

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

**IDENTIFICATION OF TF-MEDIATED DYNAMICS OF
SIGNALLING PATHWAYS
IN STEM CELL REPROGRAMMING AND DIFFERENTIATION
USING NGS DATA**

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

GEBZE
2022

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

**IDENTIFICATION OF TF-MEDIATED
DYNAMICS OF SIGNALLING PATHWAYS
IN STEM CELL REPROGRAMMING AND
DIFFERENTIATION USING NGS DATA**

ENES AK

**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

2022

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**KÖK HÜCRELERİN FARKLILAŞMASINDA
VE GERİ PROGRAMLANMASINDA SİNYAL
YOLLARININ TF-ARACILI DİNAMİKLERİNİN
NGS VERİLERİ KULLANILARAK
TANIMLANMASI**

ENES AK
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ PROGRAMI

DANIŞMANI
DR. ÖĞR. ÜYESİ. PINAR PİR

GEBZE
2022

GEBZE TEKNİK ÜNİVERSİTESİ	YÜKSEK LİSANS JÜRİ ONAY FORMU
----------------------------------	--------------------------------------

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 10/02/2022 tarih ve 2022/10 sayılı kararıyla oluşturulan jüri tarafından 24/02/2022 tarihinde tez savunma sınavı yapılan ENES AK'ın tez çalışması Gebze Teknik Üniversitesi Biyomühendislik Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : DR. ÖĞR. ÜYESİ PINAR PİR

ÜYE

: DR. ÖĞR. ÜYESİ EMRAH NİKEREL

ÜYE

: DR. ÖĞR. ÜYESİ ONUR SERÇİNOĞLU

ONAY

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

İMZA/MÜHÜR

SUMMARY

Induced pluripotent stem cells (iPSCs) are formed by restoring features of the stem cells in somatic cells from an organism. Use of induced pluripotent stem cells holds great promise in medicine, especially regenerative and personalized medicine. iPSCs originate from somatic cells which are reprogrammed into pluripotency via directed methods such as expression of specific transcription factors. iPSC-associated techniques provide tools to understand disease mechanisms, discover drug contents and develop new therapeutic interventions.

Reprogramming of iPSC from somatic cells and differentiation of somatic cells from iPSC takes place by activation of reprogramming mechanisms and their signaling pathways. Expression or repression of transcription factors (TFs) regulate these pathways. The identification of transcription factors and specific genes affecting signaling pathways is an important step in revealing cell differentiation mechanisms. Further insights about cell differentiation mechanisms can be obtained by studying the time-dependent activities of these pathways. Time dependent change in the epigenetic marks, TFs and expression of the specific genes can be profiled using omics technologies such as RNA-Seq and ChIP-Seq.

The main target of this study is a comparison of the differentiation and reprogramming processes and examining the reversibility of these two processes using transcriptome and epigenomics datasets. My aim is to propose new strategies for efficient cell differentiation, test the reversibility of differentiation processes, in addition, to shed light on mechanisms behind these processes.

Keywords: Stem cell, iPSC, Signaling pathways, Transcriptomics, Epigenomics, Bioinformatics

ÖZET

Uyarılmış pluripotent kök hücreler (iPS hücreleri), bir organizmadan somatik hücrelerdeki kök hücrelerin özelliklerinin geri kazandırılmasıyla oluşturulur. Uyarılmış pluripotent kök hücrelerin kullanımı tıpta, özellikle rejeneratif ve kişiselleştirilmiş tıpta büyük umutlar vaat etmektedir. iPSC'ler, spesifik transkripsiyon faktörlerinin ekspresyonunun artırılması gibi metotlarla pluripotensiteye yeniden programlanabilen somatik hücrelerden oluşturulmaktadırlar. iPSC kullanımına dayanan teknikler, hastalık mekanizmalarını anlamak, ilaç içeriklerini keşfetmek ve yeni terapötik müdahaleler geliştirmek için araçlar sağlar.

iPSC'nin somatik hücrelerden yeniden programlanması ve somatik hücrelerin iPSC'den farklılaşması, yeniden programlama mekanizmalarının ve sinyal yollarının aktivasyonu ile gerçekleşir. Transkripsiyon faktörlerinin (TF'ler) ifadesinin artması veya baskılanması bu düzenleyici ağları etkiler. Sinyal yollarını etkileyen transkripsiyon faktörlerinin ve spesifik genlerin tanımlanması, hücre farklılaşma mekanizmalarını ortaya çıkarmak için önemli bir adımdır ve bu yolların zamana bağlı değişimleri incelenerek hücre farklılaşma mekanizmaları hakkında daha fazla bilgi elde etmek mümkündür. Epigenetik işaretlerde, TF'lerde ve spesifik genlerin ifade seviyelerindeki zamana bağlı değişiklikler, RNA-Seq ve ChIP-Seq gibi omik teknolojileri kullanılarak ortaya konabilir.

Bu çalışmanın temel amacı, farklılaşma ve yeniden programlama süreçlerinin karşılaştırılması ve bu iki sürecin transkriptom ve epigenomik veri setleri kullanılarak tersinirliğinin incelenmesidir. Amacım verimli hücre farklılaşması için yeni stratejiler önermek, farklılaşma süreçlerinin tersine çevrilebilirliğini test etmek ve ayrıca bu süreçlerin arkasındaki mekanizmalara ışık tutmaktır.

Anahtar Kelimeler: Kök Hücre, iPSC, Sinyal Yolakları, Transkriptomiks, Epigenomiks, Biyoinformatik

ACKNOWLEDGEMENTS

I would like to express my deepest and sincere gratitude to Asst. Prof. Pınar Pir who has always supported me and allowed me to continue this path. She reminded me that in all the times I was in despair, there was always an option. I feel very lucky to have her as my advisor. I also would like to thank TUBITAK for funding this project.

