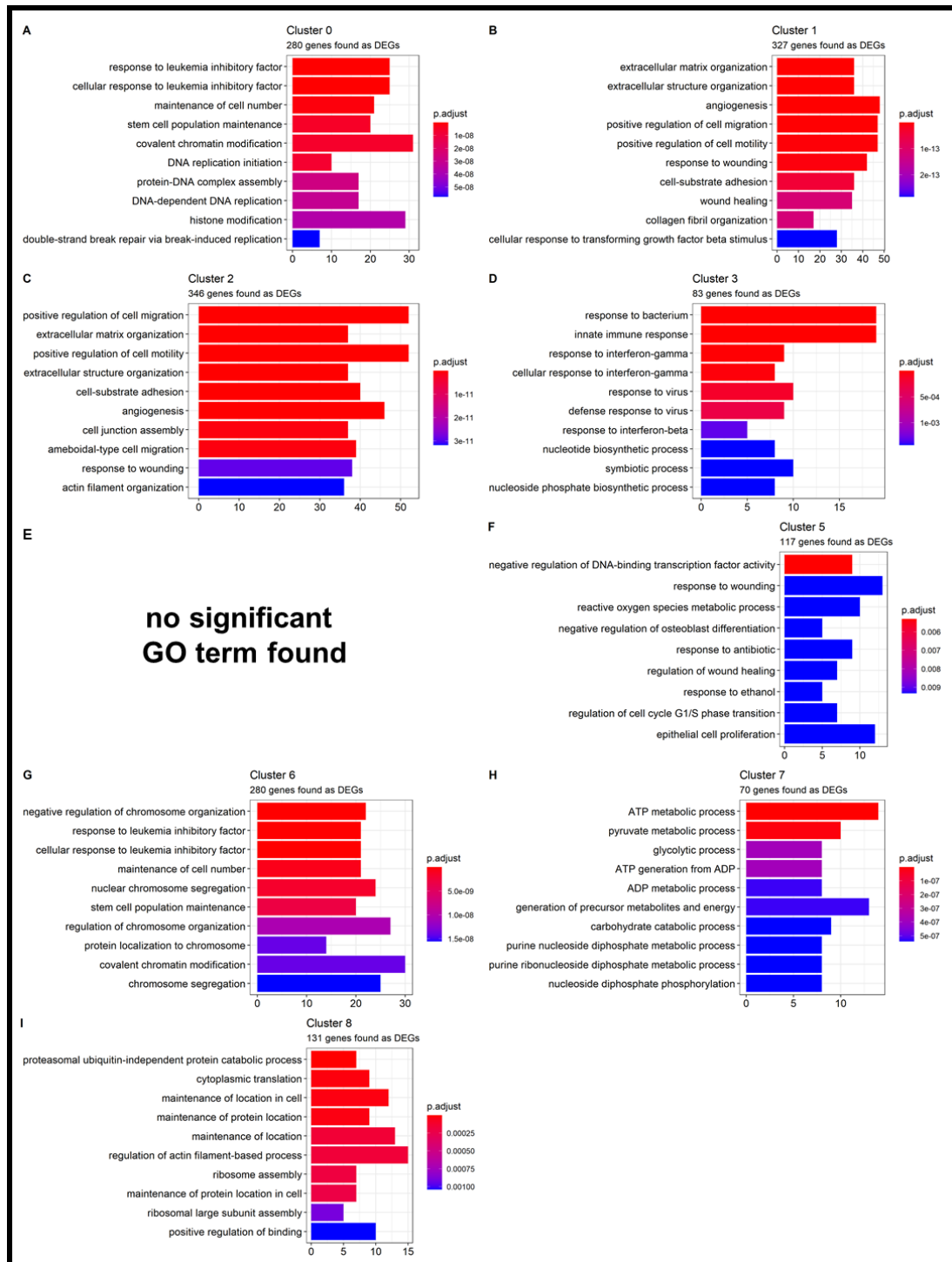


Figure 4.7: Significant GO terms for cluster 0-8 of cells from A2S-based protocol in MEF → iPSC process.

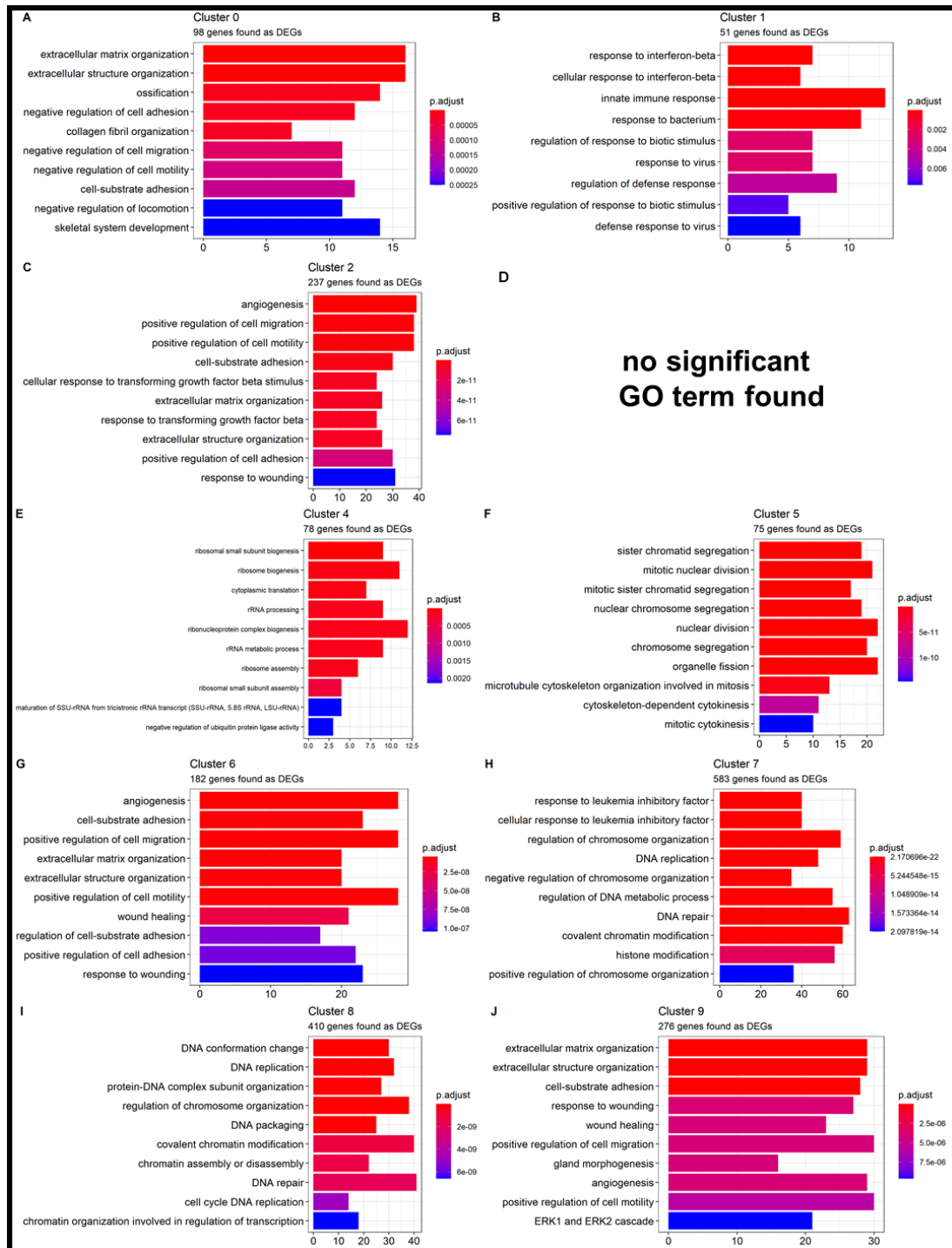


Figure 4.8: Significant GO terms for cluster 0-8 of cells from FBS-based protocol in MEF -> iPSC process.

As seen in previous datasets, GO enrichment terms found for clusters in MEF -> iPSC process show general terms in Figure 4.7 and Figure 4.8. In these plots, y axis shows the GO terms and x axis shows the number of the genes that is related with the corresponding term. Two different protocols show similar but not the same GO terms in each set. More detailed results were found with trajectory inference. With Monocle,

reprogramming trajectory was inferred for both cell sets by using count matrix and UMAP coordinates calculated in Seurat. Cells form more complex branches with FBS-based protocol, whereas branching is more distinctive with A2S-based protocol.

5. DISCUSSION

mRNA measurements (scRNA) collected from single-cell samples are an invaluable source of information in terms of indicating the diversity in cell populations that are transforming and in terms of allowing the mechanisms to be studied in single-cell resolution. Having gene expression levels from each individual cell brings a higher resolution and it allows novel ways of analyses when compared to having gene expression levels from bulk measurements. In the scope of this thesis; single-cell analyses showed how the cells were clustered and how branching of the cellular profiles to different fates took place while the cells were differentiating/reprogramming. With the further analyses, the correlation coefficients of the genes according to the pseudotimes and the genes that have the driving role in the process were found and explained in this section.

5.1. Trajectory Inference

In the differentiation process from mESC to MN, cells in day 5 start to lose their ESC profiles and divide into two main branches. One of these branches mainly consists of day 5 cells, indicating that these cells might lose their pluripotency to differentiate into neurons; whereas the other branch consists of a small part of day 5 cells and all day 12 cells, indicating that differentiation into neuron takes place in this branch. Day 12 cells are highly heterogeneous since they include different cell types other than neurons. Therefore, they divide into two small branches. Longer small branch goes to myocytes, oligodendrocytes and stromas which are completely different phenotypes compared to neurons; shorter small branch goes to astrocytes, which are closer cell type to the neurons.

Trajectory inference result of MEF -> MN process shows that MEFs from two different profiles start to differentiate and create a branch in day 2, which indicates that there is a different gene expression profile in the new branch compared to the cells in the main trajectory. Also, there is a sharp distinction in cells from day 20/22, which is caused by use of two different differentiation protocols. Based on the classification of the marker genes, Treutlein et. al reported that differentiation using BAM factors is more successful than the differentiation using *Ascl1* only. In this dataset, it is seen that

there are other cell types such as myocytes among day 22 cells differentiated using Ascl1. This situation was interpreted by authors as follows: Brn2 and Myt1l genes suppress the differentiation into other cell types, and the cells are able to differentiate into different cell types in the absence of these genes. Effects of these genes can be seen as a sharp bifurcation in the trajectory results, meaning these genes have a substantial role in separating differentiation into two different routes.

In the trajectory inference result of the cells reprogrammed with FBS-based protocol, it is seen that cells go through a complex reprogramming processes from less heterogeneous MEF population to a highly heterogeneous iPSC population, hence the trajectory inference algorithm cannot determine a main trajectory graph in this dataset, unlike in the other datasets. In the trajectory inference result of the cells reprogrammed with A2S-based protocol, reprogramming creates less heterogeneous clusters compared to the other protocol, which led A2S inhibitors made the reprogramming process more accurate. And it is also seen that day 2 cells and day 4/6 cells were separated to create new small branches. This might show that the cells on the small branches were not dedicated to be reprogrammed into iPSCs as much as the other cells in the main trajectory path.

5.2. Stochasticity and Determinism of the Processes

The fact that cell populations that are being differentiated or reprogrammed can be quite heterogeneous, the ability to order the cells according to their pseudotime values while taking into account the gene expression profiles, and the possibility of these highly correlated genes being the driver of the other differentiation processes are shown in this section. Having highest correlations in the triggered or suppressed genes supports our assumption.

Inferred trajectories of these processes show us which routes the cells can follow while differentiating or reprogramming. If these trajectories are examined closely, it can be seen that some cells can give rise to some major and minor branches in each trajectory. By examining the way of their branching, these processes can be evaluated according to their stochasticity and determinism. It can be assumed that the branches that divert from the main trajectory are formed by the cells that lost their dedication on completing the transformation to the target cell type. It can be assumed

that these cells behaved stochastically despite the external signals that direct them to target cell profile. The other cells which stayed on the main trajectory can be interpreted to behave as expected. It can be assumed that changes in these cells are deterministic and has been following the route dictated by the external signals in the growth medium.

5.2.1. Profiles of the Marker Genes

To learn more about determinism and stochasticity of the processes, marker genes from the cell types were examined. Marker genes were selected from the literature and from the results of the data analysis and maximum 6 genes were selected per cell type. Since some genes could not be detected in these single-cell datasets, some panels do not include all the marker cells that were chosen.

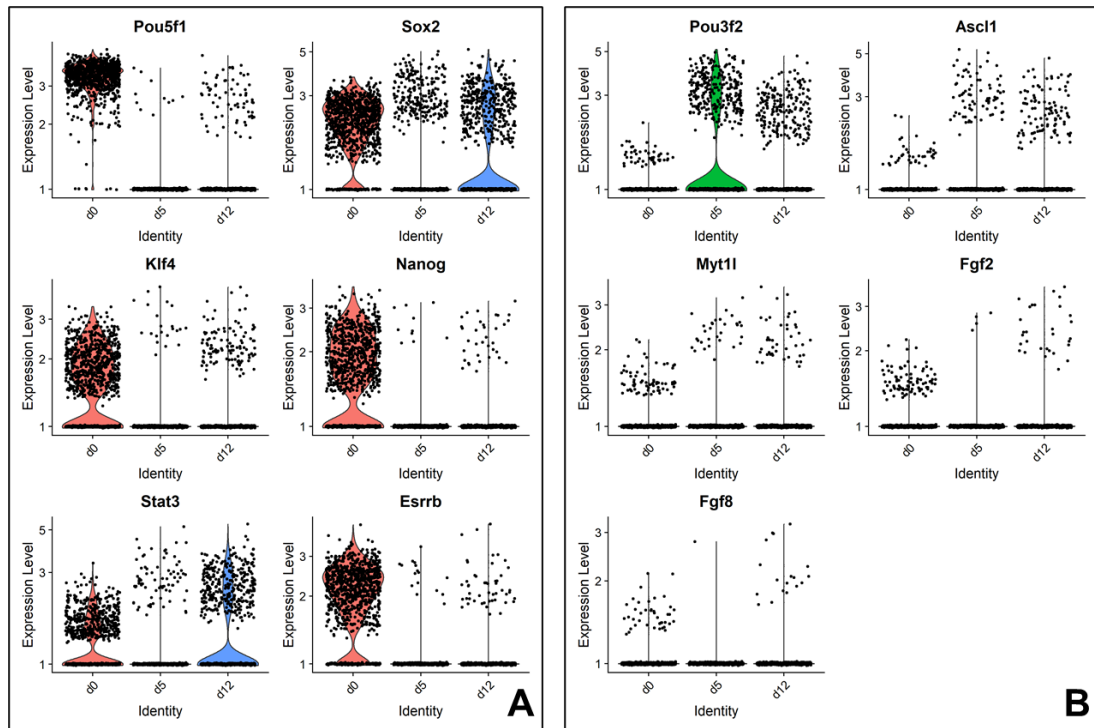


Figure 5.1: Violin plots of (A) pluripotency and (B) neuron marker genes from mESC -> MN process (Briggs et al, 3029 cells from three timepoints).

In the violin plots of Figure 5.1, while pluripotency marker genes is expressed highly in mESCs, they almost completely vanished in day 5. This indicates that pluripotency phenotype is completely gone in the first five days of the experiment. In

Figure 4.1B, neuron markers which were expressed in very few amount of cells increased in day 5, but they partially decreased in day 12. These profiles support that the cells were diverted away from being neurons in day 12. Re-expression of Stat3 and Sox2 genes indicates the emergence of different cell types that these genes are functional in.

In the violin plots of Figure 5.2A-B, markers from MEF-> MN were examined, pluripotency markers were given in Figure 5.2C for comparison. Figure 5.2C shows that a group of cells consistently express Stat3 and Klf4 genes higher, and no expression of other pluripotency markers indicates that the cells in MEF -> MN process were never got in pluripotency stage and cells differentiated directly to the neurons from MEFs. Cells that express Pou3f2 (Brn2) and Myt1l at day 22 are the cells from BAM experiment and the number of these cells are presented in the Venn schemes below.

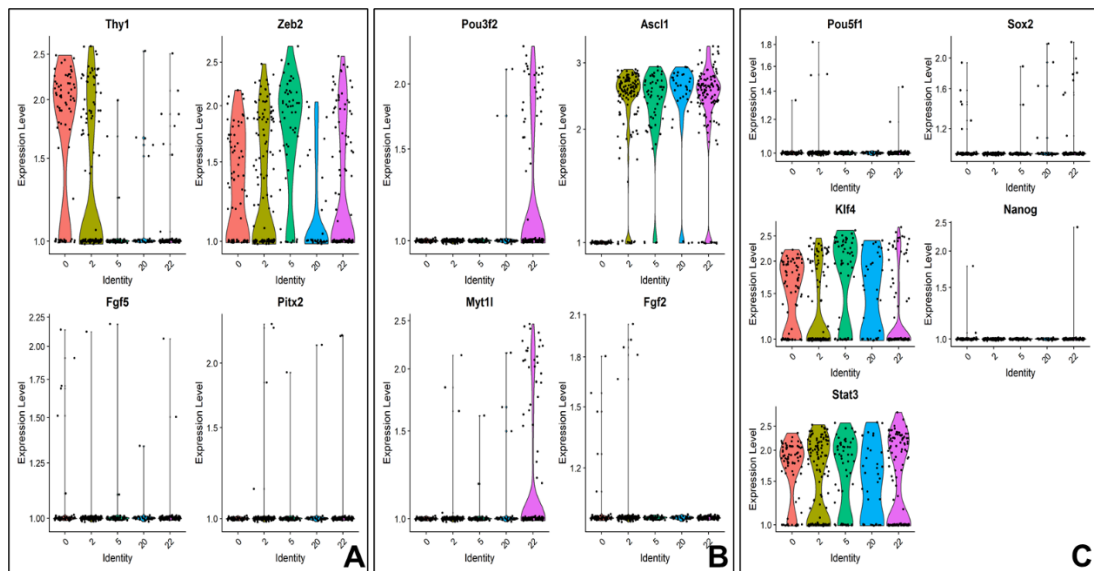


Figure 5.2: Violin plots of (A) MEF, (B) neuron and (C) pluripotency marker genes from MEF -> MN process (Treutlein et al, 405 cells from five timepoints, Ascl1 only and BAM experiments included).