I want to thank our research team and office mates for their collaborations and cheerful faces. Also, I would like to thank Gülben AVŞAR for ChIP-Seq data pre-processing.

Finally, I would like to express my gratitude to my family, who always believed and trusted me.

TABLE of CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
TABLE of CONTENTS	viii
LIST of ABBREVIATIONS	xi
LIST of FIGURES	xii
LIST of TABLES	xvi
1. INTRODUCTION	1
2. BACKGROUND ASPECTS	3
2.1. Differentiation of Stem Cells	3
2.2. Reprogramming of Stem Cells	4
2.3. Signaling Pathways	5
2.3.1. WNT/ β -catenin Signaling Pathway	6
2.3.2. Notch Signaling Pathway	7
2.3.3. TGF- β Signaling Pathway	9
2.3.4. Sonic Hedgehog Signaling Pathway	10
2.3.5. BMP Signaling Pathway	12
2.4. MET and EMT Processes	13
2.5. System Biology	14
2.6. Omics Technologies	14
2.6.1. Microarrays	15
2.6.2. RNA-Sequencing	16
2.6.3. Networks in Biological Systems	18
3. MATERIAL AND METHODS	20
3.1. Datasets	20
3.1.1. Datasets 1-2 / GSE117204 – GSE117203	20
3.1.2. Dataset 3 / GSE137001	21

3.1.3. Datasets 4-5 / GSE67462 – GSE67520	21
3.1.4. Dataset 6 / GSE111831	22
3.1.5. Dataset 7 / GSE50206	22
3.2. Pre-analysis inspection with PCA	22
3.3. Software Tools for Transcriptomics Data Analysis	24
3.3.1. RNA-Seq Data Analysis	24
3.3.1.1. RNA-Seq Data Pre-processing	24
3.3.1.1.1. FastQC	24
3.3.1.1.2. Trim Galore	25
3.3.1.1.3. HISAT2	26
3.3.1.1.4. Samtools	27
3.3.1.1.5. FeatureCounts	27
3.3.1.2. RNA-Seq Data Downstream Analysis	28
3.3.1.2.1. DESeq2	28
3.3.1.2.2. K-means clustering	28
3.3.1.2.3. Cytoscape	29
3.3.1.2.4. Dynamic Regulatory Events Miner (DREM)	29
3.3.1.2.5. KEGG	30
3.3.2. Microarray Data Analysis	30
3.3.2.1. Limma	31
4. RESULTS	32
4.1. iPSC to MEF Differentiation in Mice	32
4.2. MEF to iPSC Reprogramming in Mice	36
4.3. iPSC to Astrocyte Differentiation in Human	39
4.4. Astrocyte to iPSC Reprogramming in Human	43
4.5. Reversibility of Reprogramming and Differentiation Processes in Pluripotency Signaling Pathway	45
4.5.1. Reversibility of Pluripotency Between Reprogramming and Differentiation Process in Mice	46
4.5.2. Reversibility of Pluripotency Between Reprogramming and Differentiation Process in Human	48

4.6. Reversibility of the Epithelial-Mesenchymal Transition in Differentiation and Reprogramming	49
4.7. Comparison of the Signaling Pathways in Differentiation and Reprogramming Processes in Mice	54
4.8. Reversibility of the Signaling Pathways in Differentiation and Reprogramming Processes	56
4.9. Effects of Epigenetic Factors in MEF-iPSC Reprogramming Processes	58
4.10. Effects of Epigenetic Factors on Signaling Pathways in MEF-iPSC Reprogramming Processes	64
4.11. Comparison of the Epigenome and Transcriptome Data from the Same Process on Signaling Pathways	66
4.12. Comparison of Reversibility by Transcriptome and Epigenome Data	70
5. DISCUSSION	73
5.1. Reprogramming and Differentiation of MEF	73
5.2. Reprogramming and Differentiation of Astrocytes	74
5.3. Effects of Epigenetic Factors to Signaling Pathway in Reprogramming and Differentiation	76
5.4. Outputs of this Thesis That Can Contribute to the Literature	76
5.4.1. Correlation Between Pluripotency Signaling Pathway Gene Expressions	76
5.4.2. Reversibility of MET-EMT in Differentiation and Reprogramming	78
5.4.3. Reversibility of Signaling Pathways	79
5.4.4. Reversibility of Epigenetic Factors for Reprogramming and Differentiation	81
5.5. Strategies for Future Studies	81
6. CONCLUSIONS	83
REFERENCES	85
BIOGRAPHY	93
APPENDICES	94

LIST of ABBREVIATIONS

<u>Abbreviations</u>	<u>Descriptions</u>
BAM	: Binary Alignment Map
BP	: Biological Process
ChIP-Seq	: Chromatin Immunoprecipitation Sequencing
DREM	: Dynamic Regulatory Events Miner
ESC	: Embryonic Stem Cell
Hh	: Hedgehog signaling pathway
iPSC	: Induced Pluripotent Stem Cells
KEGG	: Kyoto Encyclopedia of Genes and Genome
LIF	: Leukemia Inhibitory Factor
MEF	: Mouse Embryonic Fibroblast
NICD	: Notch Intracellular Domain
NGS	: Next Generation Sequencing
PE	: Paired End
SAM	: Sequence Alignment Map
SE	: Single End
TGF- β	: Transforming Growth Factor- β
WNT	: Wingless-related integration site

LIST of FIGURES

<u>Figure No:</u>	<u>Page</u>
2.1: A pictorial version of depiction of Waddington.	3
2.2: According to Waddington's landscape, depiction of cellular reprogramming.	5
2.3: WNT signaling pathway.	7
2.4: Notch signaling pathway.	8
2.5: TGF- β signaling pathway.	10
2.6: The crosstalk between Shh pathway and other signaling pathways.	11
2.7: Smad-dependent pathway and non-smad pathways.	13
2.8: Workflow of microarray.	16
2.9: Overview of the RNA-Seq workflow.	17
3.1: iPSC to MEF differentiation experiment design by Thakurela et al.	21
3.2: MEF to iPSC reprogramming experiment design by Velychko et al.	21
3.3: PCA charts of the Velychko et al 2019 and Thakurela et al 2019 datasets.	23
3.4: An example quality report generated by FastQC.	25
3.5: Cytoscape interface.	29
3.6: DREM2 interface.	30
4.1: Clustering result of Thakurela et al dataset.	32
4.2: GO BP analysis results of selected gene clusters of iPSC to MEF process.	34
4.3: The result of co-regulatory analysis of significantly changed genes of Thakurela et al dataset.	35
4.4: Genes that correlate in the iPSC to MEF process and interact in Pluripotency signaling pathways.	36

4.5:	Clustering result of the Velychko et al dataset.	37
4.6:	GO BP analysis results of selected gene clusters of MEF to iPSC process.	38
4.7:	The result of co-regulatory analysis of significantly changed genes of Velychko et al dataset.	39
4.8:	Clustering result of Tekin et al dataset.	40
4.9:	GO BP analysis results of selected gene clusters of the iPSC to astrocyte process.	41
4.10:	The result of co-regulatory analysis of significantly changed genes of Tekin et al dataset.	42
4.11:	Clustering result Takahashi et al dataset.	43
4.12:	GO BP analysis results of selected gene clusters of astrocytes to iPSC process.	44
4.13:	The result of co-regulatory analysis of significantly changed genes of the Takahashi et al dataset.	45
4.14:	Pluripotency and related signaling pathways were mapped with significantly changed genes with increased expression level in mice.	46
4.15:	Pluripotency and related signaling pathways were mapped with significantly changed genes with decreased expression level in mice.	47
4.16:	Pluripotency and related signaling pathways were mapped with significantly changed genes with increased expression level in human.	48
4.17:	Pluripotency and related signaling pathways were mapped with significantly changed genes with decreased expression level in human.	49
4.18:	Expression profiles of MET and EMT genes in MEF – iPSC processes	50
4.19:	Expression profiles of MET and EMT genes in iPSC - MEF processes	51

4.20:	EMT and MET genes in the process of reprogramming from human astrocyte cells to stem cells (Takahashi et al. 2014).	52
4.21:	MET and EMT marker genes in the process of differentiation from stem cells to human astrocytes (Tekin et al. 2016).	53
4.22:	WNT and TGF- β KEGG signaling pathways.	54
4.23:	Notch signaling pathway.	55
4.24:	Activation and suppression of signaling pathways at time intervals during differentiation and reprogramming processes in both organisms, human and mouse.	57
4.25:	Clustering result of Chen et al dataset.	59
4.26:	GO BP analysis results of selected gene clusters.	60
4.27:	The result of co-regulatory analysis of significantly changed genes of Chen et al. dataset.	61
4.28:	Chen et al. ChIP-Seq dataset clustering results.	62
4.29:	GO BP analysis results of selected gene clusters of Chen et al ChIP-Seq dataset.	63
4.30:	Heatmap showing the amount of acetylation in the process of genes involved in Pluripotency, WNT, Notch, TGF- β signaling pathways in the reprogramming and differentiation processes.	65
4.31:	Signaling pathway mapped with significantly changed genes with increased expression levels. The dataset was generated by Chen et al., 2019.	67
4.32:	Signaling pathway mapped with significantly changed genes with decreased expression levels. The dataset was generated by Chen et al., 2019.	68
4.33:	Signaling pathway mapped with significantly changed genes with increased expression levels. The datasets generated by Thakurela et al., 2019.	69
4.34:	Signaling pathway mapped with significantly changed genes with decreased expression levels. The datasets generated by Thakurela et al., 2019.	70

4.35:	Visualization of activation and inactivation at time intervals of signaling pathways in the mouse organism during differentiation and reprogramming processes.	71
5.1:	The visualization of the MET-EMT processes in mice.	79



LIST of TABLES

<u>Table No:</u>		<u>Page</u>
3.1:	Datasets used in this thesis.	20
4.1:	Marker genes selected from relevant signaling pathways found in human.	56
4.2:	Marker genes selected from relevant signaling pathways found in mouse.	56

1. INTRODUCTION

Stem cells are undifferentiated cell types that have self-renewing feature and also that can differentiate into functional cells such as heart, fibroblast, astrocyte cells. Thanks to their self-renewal feature, undifferentiated stem cells are continuously regenerated, hence the stem cell population is preserved. Stem cells have different levels of differentiation potential, and they can be classified based on their differentiation potential such as Totipotent, Pluripotent, Multipotent and Unipotent stem cells. For instance, according to literature, zygote and blastomeres are totipotent, therefore they can differentiate to all embryonic and extra-embryonic tissues [1]. However, Unipotent and Multipotent stem cells have limited self-renewing ability and they can differentiate into only specific cell types. In addition, pluripotent stem cells (PSCs) can differentiate into all three germ cell layers, therefore they can differentiate into all types of cells in the organism however they cannot give rise extra embryonic cell types such as placental cells [2].

Stem cell research and its clinical applications become more valuable in the recent years [3]. Stem cells can be found in every tissue in the body and facilitate tissue regeneration. Stem cell technology is used in treatment of different types of diseases and disorders such as type 1 diabetes or macular degeneration of retina [4]. Injured tissues are healed by regeneration, therefore stem cells are very important components of the body [5].

Embryonic stem (ES) cells can be obtained from the inner cell mass of mammalian blastocysts. However, there are ethical problems related to the use of human embryos in medicine, in addition tissue rejection may also be seen during stem cell therapy, after transplantation. These problems can be overcome with the generation of pluripotent stem cells from the patients' own cells [6]. Induced pluripotent stem cells (iPSCs) reprogrammed from somatic cells have features similar to those of embryonic stem cells. These features are self-renewal and ability to differentiate into all somatic cell types [7]. iPSC technology is expected to innovate the regenerative medicine in the near future.