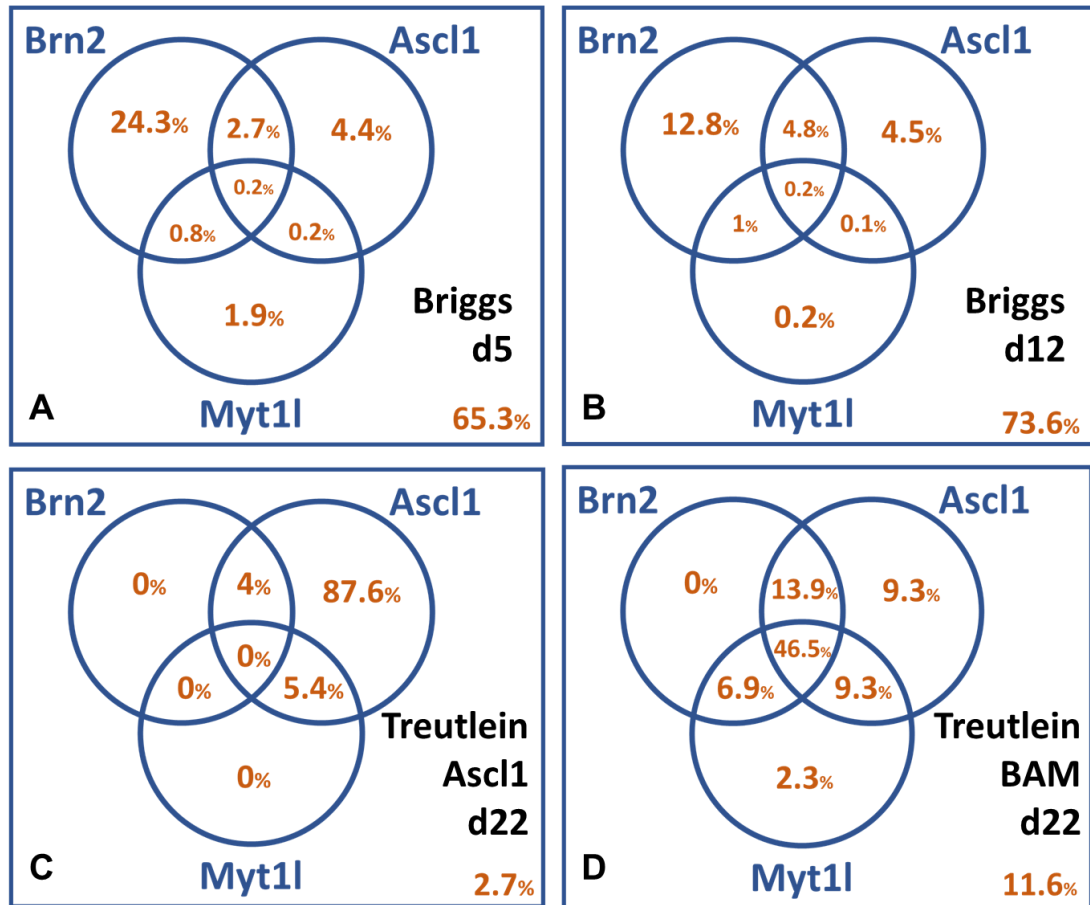


Figure 5.3: The percentage of cells that expressed marker genes at high levels in (A, B) mESC → MN and (C, D) MEF → MN processes.

The percentage of cells that express neuron markers in the experiment of mESC → MN is given in Figure 5.3A-B. the percentages outside of the circles represent the cells that cannot express any of these genes. The percentages of cells that express all three of the markers mutually is very low. In the experiment of MEF → MN where Ascl1 was triggered only, most of the cells were able to express Ascl1, and none of the cells expressed the other two markers mutually as seen in Figure 5.3C. However, three of the markers could be expressed in 46.5% of the cells from the experiment of MEF → MN where BAM factors were triggered. The expression levels of these genes in motor neuron generation processes were given in Figure 5.4. Gene expression levels increased while cells are differentiating, this increase is particularly evident in the dataset of Briggs et al, where the starting cells are mESC, and having a positive correlation of gene expression levels with the pseudotime values also suits these profiles. It is seen that Gsk3b, located under the WNT pathway, is an important gene in the formation of neuron profile as well as playing a role in the exit from

pluripotency, and this effect can be observed at the level of gene expression. It is likely this event to take place through Myc, a gene that is suppressed by GSK3B. The expression levels of the Myc gene were negatively correlated (-0.496 and -0.211) with the pseudotime values in Ascl1 and BAM experiments. It was reported that the Myc gene affects neuron development [Zinin et al., 2014], these analyses indicate that Myc is regulated by Gsk3b in this process. By using this data within the scope of this thesis, it can be anticipated that the Myc gene has a driving role in the formation of neuron cells, and triggering this gene directly or through suppression of Gsk3b will accelerate the process and make it more accurate.

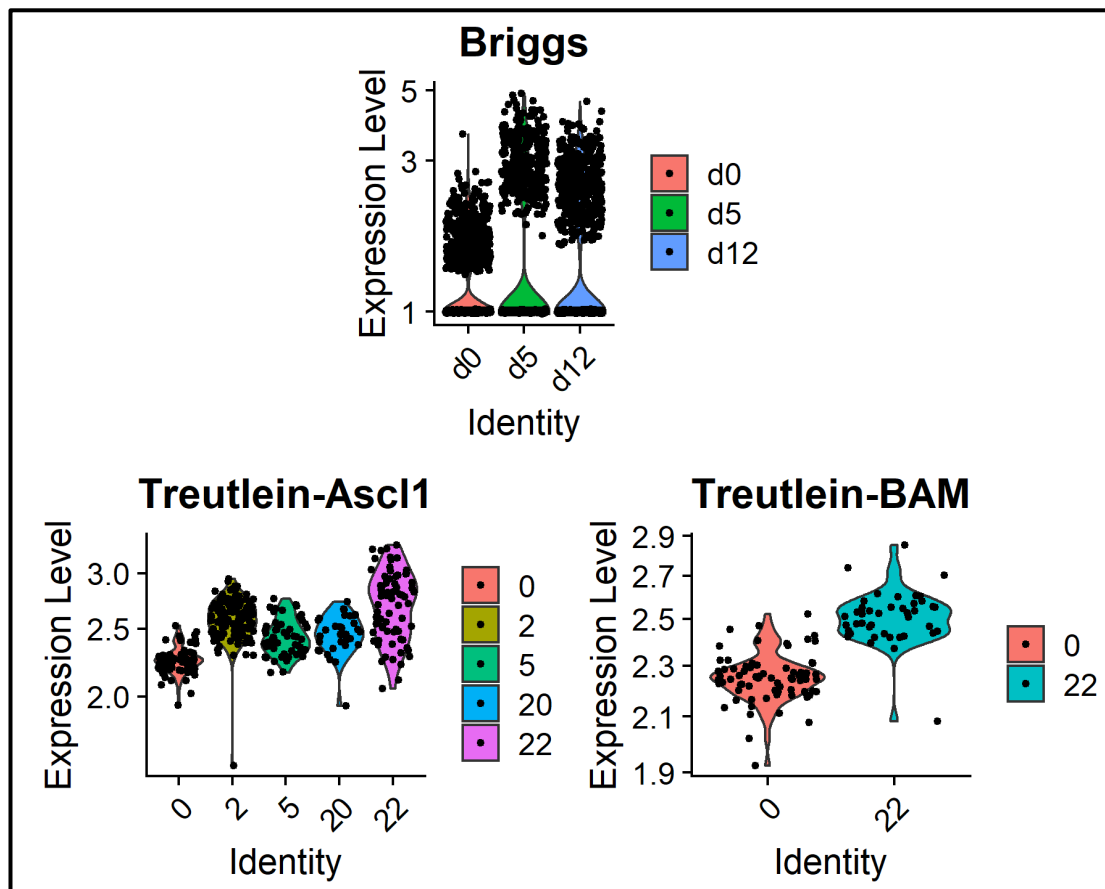


Figure 5.4: Expression levels of Gsk3b gene in mESC -> MN and MEF -> MN processes.

The change of the marker genes of MEFs and mESCs with respect to time in the MEF -> iPSC process was given as violin plots in Figure 5.5 and Figure 5.6. In the study from Tran et al, all of their results include iPSCs; however, their experiments and dataset do not include iPSCs but mESCs. Since iPSCs and mESCs are close to each other according to their pluripotency, mESCs can be assumed as the final

timepoint for this dataset, and further analyses except the calculation of correlation based on pseudotime values were done based on this assumption. In the reprogramming of MEF to iPSC with A2S-based protocol, expression of MEF markers were decreasing by time, which shows that MEF phenotype was disappearing. Also, there is a noticeable increase in the expression of pluripotency markers by time. But a similar decrease in the expression MEF markers cannot be seen in the reprogramming with FBS-based protocol, and the increase in the expression of pluripotency markers is also not noticeable with this protocol.