Although iPSCs seems to be the future of regenerative medicine, the efficiency of reprogramming is low, around 0.1% to 3% [8]. Although many methods have been developed for reprogramming and transformation processes of the cells, the relationship

between signaling pathways and epigenetic structures that orchestrate the process has not yet been understood [9].

In this thesis study, systems biology approach was used. Systems biology is a new branch compared to other biological sciences. Systems biology investigates cells, tissues and organs as a system rather than focusing on the characteristics of isolated parts of a cell or organism [10]. High-throughput omics data, such as proteomics, transcriptomics and epigenomics are often utilized in system biology. Thanks to omics technologies, large amount of biological data is produced. These data provide information on thousands of components of the biological systems simultaneously and can be used to compare the cell and tissue profiles [11].

In this thesis study, time course omics datasets from NCBI GEO public databases were used to examine reversibility of two different processes and the effect of signaling pathways such as WNT, Notch on these processes.

2. BACKGROUND ASPECTS

2.1. Differentiation of Stem Cells

Conrad Waddington has published his famous depiction of likening the process of differentiation to a ball rolling down a hill in 1957. In this depiction, the ball on top of the hill represents the stem cell, the depiction can be seen in Figure 2.1 [12]. Embryo contains cells that have the potential to differentiate into all mature cell types in the body and as the cells differentiate, they lose their potential of differentiation, similar to loss of potential energy as the ball rolls down the hill [13].

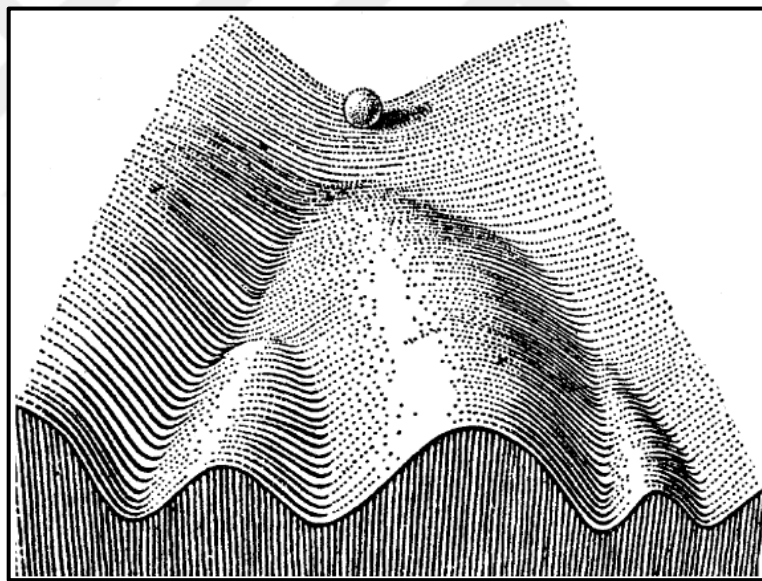


Figure 2.1: A pictorial version of depiction of Waddington.

Stem cell differentiation is crucial for organogenesis and tissue homeostasis. It is an active research area to investigate the differentiation mechanisms to create *in vitro* models of these processes. This will allow potentially unlimited access to certain tissues [14]. Stable human ESC lines were generated in 1998 for the first time and they were cultured while maintaining their pluripotency successfully. Spontaneous differentiation of ESCs *in vitro* occurs, a result of discarding the pluripotency features from their culture's environment. Several studies have shown that downregulation or upregulation of specific

transcription factors, depending on the cell type to which the stem cell will differentiate, affects the specific metabolic and signaling pathways, that lead to differentiation of stem cells [15]. While overexpression of particular genes trigger differentiation, silencing of certain genes such as Sox2, Nanog, Oct4 also support the differentiation process [14].

Biological cues are required for stem cell differentiation. These cues are controlled by different types of cell-soluble factors [16]. Differentiating cells are in perpetual communication with their environment, receiving and sending different types of signals. Signaling pathways such as wingless (WNT), transforming growth factor- β (TGF- β) and Notch pathways regulates the differentiation processes as a response to biological cues received by the cells [17]. Moreover, these pathways also orchestrate different types of processes such as cell proliferation in addition to differentiation of stem cells [18].

2.2. Reprogramming of Stem Cells

More than 50 years ago, nuclear reprogramming technology was established. Somatic cells can be rejuvenated by erasing epigenetic memories whereby new pluripotent features are generated in the cell [19]. Although in recent years, cell reprogramming processes has been under investigation, mechanisms of reprogramming are still largely unknown [15].

John Gurdon and his colleagues have produced clones from somatic cells of African clawed frogs that latin name is *Xenopus laevis* [20]. Davis and his colleagues have discovered the myogenic differentiation 1 (MYOD1) gene by using complementary DNA (cDNA). According to this study, overexpression of MYOD1 is enough to facilitate reprogramming of fibroblasts to myoblasts [21]. The study by Davis and his colleagues has been an encouraging in demonstrating that the pluripotent cells can be produced without the need for fertilized eggs.

As a major milestone, Ian Wilmut and his colleagues have reported the cloning of a sheep, globally renowned Dolly. This achievement has shown that removal of the epigenetic memories that alter somatic cell fate can be possible in mammals too [22]. In this way, reprogramming of a cell to the pluripotent state has been made possible by the fusing of embryonic stem cells and somatic cells. According to these approaches,

fertilized eggs and pluripotent cells contain "reprogramming factors" that can remove the memories of somatic cells.

A decade later, Takahashi and Yamanaka achieved successful induction of pluripotency in mouse embryonic fibroblasts (MEF) via over expression of four embryonic transcription factors (Oct4, Sox2, Klf4, c-Myc) in 2006. These transcription factors are currently known as OSKM or “Yamanaka factors”. Reprogramming mechanism of somatic cells into embryonic stem cells was awarded with the Nobel Prize in Medicine for Yamanaka in 2012, shared with John B. Gurdon [8]. An illustration of reprogramming can be seen in Figure 2.2 [19].

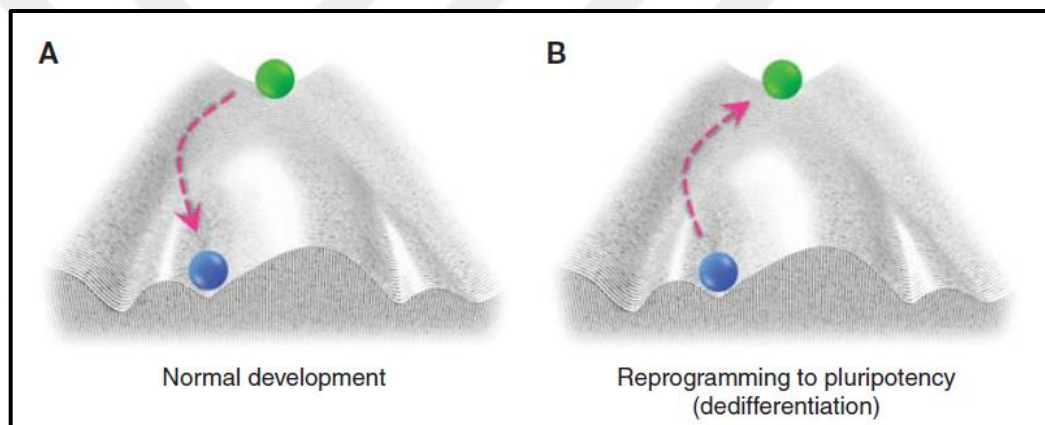


Figure 2.2: According to Waddington's landscape, depiction of cellular reprogramming. A) Normal development of a cell. Development starting with pluripotent cell (green ball) at the top of the hill and proceed to differentiated state (blue ball). B) A differentiated cell (blue ball) can be reprogramed into the pluripotent cell (green ball) by using the Yamanaka transcription factors.

2.3. Signaling Pathways

Differentiating cells are in constant communication with their neighbours. Different types of signals are received and sent during the differentiation process. Differentiation processes are managed by signaling pathways [17]. WNT, TGF- β , Notch are some of the pathways that affect cell fate during differentiation. In addition, these pathways also regulate different types of processes such as cell proliferation and reprogramming of cells [18].

2.3.1. WNT/ β -catenin Signaling Pathway

WNT/ β -catenin signaling pathway is one of the essential pathways for embryo development and adult tissue homeostasis in most tissues and organ systems [23]. WNT gene was discovered firstly as Int1 in mice [24, 25]. Then a homolog gene of Int1 was discovered in *Drosophila melanogaster*, presence of this gene leads to wingless phenotype in *D. melanogaster*. After this discovery, name of the Int1 gene was changed as WNT [23, 26]. Further studies revealed that, WNT proteins control a canonical signaling pathway together with β -catenin, hence this pathway is called as WNT/ β -catenin signaling pathway [27].

WNT/ β -catenin signaling pathway was highly conserved in evolution. The signaling pathway plays an important role in many different types of biological processes such as development, growth, stem cell maintenance and metabolism. Dysregulation of WNT/ β -catenin pathway may lead to diseases such as cancer, diabetes and obesity [28].

WNT/ β -catenin signaling pathway plays an active role throughout the basilar papilla (BP) in hair cells (HCs) and their supporting cells. Jacques and his colleagues have conducted an experiment to reveal the function of WNT/ β -catenin pathway in developing HCs. In this experiment, team of Jacques performed gain and loss of function experiments of WNT in the chick BP and zebrafish lateral line system, respectively. While, proliferation and HC formation are decreased as a result of pharmacological inhibition of WNT signaling pathway with WNT antagonist IWR-1 chemical in BP and lateral line, pharmacological activation of this pathway significantly increased the number of HCs in the lateral line and BP [29].

It has been reported that there is a significant correlation between dysregulation of the WNT signaling pathway and different types of diseases. Also, WNT signaling pathway plays an important role in embryogenesis and cell fate, hence, better understanding the WNT signaling pathway can be a crucial gain for medicine and developmental biology. An illustration of WNT signaling pathway can be seen in Figure 2.3 [30].

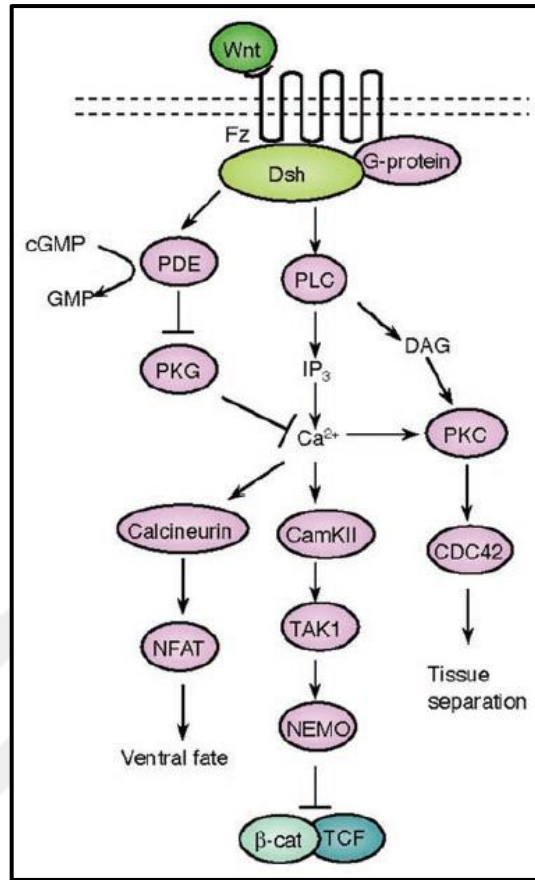


Figure 2.3: WNT signaling pathway.