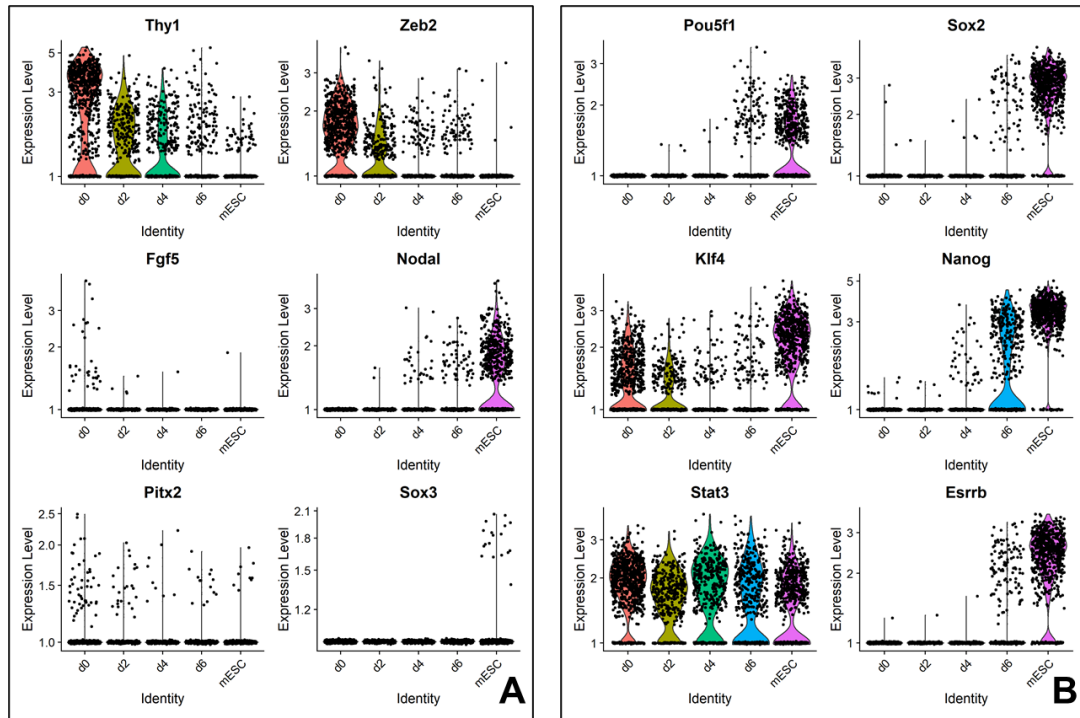


Figure 5.5: Violin plots according to the expression levels of (A) MEF, (B) pluripotency markers in MEF -> iPSC process (Tran et al, 2642 cells from five timepoints, the experiment that OSKM was triggered and inhibitor was added – A2S-based protocol).

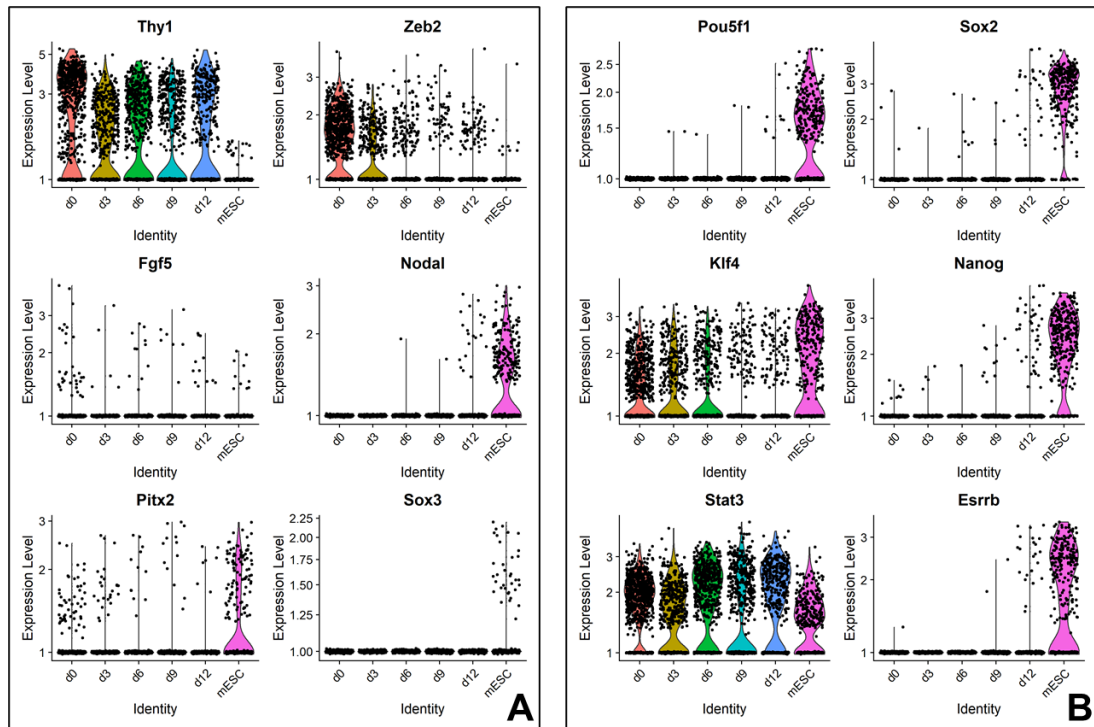


Figure 5.6: Violin plots according to the expression levels of (A) MEF, (B) pluripotency markers in MEF → iPSC process (Tran et al, 3127 cells from five timepoints, the experiment that OSKM was triggered in FBS medium – FBS-based protocol).

The percentage of cells that expressed marker genes at a high level in the MEF → iPSC process conducted by Tran et al were given in Figure 5.7. In these experiments, the Nanog gene was not triggered, but Nanog expression was high in most cells. And although the other two genes were triggered, they were not expressed in many cells. In the A2S experiment where the inhibitors were also used, three of the markers could be expressed in most cells.

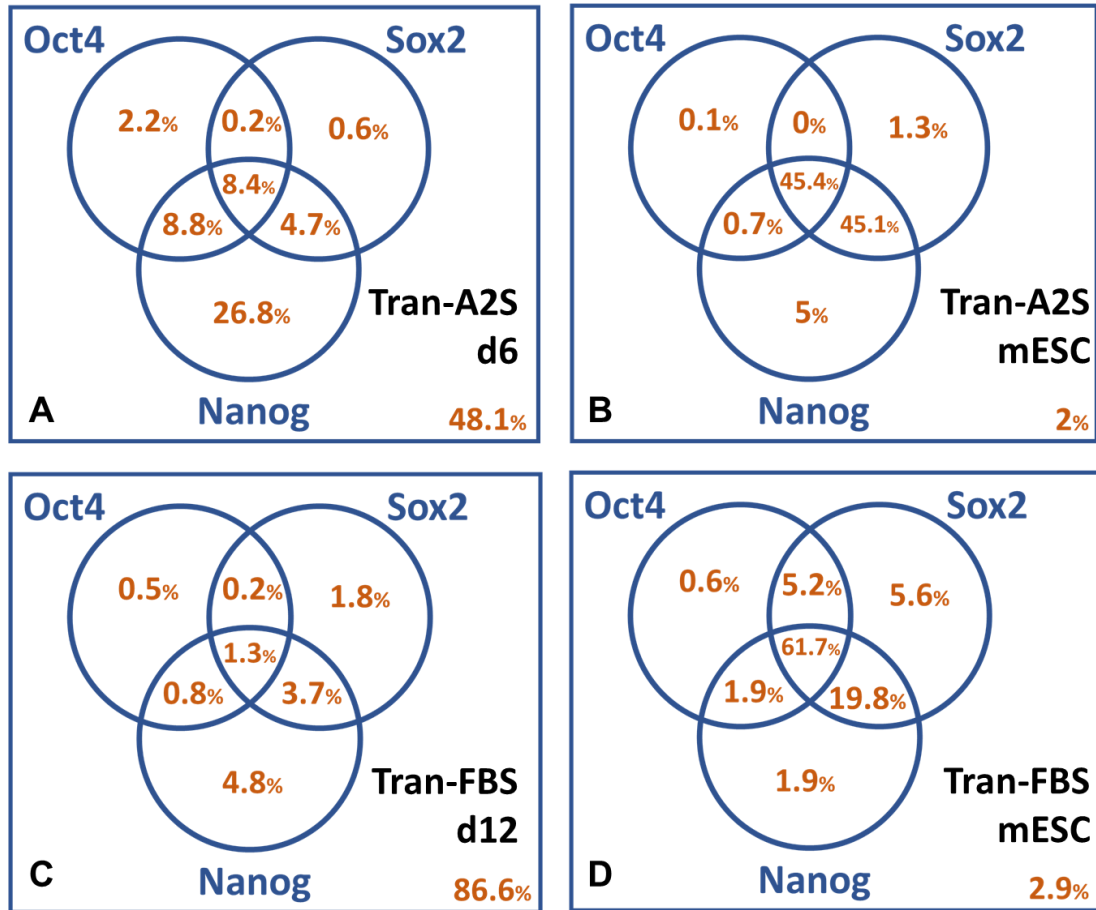


Figure 5.7: The percentage of cells that expressed marker genes at high levels from (A, B) A2S-based protocol and (C, D) FBS-based protocol in MEF → iPSC process.

5.3. Analysis of the Driver Genes

The datasets analysed in this study were from diverse experimental designs, hence integration of the raw datasets was not possible. However, it is possible to compare how the trajectories are regulated in these three processes by using the changes in cell states with respect to the pseudotime.

The "driver genes" concept has been developed to identify the genes that are assumed to be related to differentiation processes. Genes that have this relation can be used to separate the cell profiles and can be directly related to pseudotime values that were used to order cells on the trajectories. To find the driver genes, Spearman correlation coefficients were calculated with the gene expression levels and the pseudotime values. Spearman correlation was chosen instead of Pearson to prevent biases in correlation coefficients of very highly expressed genes, and to avoid the assumption of linear relationship between gene expression levels and pseudotime.

values. If the expression level of a gene increases or decreases in cells compared to other genes as pseudotime values of cells increase, regardless of whether this increase or decrease is linear, that gene may be a 'driver gene' or it may be regulated by one of the driver genes. Functions of the genes with high correlation coefficients were examined together to find driver genes or the genes that regulate driver genes. Due to the prominence of pluripotency pathways in the investigated processes, only the correlations of genes in this pathway with pseudotime values are given in Table 5.1. The correlation coefficients larger than 0.2 and smaller -0.2 were shown in pink and green colours on Table 5.1, the threshold of 0.2 corresponds to smallest absolute correlation coefficient that is significantly different than 0 (Fisher's z-test, adjusted p-value = 0.05, corrected by using Bonferroni method), and important transcription factors were written in bold.

mESC -> MN column in Table 5.1 refers to the cells that exit from pluripotency and show that important transcription factors (TFs) for pluripotency have negative values (according to pseudotime values). Some of these TFs are Nanog, Sox2, Klf4, Pou5f1 (Oct4) which are OSKM genes but excluding M, Esrrb and Tbx3. While their pseudotime values increased, expression levels of these genes decreased. Therefore, this decrease of the activities of these TFs can be a “driver” for this cell differentiation process. Same genes in MEF -> iPSC column have a positive correlation and they have a driver role in the reprogramming process. Klf4 gene is an exception here and it is understood that this gene has no role in MEF -> iPSC process. Myc gene is the last one in OSKM factors, however it is seen that this gene has no active role in reprogramming and differentiation processes according to these datasets. These results indicate that K and M of OSKM factors have no critical role in these processes [Buganim et al., 2013].