2.3.2. Notch Signaling Pathway

Cell proliferation and differentiation are regulated by several signaling pathways, one of them is Notch signaling pathway. The Notch signaling pathway was highly conserved across species. It determines the cell fate, thereby its activity is crucial especially in embryonic development [31]. In mice and human, four single-pass transmembrane Notch receptors are located. These receptors are activated after binding of the ligands Dll1, Dll2 or Dll3, to receptors Jagged1 or Jagged2 of neighbouring cells [32]. As a result of the interaction between receptors and ligands, proteolytic cleavage of a Notch receptor via the g-secretase complex takes place. Then conformational change occurs in the Notch receptor and the Notch intracellular domain (NICD) is released from the cellular membrane into the cytoplasm [31]. Notch signaling pathway can be seen in Figure 2.4 [33].

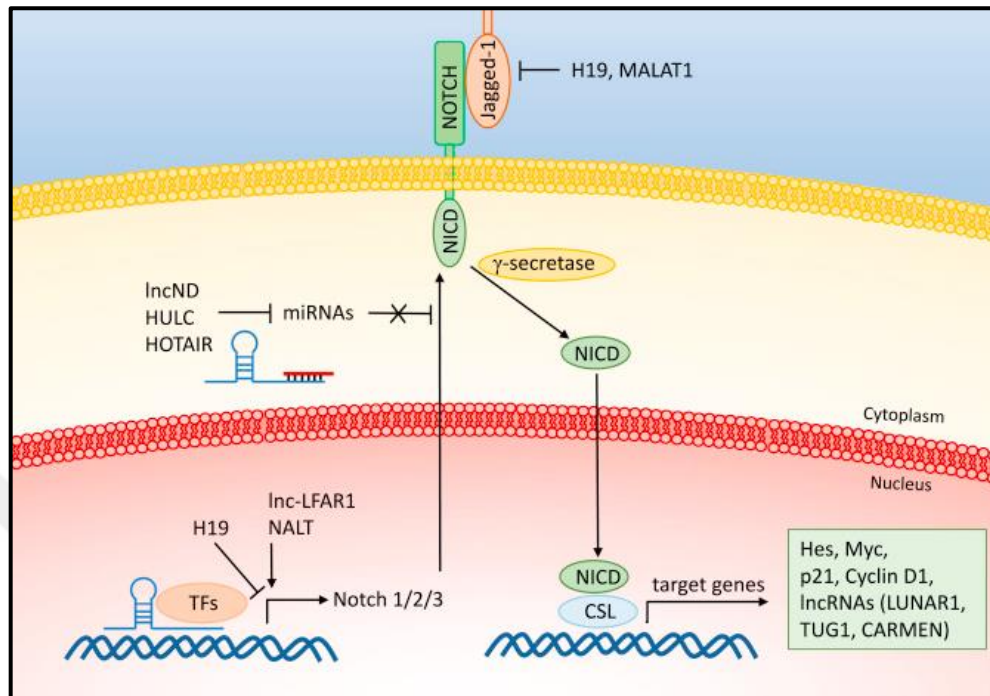


Figure 2.4: Notch signaling pathway.

Signaling of Notch1 is shown to have negative correlation with cell growth and motility in mouse embryonic fibroblasts. In addition, enforced overexpression of Notch1 leads to decrease of proliferation of human dermal fibroblast. According to Liu and his colleagues' study, in the proliferating fibroblasts, there are not any significantly change in gene expression in Notch signaling. On the contrary, the expression of Notch1, Notch3, Notch4, Dll-1, Jagged-2, ADAM-17 and Hey-1 were significantly up-regulated in the G0 phase of the fibroblasts [34].

Notch signaling pathway is crucial for many processes, such as development, repair and homeostasis of the tissues. Further work on the cell-type-specific roles of the Notch signaling pathway can help us understand the fundamentals of many disorders and developmental abnormalities [31].

2.3.3. TGF- β Signaling Pathway

More than three decades ago the TGF- β family was discovered and its biological activity has been a pivotal topic in T-cell survival and proliferation research [35]. Also, TGF- β signaling has a crucial role in regulating tissue development, homeostasis, repair and host immunity. Dysregulation of the TGF- β signaling pathway leads to many diseases such as fibrotic diseases, cancer and connective tissue disorders [36]. The mammalian TGF- β isoforms are TGF- β 1, β 2, β 3 but the key member in TGF- β superfamily is Transforming growth factor-beta1 (TGF- β 1) [37, 38]. Secreted factors' TGF- β bind and activate cell surface complexes of receptors that are classified as type I and type II. This classification is performed according to their sequence similarities. TGF- β type I and TGF- β type II receptors form a complex. As a result of this binding, phosphorylation and activation of type I receptors take place. Thus receptor regulated RSmads can bind to receptors [35]. Then, R-Smads are activated via phosphorylation at carboxy-terminal serines. Subsequently, receptor complexes release these R-Smads. Active R-Smads are transported into the nucleus and form transcriptional complexes with Smad4 and different types of DNA-binding cofactors, also non-DNA-binding co-activators and co-repressors. These complexes regulate the expression of particular target genes [35, 39]. TGF- β signaling pathway can be seen in Figure 2.5 [35, 38].

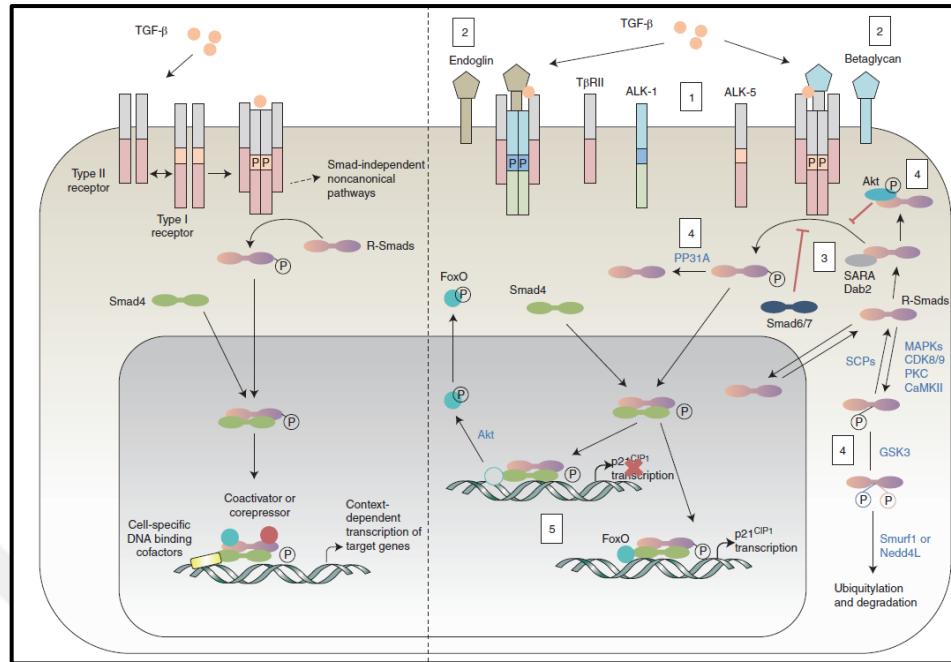


Figure 2.5: TGF- β signaling pathway.

In the recent years, our understanding of stem cells and developmental biology improving rapidly. Advancing our knowledge on functions of TGF- β pathways will aid both basic research of stem cells and their potential applications [40].

2.3.4. Sonic Hedgehog Signaling Pathway

Hedgehog signaling pathway (Hh) was first discovered in *Drosophila*. Hh plays a key role in embryonic and postnatal development. Generally, it regulates different types of processes such as cell proliferation, tissue repair and differentiation [41]. There are three mammalian Hh proteins that are Shh, Indian-Hedgehog (Ihh) and Desert-Hedgehog (Dhh). While Ihh supports the development of skeletal system, Shh has roles in the nervous system. Functions of Dhh are generally limited with gonad cells, for instance, Dhh is active in granulosa cells of ovaries and sertoli cells of the testis [42].

Hedgehog signaling pathway plays an important role in initiation of fibrosis *in vivo* and *in vitro* [43]. According to the study of Horn and her colleagues, overexpression of Shh leads to hedgehog pathway activation with accumulation of the Gli-2 and increased expression of target genes of hedgehog. Shh leads to release of collagen in cultured

fibroblasts and supports differentiation of fibroblasts into myofibroblasts [44]. There are two activation mechanisms of Shh pathway. One of them is canonical signaling activation that is receptor-induced signaling and the other one is non-canonical signaling activation which occurs through Gli-independent mechanisms [42, 45]. There are two types of Gli-independent mechanisms: In type I, Smo modulates Ca^{2+} and actin cytoskeleton and in type II, cell proliferation and survival are increased, independent of Smo [42].

Many biological processes such as stem cell self-renewal, cell growth and differentiation are regulated by a number of major signaling pathways. These pathways work together with each other and orchestrate the cell fate [41]. Performed studies show that Shh signaling pathway interacts with other signaling pathways, such as TGF- β , Notch and WNT/ β -catenin. Mechanisms of interaction of Shh with the other signaling pathways can be seen in Figure 2.6 [42].

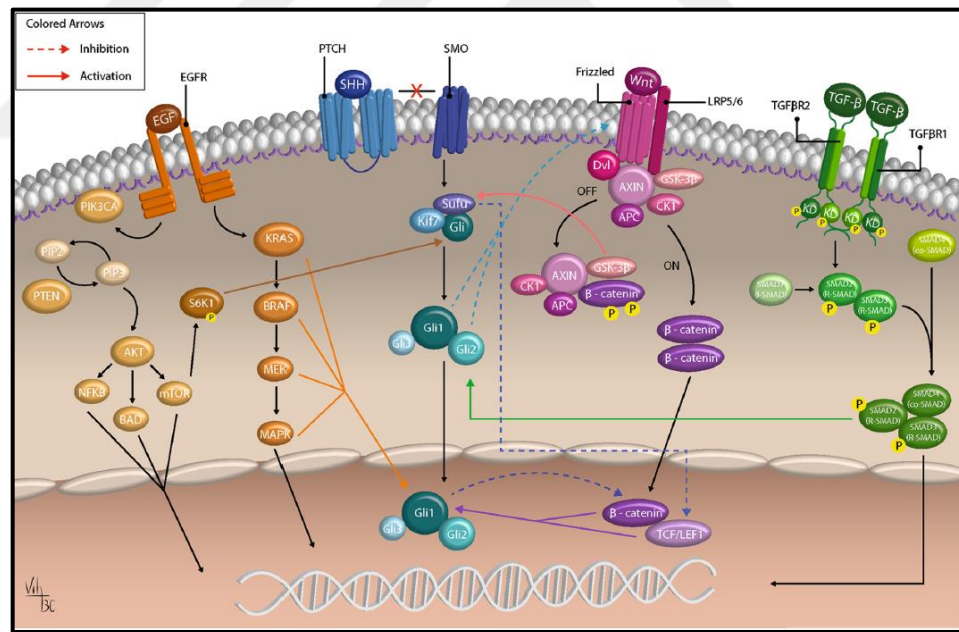


Figure 2.6: The crosstalk between Shh pathway and other signaling pathways.