Stat3 gene, which regulates the driver genes of Oct4, Sox2 and Nanog, has a driver role only in MEF -> iPSC process and has no effect on gene expression levels in mESC -> MN process.

Table 5.1: Correlation coefficients of the gene expression in pluripotency pathways with pseudotime values.

PP Gene	iPSC -> MN	MEF -> MN	MEF -> iPSC (A2S)	MEF -> iPSC (FBS)
Acvr1	0.070	-0.051	-0.500	-0.379
Acvr1b	-0.062	-0.062	0.040	0.019
Acvr1c	-0.242	0.068	0.000	0.000
Acvr2a	0.041	0.098	-0.233	-0.191
Acvr2b	-0.148	0.044	0.244	0.274
Akt1	0.182	-0.133	-0.646	-0.492
Akt2	-0.191	0.402	-0.118	-0.222
Akt3	0.017	0.174	-0.280	-0.209
Apc	-0.022	0.098	-0.108	-0.150
Apc2	0.102	0.101	-0.035	0.048
Axin1	-0.287	-0.185	-0.077	-0.139
Axin2	0.074	0.184	0.266	0.058
Bmp4	-0.164	-0.234	-0.170	-0.186
Bmpr1a	-0.055	0.132	-0.338	-0.214
Bmpr1b	-0.120	0.051	-0.348	-0.313
Bmpr2	0.085	0.105	-0.580	-0.428
Ctnnb1	-0.117	0.056	-0.164	-0.256
Dusp9	-0.426	-0.151	0.192	0.120
Dvl1	-0.054	-0.014	-0.288	-0.261
Dvl2	-0.250	-0.068	0.174	0.079
Dvl3	-0.179	0.201	0.015	-0.006
Esrrb	-0.695	0.000	0.733	0.434
Fzd1	0.270	0.050	-0.444	-0.272
Fzd10	-0.111	0.000	0.383	0.135
Fzd2	0.327	-0.095	-0.566	-0.434
Fzd3	-0.150	0.035	-0.279	-0.081
Fzd4	-0.088	-0.007	-0.059	-0.077
Fzd5	-0.465	-0.074	0.359	0.183
Fzd6	-0.123	-0.261	-0.232	-0.165
Fzd7	-0.046	-0.094	-0.587	-0.255
Fzd8	0.035	0.026	-0.307	-0.268
Fzd9	-0.036	0.137	0.124	0.074
Grb2	-0.284	0.092	-0.032	-0.195
Gsk3b	-0.021	0.256	-0.128	-0.129
Id1	-0.377	-0.066	-0.574	-0.357
Id2	-0.004	0.116	-0.378	-0.176
Id3	0.103	-0.027	-0.570	-0.524
Id4	0.225	0.074	-0.397	-0.234
Il6st	0.136	0.062	-0.372	-0.388
Inhba	-0.131	-0.574	-0.534	-0.472
Inhbb	-0.106	-0.105	-0.262	-0.231
Inhbc	-0.097	0.000	0.000	0.000
Inhbe	-0.078	0.000	0.000	0.000
Jak1	-0.056	0.004	-0.665	-0.568

Table 5.1 Continued.

Jak2	0.149	-0.129	-0.385	-0.255
Jak3	-0.185	-0.005	0.234	0.057
Klf4	-0.521	-0.079	0.186	-0.126
Lif	-0.091	-0.080	-0.123	-0.096
Lifr	-0.229	-0.138	0.045	0.034
Map2k1	-0.195	0.219	-0.211	-0.324
Map2k2	-0.414	-0.077	-0.202	-0.261
Mapk1	-0.150	-0.123	-0.307	-0.279
Mapk3	-0.028	0.019	-0.413	-0.203
Myc	-0.068	-0.216	-0.490	-0.253
Nanog	-0.636	-0.051	0.806	0.478
Nodal	-0.371	0.000	0.550	0.382
Pik3ca	-0.091	0.176	-0.401	-0.295
Pik3cb	-0.255	0.035	-0.100	-0.006
Pik3r1	0.142	0.090	-0.224	-0.129
Pik3r2	-0.077	-0.002	-0.078	-0.051
Pik3r3	0.024	0.161	0.016	-0.037
Pou5f1	-0.768	-0.040	0.506	0.418
Smad1	0.088	0.089	-0.279	-0.215
Smad2	-0.135	-0.066	-0.198	-0.218
Smad3	0.066	0.028	-0.363	-0.296
Smad4	-0.013	-0.066	-0.286	-0.245
Smad5	0.053	0.021	-0.352	-0.237
Smad9	-0.150	-0.039	-0.229	-0.180
Sox2	-0.576	0.056	0.741	0.505
Stat3	-0.051	-0.140	-0.437	-0.322
Tbx3	-0.448	0.030	0.536	-0.096
Tcf3	-0.151	-0.072	-0.114	-0.097
Tcf7	-0.104	-0.008	0.159	0.205
Wnt1	-0.044	0.000	0.000	0.000
Wnt10a	-0.057	-0.177	-0.030	-0.050
Wnt10b	-0.073	0.000	-0.034	-0.011
Wnt11	-0.027	0.047	-0.253	-0.256
Wnt16	-0.165	-0.078	-0.044	-0.009
Wnt2	-0.035	-0.080	-0.225	-0.139
Wnt2b	-0.079	0.015	-0.061	-0.046
Wnt3	-0.086	0.039	0.052	0.071
Wnt3a	-0.168	0.000	0.058	0.085
Wnt4	-0.157	-0.020	-0.281	-0.191
Wnt5a	0.204	-0.248	-0.690	-0.602
Wnt5b	0.152	0.105	0.004	-0.034
Wnt6	-0.141	0.000	0.065	0.039
Wnt7a	-0.027	0.023	0.019	0.018
Wnt7b	-0.301	0.000	0.077	0.075
Wnt8a	-0.066	0.000	0.065	0.000
Wnt9a	0.048	0.012	-0.275	-0.180

MEF → MN process follows a route different from the two other processes and *Inhba* gene seems to be the most effective gene on this process. This gene takes part in TGFβ pathway and is thought to have a driver role on suppression of TGFβ pathway in differentiation process from MEFs to MNs.

By examining signaling molecule families in Table 5.1, it is also seen that which protein of each family has a driver role in the processes. For example; genes such as *Inhba* gene in *Inhb* family, *Wnt5a* in Wnt ligands family, *Id4* in *Id* family, *Jak1* in *Jak* family are seen as having decisive roles in the development trajectory compared to other family members. Other members of these families have roles in the functions of same signaling pathway of different cell types, but they have no effect in the processes we examine in the scope of this thesis.

Taking into account the analysis results on the accuracy of determining the “driver genes”, successful aspects and shortcomings of these experiments were also examined. In the experiment done by Briggs et al, it was planned to differentiate cells into motor neurons, but these cells start to lose their MN profiles in day 5. In two experiments done by Treutlein et al, *Ascl1* gene was not enough to take control of iPSC to MN differentiation, but this gene became useful and the differentiation protocol became more effective when combined with *Pou3f2* (*Brn2*) and *Myt1l*. Driver genes that affect neuronal differentiation were studied by comparing driver genes in these three processes.

In the experiment that *Ascl1* gene was triggered only, 35 genes had correlation higher than 0.5; but in the experiment that BAM factors were triggered, 525 genes had correlation higher than 0.5. GO term enrichment analysis of these 525 genes show that these genes are directly related to the neuron development and it was understood that BAM factors tightly regulate neuronal differentiation processes. However, triggering only the *Ascl1* gene could not regulate the processes in a similar way. Genes that have correlation higher than 0.5 in both *Ascl1*- and BAM-triggered experiments were added to Table 5.2 to compare *Brn2*, *Myt1l* and the other datasets. *Gsk3b* is a gene that takes part in pluripotency pathways, is suppressed by Wnt signaling pathway and makes the formation of pluripotency and conservation easier by suppressing the gene using inhibitors. This gene has the correlation coefficient of 0.26 in the analysis where *Ascl1* and BAM experiments of MEF to MN differentiation were evaluated together in Table 5.2 and has high and positive correlation in both experiments. It can be assumed that

Gsk3b gene can make the exit from pluripotency faster when triggered and can have a driver role in neuron formation.

The number of genes that have negative correlation and has a decreasing gene expression level during the differentiation process is 443 in Ascl1 experiment, 304 in BAM experiment, and the number of common genes is 82. Among these genes, “extracellular matrix component” term is significantly common. It is understood that while the cells are losing their MEF profile, the expression of a group of extracellular matrices also decreases and this group also plays a driver role in this process, or is suppressed by the effect of a strong driver regulator.

Table 5.2. Correlation coefficients of the expression of driver genes in terms of pseudotime values in Ascl1 and BAM-triggered experiments.

PP Gene	MEF->MN (ASCL1)	MEF->MN (BAM)	iPSC->MN	MEF->iPSC (A2S)	MEF->iPSC (FBS)
Ascl1	0.642	0.685	0.075	0.000	0.000
Myt1l	0.099	0.638	-0.126	0.016	0.000
Pou3f2	0.056	0.720	0.292	-0.010	0.000
Ncoa1	0.718	0.507	-0.176	-0.041	-0.074
Snca	0.595	0.705	-0.132	0.022	-0.018
Tmem181c	0.585	0.550	0.000	0.000	0.000
Gsk3b	0.584	0.570	-0.150	-0.154	-0.032
Klhl24	0.579	0.505	-0.308	-0.040	0.190
rtTA	0.535	0.760	0.000	0.000	0.000
Prune2	0.533	0.606	-0.185	-0.132	-0.158
Zfp238	0.532	0.506	0.000	0.000	0.000
Gdpd1	0.527	0.618	-0.045	0.004	-0.075
Hsf2	0.524	0.575	-0.657	0.013	-0.088
Fry	0.522	0.691	-0.114	0.098	0.081

Results of these analyses show that there are some genes that drive these processes and considering these genes while developing new differentiation and reprogramming strategies would be highly beneficial to increase the efficiency and efficacy of these processes.

5.4. Comparison of Previous Analyses to Current Analyses and Technical Challenges

This section provides a comparison of the initial analysis of the datasets and the re-analyses performed in this study. The technical limitations of the study are also briefly discussed.

The study that reported the first data set used in this study (Briggs et al., 2017) aimed comparison of standard and direct protocols for differentiation of ESCs to motor neurons (MNs). Three biological questions were answered in the study: 1. How does direct protocol bypass progenitor states, 2. Which intermediates were skipped, and finally, 3. Does direct protocol fully converge to the final state? Trajectories and efficiencies of the two protocols were compared to be able to address these questions. Interestingly, the marker genes were investigated by averaging their expression values in sub-populations and presented as a heatmap of average values, which hindered the potential of the study to reveal the transcriptional heterogeneity of the subpopulations. Relating the gene expression to trajectories (or pseudotimes) was not attempted in the study. Further, the trajectories were not investigated as a continuum of cell states, but instead treated as a series of discrete sub-states. In this re-analysis study, investigation of gene expression data from cells that were exposed to direct protocol revealed not only the subpopulations but their relative states on the pseudotime, genes that are directly correlated to the pseudotime and the heterogeneity of gene expressions.

The study that reported the second data set used in this study (Treutlein et al., 2016) aimed comparison of two protocols for differentiation of MEF cells to MN cells. First protocol involved induction of transcription factor *Ascl1* only, whereas second one involved induction of all three BAM factors. The second protocol led to fewer off-target cells as the two additional factors, *Brn2* and *Myt1l* blocked the pathways which led the cells to myocyte fate. The clusters of cells and GO terms related to genes specific to each cluster were investigated. Two phases were observed during the differentiation process, initiation and maturation phases. In addition, a covariance network of the transcription factors was constructed, the data was partitioned into three stages and hence, three stage-specific networks were constructed. This approach assumes that cells change their regulation programme through out the process, however does not consider the continuity of cell states which was demonstrated in the trajectory analysis. The trajectories of the cells were investigated in relation to

neuronal phenotype, however the quantitative relationship between the pseudotime and gene expression was not studied.

The study that reported the third data set used in this study (Tran et al., 2019), compared two reprogramming processes from MEF to iPSCs, in the first protocol only OSKM factors were induced, in the second protocol, in addition to OSKM induction, inhibitors were used. The second protocol was more efficient as the cells were not able to divert to alternative cell fates. The clusters of cells and GO terms related to genes specific to each cluster was investigated. The trajectories of the cells and day-specific regulation networks were constructed. Once more, although the trajectories indicated a continuum of cell states, genes were not quantitatively investigated with respect to pseudotimes.

In this study, all three datasets were investigated via the same data analysis pipeline, more advanced versions of the tools were used such as use of UMAP instead of tSNE used in the studies described above. Monocle 3 was used for construction of trajectories, which has improvements over the previous versions in terms of data handling and trajectory calculations. Investigation of cell clusters and GO terms enriched among genes specific to each cluster was done as a pre-analysis of the data sets. The trajectories were constructed, their structure allowed analysis of heterogeneity of the routes followed by the cells, which is closely related to the experimental protocol. Processes tightly regulated via the experimental protocol led to less heterogeneous routes whereas loosely regulated processes led to highly heterogeneous routes. Nevertheless, the trajectories revealed that cells states are in a continuum, hence genes which are expressed along that continuum should be driver genes of the process or under direct regulation of the driver genes. Such genes were found by using rank-based Spearman correlation coefficients to avoid linearity assumption. One interesting discovery, not attempted by any of the three studies, was identification of signalling pathway components being expressed in correlation with the pseudotime. This result indicated that during the differentiation and reprogramming processes, the signalling pathways are not only triggered via ligand-receptor interactions or transcription factors but also via transcriptional regulation of the pathway components. Induction of these components in the experimental protocols can further increase efficiency of these processes.

Working with scRNA data has a number of technical difficulties and currently there are no optimal solutions for these difficulties. One such difficulty is integration

of datasets from different sources. This study was initially designed with the aim of integrating the datasets prior to data analysis and performing a unified analysis of all processes, including the trajectory inference steps. However, data from different experiments has to be batch corrected prior to integration, and batch correction methods tested did not produce satisfactory results (results not shown) in pre-analysis steps such as clustering. Therefore, the datasets were analysed separately.

scRNA data is often used to identify the cell types in a tissue or cell culture. However, cell type identification has also some challenges. scRNA data has a large fraction of missing values, therefore use of few well-known cell markers is usually not adequate. Most studies, including the three studies mentioned above, do the cell type assignment solely based on clustering. However, clustering can be biased due to a subset of very variant genes or drop-out events. The analyses in this study demonstrated that clusters may be composed of different cells types (results not shown) and cell type identification has to be done via more sophisticated algorithms rather than clustering. The purpose of this study was not cell type identification, on the contrary, analysis of heterogeneity in the whole population was aimed, hence cell type identification was not done.

Trajectory inference methods are still being improved for better performance. One difficulty in this study was using trajectory inference methods on time series data. The inferred trajectories often violate the natural ordering (order of time) of the cells, indicating that very variant genes or perhaps drop out events heavily influence the algorithm and introduce biases to the results. Hence, in this study the parameters had to be optimized for results that satisfy the natural ordering and reveal the true heterogeneity in the routes. Better performing inference methods need to be developed for more reliable trajectory results.

6. CONCLUSION

Single-cell studies has started to change the perspective of the scientists on how the biological processes can be investigated. Increase in the data resolution improved the interpretation of these processes. In the scope of this thesis, datasets from differentiation and reprogramming processes by single-cell sequencing were retrieved from literature. By using the cellular heterogeneity, it can be seen how single cells behave in each timepoint during the differentiation and cellular states that can not be detected with conventional techniques can be uncovered. Studying single-cell RNA-seq data allows us to reveal how the cells are grouped, what those groups approximately represent, what the main elements that cause this grouping are, which routes the cells follow and what kind of branches form during their transformation and finally, how cellular states represented by these branches are driven by master regulators. This thesis aims to answer these questions.

Datasets studied in the scope of this thesis are created by collecting cells from mESC to MN differentiation, MEF to MN differentiation, MEF to iPSC reprogramming processes. Analysis of mESC to MN differentiation shows that cells start to become neurons in earlier timepoint and the cells evolve to non-neuron cells in later timepoints. Analysis of MEF to MN differentiation takes place efficiently when the cells are induced with BAM factors instead of *Ascl1* factor only. Analysis of MEF to iPSC reprogramming shows that the cells become more dedicated to reprogrammed when A2S-based protocol was applied, and the cells behave with more complexity when FBS-based protocol was applied. Further, an evaluation of the deterministic and stochastic progress of cell states, and regulators of the trajectories were presented.

These results are based on one dataset for each differentiation route due to absence of such studies while this thesis study was conducted. More datasets from the same differentiation routes can be studied in a streamlined workflow to find more robust results. Even though some basic analysis were already done in the studies where these datasets were generated, any additional analysis was not included in these studies. Necessary and more detailed analysis were done in the scope of this thesis.

This thesis focuses on the analysis of differentiation/reprogramming processes by using clustering and trajectory inference. These analyses show the basic characteristics of the transformation processes. More insights can be found by studying

ligand-receptor interactions that occur in each cell, which can unravel what kind of community behaviors these cells have during the differentiation.

REFERENCES

- Altschuler S. J., Wu L. F., (2010), "Cellular heterogeneity: do differences make a difference?", *Cell*, 141 (4), 559-563.
- Bang A. G., Carpenter M. K. (2009), "Characteristics and Characterization of Human Pluripotent Stem Cells". Elsevier Inc.
- Barky A. R. E., Ali E. M. M., Mohamed T. M., (2017), "Stem cells, classifications and their clinical applications", *Am. J. Pharmacol. Ther*, 1 (1), 001-007.
- Becht E., McInnes L., Healy J., Dutertre C.-A., Kwok I. W. H., Ng L. G., Ginhoux F., Newell E. W., (2019), "Dimensionality reduction for visualizing single-cell data using UMAP", *Nature biotechnology*, 37 (1), 38-44.
- Berge K. V. d., Hembach K. M., Sonesson C., Tiberi S., Clement L., Love M. I., Patro R., Robinson M. D., (2019), "RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis", *Annual Review of Biomedical Data Science*, 2 (1), 139-173.
- Bose S., Wan Z., Carr A., Rizvi A. H., Vieira G., Pe'er D., Sims P. A., (2015), "Scalable microfluidics for single-cell RNA printing and sequencing", *Genome Biology*, 16 (1), 120.
- Brennecke P., Anders S., Kim J. K., Kołodziejczyk A. A., Zhang X., Proserpio V., Baying B., Benes V., Teichmann S. A., Marioni J. C., Heisler M. G., (2013), "Accounting for technical noise in single-cell RNA-seq experiments", *Nature Methods*, 10 (11), 1093-1095.
- Briggs J. A., Li V. C., Lee S., Woolf C. J., Klein A., Kirschner M. W., (2017), "Mouse embryonic stem cells can differentiate via multiple paths to the same state", *eLife*, 6, e26945.
- Buganim Y., Faddah D. A., Jaenisch R., (2013), "Mechanisms and models of somatic cell reprogramming", *Nature Reviews Genetics*, 14 (6), 427-439.
- Caglayan S., Ahrens T. D., Cieřlar-Pobuda A., Staerk J. (2019), "Modern Ways of Obtaining Stem Cells". Elsevier.
- Cao J., Packer J. S., Ramani V., Cusanovich D. A., Huynh C., Daza R., Qiu X., Lee C., Furlan S. N., Steemers F. J., Adey A., Waterston R. H., Trapnell C., Shendure J., (2017), "Comprehensive single-cell transcriptional profiling of a multicellular organism", *Science*, 357 (6352), 661-667.
- Cao J., Spielmann M., Qiu X., Huang X., Ibrahim D. M., Hill A. J., Zhang F., Mundlos S., Christiansen L., Steemers F. J., Trapnell C., Shendure J., (2019), "The single-cell transcriptional landscape of mammalian organogenesis", *Nature*, 566 (7745), 496-502.
- Carlson M., org.Mm.eg.db: Genome wide annotation for Mouse., in: (Ed.)^(Eds.), 2019, pp. R package version 3.8.2.

Carrel A., Ebeling A. H., (1926), "THE FUNDAMENTAL PROPERTIES OF THE FIBROBLAST AND THE MACROPHAGE : I. THE FIBROBLAST", *The Journal of experimental medicine*, 44 (2), 261-284.

Chagastelles P. C., Nardi N. B., (2011), "Biology of stem cells: an overview", *Kidney Int Suppl* (2011), 1 (3), 63-67.

Chen G., Ning B., Shi T., (2019), "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis", *Frontiers in genetics*, 10 (317).

Chen H., Albergante L., Hsu J. Y., Lareau C. A., Lo Bosco G., Guan J., Zhou S., Gorban A. N., Bauer D. E., Aryee M. J., Langenau D. M., Zinovyev A., Buenrostro J. D., Yuan G.-C., Pinello L., (2019), "Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM", *Nature Communications*, 10 (1), 1903.

Darnell H. L., Arnold B., Zipursky S. L., Paul M., David B., James (2000), "Overview of Neuron Structure and Function". "Molecular Cell Biology", W. H. Freeman.

Eden E., Navon R., Steinfeld I., Lipson D., Yakhini Z., (2009), "GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists", *BMC Bioinformatics*, 10 (1), 48.

Elbuluk A., Einhorn T. A., Iorio R., (2017), "A comprehensive review of stem-cell therapy", *JBJS Reviews*, 5 (8), 1-9.

Evans M. J., Kaufman M. H., (1981), "Establishment in culture of pluripotential cells from mouse embryos", *Nature*, 292 (5819), 154-156.

Harris M. A., Clark J., Ireland A., Lomax J., Ashburner M., Foulger R., Eilbeck K., Lewis S., Marshall B., Mungall C., Richter J., Rubin G. M., Blake J. A., Bult C., Dolan M., Drabkin H., Eppig J. T., Hill D. P., Ni L., Ringwald M., Balakrishnan R., Cherry J. M., Christie K. R., Costanzo M. C., Dwight S. S., Engel S., Fisk D. G., Hirschman J. E., Hong E. L., Nash R. S., Sethuraman A., Theesfeld C. L., Botstein D., Dolinski K., Feierbach B., Berardini T., Mundodi S., Rhee S. Y., Apweiler R., Barrell D., Camon E., Dimmer E., Lee V., Chisholm R., Gaudet P., Kibbe W., Kishore R., Schwarz E. M., Sternberg P., Gwinn M., Hannick L., Wortman J., Berriman M., Wood V., de la Cruz N., Tonellato P., Jaiswal P., Seigfried T., White R., (2004), "The Gene Ontology (GO) database and informatics resource", *Nucleic Acids Res*, 32 (Database issue), D258-261.

Hou P.-S., Chuang C.-Y., Yeh C.-H., Chiang W., Liu H.-J., Lin T.-N., Kuo H.-C., (2017), "Direct Conversion of Human Fibroblasts into Neural Progenitors Using Transcription Factors Enriched in Human ESC-Derived Neural Progenitors", *Stem Cell Reports*, 8 (1), 54-68.

Ivey M. J., Tallquist M. D., (2016), "Defining the Cardiac Fibroblast", *Circ J*, 80 (11), 2269-2276.

K Takahashi S. Y., (2006), "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors", *Cell*, 126, 663-676.

Kendall R. T., Feghali-Bostwick C. A., (2014), "Fibroblasts in fibrosis: novel roles and mediators", *Frontiers in Pharmacology*, 5, 123.

Kiselev V. Y., Andrews T. S., Hemberg M., (2019), "Challenges in unsupervised clustering of single-cell RNA-seq data", *Nature Reviews Genetics*, 20 (5), 273-282.

Klein Allon M., Mazutis L., Akartuna I., Tallapragada N., Veres A., Li V., Peshkin L., Weitz David A., Kirschner Marc W., (2015), "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells", *Cell*, 161 (5), 1187-1201.

Lowe R., Shirley N., Bleackley M., Dolan S., Shafee T., (2017), "Transcriptomics technologies", *PLOS Computational Biology*, 13 (5), e1005457.

Luecken M. D., Theis F. J., (2019), "Current best practices in single-cell RNA-seq analysis: a tutorial", *Molecular Systems Biology*, 15 (6), e8746.

Macosko E. Z., Basu A., Satija R., Nemesh J., Shekhar K., Goldman M., Tirosh I., Bialas A. R., Kamitaki N., Martersteck E. M., (2015), "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets", *Cell*, 161 (5), 1202-1214.

Maere S., Heymans K., Kuiper M., (2005), "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks", *Bioinformatics*, 21 (16), 3448-3449.

Mah N., Wang Y., Liao M.-C., Prigione A., Jozefczuk J., Lichtner B., Wolfrum K., Haltmeier M., Flöttmann M., Schaefer M., Hahn A., Mrowka R., Klipp E., Andrade-Navarro M. A., Adjaye J., (2011), "Molecular Insights into Reprogramming-Initiation Events Mediated by the OSKM Gene Regulatory Network", *PLOS ONE*, 6 (8), e24351.

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y.-J., Chen Z., Dewell S. B., Du L., Fierro J. M., Gomes X. V., Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L. I., Jarvie T. P., Jirage K. B., Kim J.-B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomasz A., Vogt K. A., Volkmer G. A., Wang S. H., Wang Y., Weiner M. P., Yu P., Begley R. F., Rothberg J. M., (2005), "Genome sequencing in microfabricated high-density picolitre reactors", *Nature*, 437 (7057), 376-380.

Medvedev S. P., Shevchenko A. I., Zakian S. M., (2010), "Induced Pluripotent Stem Cells: Problems and Advantages when Applying them in Regenerative Medicine", *Acta naturae*, 2 (2), 18-28.

Mi H., Muruganujan A., Casagrande J. T., Thomas P. D., (2013), "Large-scale gene function analysis with the PANTHER classification system", *Nature protocols*, 8 (8), 1551-1566.

Mitalipov S., Wolf D. (2009), "Totipotency, Pluripotency and Nuclear Reprogramming". Springer Berlin Heidelberg.

Murray L. A., Knight D. A., Laurent G. J. (2009), "Chapter 15 - Fibroblasts". In: P. J. Barnes, J. M. Drazen, S. I. Rennard and N. C. Thomson, Editors, "Asthma and COPD (Second Edition)", Academic Press.

Poirion O. B., Zhu X., Ching T., Garmire L., (2016), "Single-cell transcriptomics bioinformatics and computational challenges", *Frontiers in genetics*, 7, 163.

Qiu L. Q., Lai W. S., Stumpo D. J., Blackshear P. J., (2016), "Mouse Embryonic Fibroblast Cell Culture and Stimulation", *Bio Protoc*, 6 (13).

Rai M. F., Tycksen E. D., Sandell L. J., Brophy R. H., (2018), "Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears", *Journal of Orthopaedic Research*, 36 (1), 484-497.

Raudvere U., Kolberg L., Kuzmin I., Arak T., Adler P., Peterson H., Vilo J., (2019), "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)", *Nucleic acids research*, 47 (W1), W191-W198.

Reubinoff B. E., Pera M. F., Fong C.-Y., Trounson A., Bongso A., (2000), "Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro", *Nature biotechnology*, 18 (4), 399-404.

Rhee S. Y., Wood V., Dolinski K., Draghici S., (2008), "Use and misuse of the gene ontology annotations", *Nat Rev Genet*, 9 (7), 509-515.

Rosenberg A. B., Roco C. M., Muscat R. A., Kuchina A., Sample P., Yao Z., Graybuck L. T., Peeler D. J., Mukherjee S., Chen W., Pun S. H., Sellers D. L., Tasic B., Seelig G., (2018), "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding", *Science*, 360 (6385), 176-182.

Rowe R. G., Daley G. Q., Induced pluripotent stem cells in disease modelling and drug discovery, in: (Ed.)^(Eds.), *Nature Publishing Group*, 2019, pp. 377-388.

Saelens W., Cannoodt R., Todorov H., Saeys Y., (2019), "A comparison of single-cell trajectory inference methods", *Nature biotechnology*, 37 (5), 547-554.

Shendure J., Porreca G. J., Reppas N. B., Lin X., McCutcheon J. P., Rosenbaum A. M., Wang M. D., Zhang K., Mitra R. D., Church G. M., (2005), "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome", *Science*, 309 (5741), 1728-1732.

Shi Y., Inoue H., Wu J. C., Yamanaka S., Induced pluripotent stem cell technology: A decade of progress, in: (Ed.)^(Eds.), *Nature Publishing Group*, 2017, pp. 115-130.

Sims D., Sudbery I., Ilott N. E., Heger A., Ponting C. P., (2014), "Sequencing depth and coverage: key considerations in genomic analyses", *Nature Reviews Genetics*, 15 (2), 121-132.

Stumpf P. S., Smith R. C. G., Lenz M., Schuppert A., Müller F.-J., Babbie A., Chan T. E., Stumpf M. P. H., Please C. P., Howison S. D., Arai F., MacArthur B. D., (2017), "Stem Cell Differentiation as a Non-Markov Stochastic Process", *Cell systems*, 5 (3), 268-282.e267.

Svensson V., Vento-Tormo R., Teichmann S. A., (2018), "Exponential scaling of single-cell RNA-seq in the past decade", *Nature protocols*, 13 (4), 599-604.

Takahashi K., Yamanaka S., (2006), "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors", *Cell*, 126 (4), 663-676.

Tang F., Barbacioru C., Wang Y., Nordman E., Lee C., Xu N., Wang X., Bodeau J., Tuch B. B., Siddiqui A., Lao K., Surani M. A., (2009), "mRNA-Seq whole-transcriptome analysis of a single cell", *Nature Methods*, 6 (5), 377-382.

Thanapitak S., (2012), "Bionics Chemical Synapse".

The Gene Ontology Consortium, (2018), "The Gene Ontology Resource: 20 years and still GOing strong", *Nucleic acids research*, 47 (D1), D330-D338.

Traag V. A., Waltman L., van Eck N. J., (2019), "From Louvain to Leiden: guaranteeing well-connected communities", *Scientific Reports*, 9 (1), 5233.

Tran K. A., Pietrzak S. J., Zaidan N. Z., Siahpirani A. F., McCalla S. G., Zhou A. S., Iyer G., Roy S., Sridharan R., (2019), "Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency", *Cell Reports*, 27 (6), 1726-1741.e1725.

Tran K. A., Pietrzak S. J., Zaidan N. Z., Siahpirani A. F., McCalla S. G., Zhou A. S., Iyer G., Roy S., Sridharan R., (2019), "Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency", *Cell Rep*, 27 (6), 1726-1741.e1725.

Treutlein B., Lee Q. Y., Camp J. G., Mall M., Koh W., Shariati S. A. M., Sim S., Neff N. F., Skotheim J. M., Wernig M., Quake S. R., (2016), "Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq", *Nature*, 534 (7607), 391-395.

Vailati-Riboni M., Palombo V., Loor J. J. (2017), "What Are Omics Sciences?". In: B. N. Ametaj, Editor, "Periparturient Diseases of Dairy Cows: A Systems Biology Approach", Springer International Publishing.

Vierbuchen T., Ostermeier A., Pang Z. P., Kokubu Y., Sudhof T. C., Wernig M., (2010), "Direct conversion of fibroblasts to functional neurons by defined factors", *Nature*, 463 (7284), 1035-1041.

Vierbuchen T., Ostermeier A., Pang Z. P., Kokubu Y., Südhof T. C., Wernig M., (2010), "Direct conversion of fibroblasts to functional neurons by defined factors", *Nature*, 463 (7284), 1035-1041.

Web 1, (2019), <https://cole-trapnell-lab.github.io/monocle3/docs/trajectories/>, (Access Date: 25/03/2020).

Weber A. P. M., (2015), "Discovering New Biology through Sequencing of RNA", *Plant Physiology*, 169 (3), 1524-1531.

Weinreb C., Wolock S., Klein A. M., (2017), "SPRING: a kinetic interface for visualizing high dimensional single-cell expression data", *Bioinformatics*, 34 (7), 1246-1248.

Yu G., Wang L.-G., Han Y., He Q.-Y., (2012), "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters", *OMICS: A Journal of Integrative Biology*, 16 (5), 284-287.

Zakrzewski W., Dobrzyński M., Szymonowicz M., Rybak Z., (2019), "Stem cells: past, present, and future", *Stem Cell Research & Therapy*, 10 (1), 68-68.

Zhang B., Wolynes P. G., (2014), "Stem cell differentiation as a many-body problem", *Proc Natl Acad Sci U S A*, 111 (28), 10185-10190.

Zhang L., Hu J., Athanasiou K. A., (2009), "The Role of Tissue Engineering in Articular Cartilage Repair and Regeneration", *Critical Reviews™ in Biomedical Engineering*, 37 (1-2), 1-57.

Zhang S., Li X., Lin Q., Wong K.-C., (2020), "Review of Single-cell RNA-seq Data Clustering for Cell Type Identification and Characterization", *arXiv preprint arXiv:2001.01006*.

Ziegenhain C., Vieth B., Parekh S., Hellmann I., Enard W., (2018), "Quantitative single-cell transcriptomics", *Briefings in Functional Genomics*, 17 (4), 220-232.

Zinin N., Adameyko I., Wilhelm M., Fritz N., Uhlén P., Ernfors P., Henriksson M. A., (2014), "MYC proteins promote neuronal differentiation by controlling the mode of progenitor cell division", *EMBO reports*, 15 (4), 383-391.

BIOGRAPHY

Batuhan akır was born in Üsküdar/İstanbul in 1994. He completed his undergraduate studies in Department of Bioengineering at Yıldız Technical University. In 2017, he started to study for his master’s degree in Bioinformatics and Systems Biology Programme in Gebze Technical University. In 2019, he did an internship in Cellular Genetics Informatics Department at the Wellcome Sanger Institute, UK.

APPENDIX

Appendix A: Publications Done In the Scope of This Thesis Study

Çakır B., Prete M., Huang N., Van Dongen S., Pir P., Kiselev V. Y. (2020), “Comparison of visualisation tools for single-cell RNAseq data”, NAR Genomics and Bioinformatics, 2 (3), lqaa052.

Çakır B., Pir P. (2019), “Nöronal hücre farklılaşma süreçlerinin tek hücre transkriptom analizi”, VI. Türk Tıp Dünyası Kurultayı, İstanbul, TÜRKİYE, 29-31 October.

Çakır B., Pir P. (2019), “Single-cell transcriptome analysis of neuronal cell differentiation processes”, The International Symposium on Health Informatics and Bioinformatics 2019, 85-86, İzmir, TURKEY, 17-19 October.

Appendix B: Other Appendices

Supplementary files mentioned in the thesis are provided in the electronic format and listed below:

- Supplementary File 1: Markers of each cluster in mESC → MN process
- Supplementary File 2: GO enrichment analysis terms of the cluster markers in mESC → MN process
- Supplementary File 3: Markers of each cluster in MEF → MN process
- Supplementary File 4: GO enrichment analysis terms of the cluster markers in MEF → MN process
- Supplementary File 5: Markers of each cluster in MEF → mESC process with A2S-based protocol
- Supplementary File 6: GO enrichment analysis terms of the cluster markers in MEF → mESC process with A2S-based protocol
- Supplementary File 7: Markers of each cluster in MEF → mESC process with FBS-based protocol
- Supplementary File 8: GO enrichment analysis terms of the cluster markers in MEF → mESC process with FBS-based protocol

Files that contain the markers of each cluster consist of 6 columns. “p_val” column shows unadjusted p-values calculated with Wilcoxon Rank Sum test, “avg_logFC” column shows log fold-change of the average expression between the cells in the related cluster and all cells. “pct.1” column shows the percentage of cells where the corresponding gene is detected in the related cluster. “pct.2” column shows the percentage of cells where the corresponding gene is detected in the all cells. “p_val_adj” shows adjusted p-values based on Bonferroni correction.

Files that contain the GO enrichment analysis terms of each cluster markers consist of 9 columns. “ID” shows the identifier numbers of the GO term in the corresponding line. “Description” shows what that GO term describes. “GeneRatio” shows the number of genes that represent the related GO term in the given list of genes divided by total number of the genes in the given list for GO enrichment analysis. “BgRatio” shows the number of genes that represent the related GO term in the all genes in the dataset divided by total number of the all genes. “pvalue” shows the p-values of the corresponding term calculated with Fisher’s Exact Test. “p.adjust” shows the adjusted p-values based on Benjamini-Hochberg correction. “qvalue” shows the adjusted p-values based on false discovery rate approach. “geneID” shows the gene symbols in the given list. “Count” shows the number of genes that represent the related GO term in the given list of genes.