The research on these complex crosstalk mechanisms would shed light on unknown interactions, for the better understanding of developmental biology and increasing the efficiency of stem cell differentiation processes.

2.3.5. BMP Signaling Pathway

Although Bone Morphogenetic Proteins (BMPs) is a subgroup of the TGF- β signaling pathway, it has a large number of components and plays an important role in many aspects of the cell-cell interaction in higher organisms. BMPs perform crucial functions throughout embryonal development, in addition, they are also involved in regeneration and tissue homeostasis in adult organisms [46].

BMPs were isolated for the first time from demineralized bone matrix. At the end of the 1980s, the BMPs were purified to homogeneity [47]. There are 33 TGF- β member genes of TGF- β and 15 genes from these are also BMP related [48]. There are 8 Smad proteins in mammals (Smad1 through Smad8, Smad8 is also known as Smad9). Receptor-regulated Smads (Smad1, Smad5, Smad8) are phosphorylated by activated BMP type I. Activated and phosphorylated R-Smad (Smad1, Smad5, Smad8 are also known as R-Smad) proteins create a complex with Smad4 and then get transported into the nucleus. During the differentiation of osteoblast, phosphorylation and nuclear accumulation of Smads are induced by BMP-6. Smad1 and Smad5 are involved in this mechanism, while Smad6 is not [48, 49].

BMPs also activate several types of signaling via Smad-independent pathways. For instance, c-Jun amino-terminal kinase (JNK), mitogen-activated protein kinases (MAPKs), phosphoinositol-3 kinase (PI3K), small GTPases and Akt are among targets of BMPs [50]. Smad-dependent and Smad-independent BMP signaling pathway can be seen in Figure 2.7 [47]. Many BMPs are expressed in a variety of tissues during embryogenesis but after birth, some BMPs become limited to specific tissues. For instance, while BMP-3, -4, -5, and -6 are expressed in lung, BMP-7 is highly expressed in kidney in adult mice [51]. BMP-3 is highly expressed by osteocytes and osteoblasts in mice. During fracture healing, BMP-4 expression is transiently induced in callus-forming cells in the early phase. BMP-6 is abundantly expressed in hypertrophic chondrocytes during endochondral ossification [48]. Smad-dependent and independent BMPs signaling pathway can be seen in Figure 2.7 [47].

Medical use of inhibitors or/and activators of BMP signaling pathway will provide the new treatments for different types of diseases. Thus, better understanding of the BMP

signaling pathway may lead to better treatment methods with minimal or no unwanted side effects [46, 48].

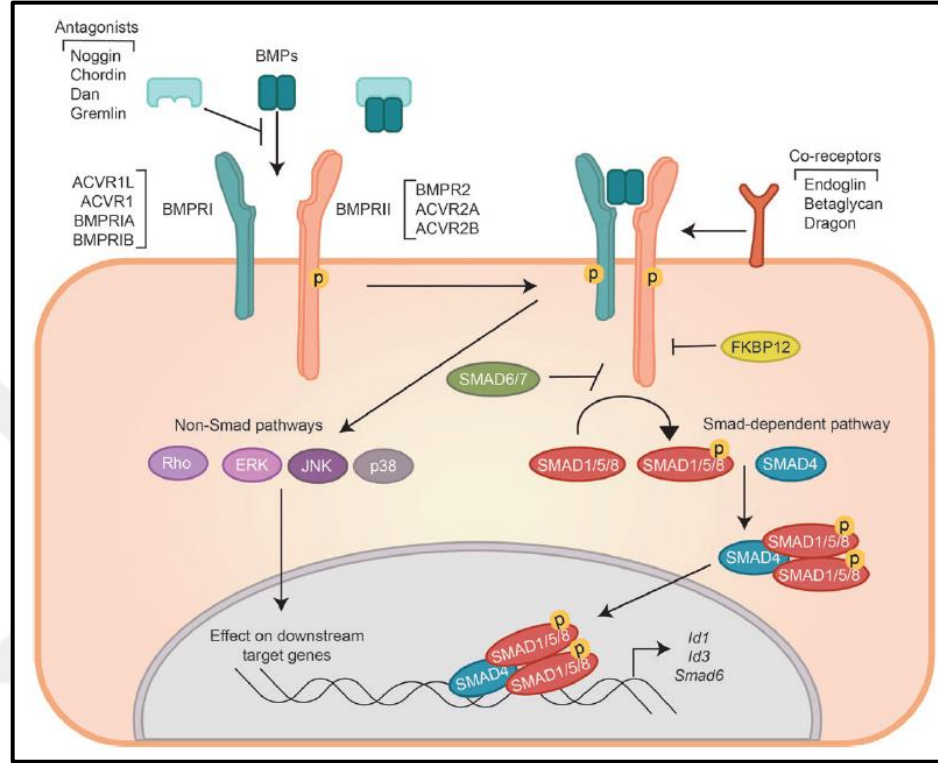


Figure 2.7: Smad-dependent pathway and non-smad pathways.

2.4. MET and EMT Processes

EMT is a process in which epithelial cells lose their cell polarity and adhesion between each other. In this process, epithelial cells gain the migratory features required for mesenchymal cell phenotype. EMT takes place in embryogenesis, and this process has been described as a gradual process of differentiation [52]. These changes are caused by the action of several transcription factors, such as Snai1 (Snail), Snai2 (Slug), Twist, ZEB1, and ZEB2, which are potent repressors of epithelial gene expression [52]. These transcription factors and other additional TFs are also surveyed in this thesis to find relationships between EMT-differentiation and MET-reprogramming processes.

On the contrary, many observations from the reprogramming experiments also provide evidence that reprogramming results when EMT factors are expressed.

Mesenchymal gene expression increases early in fibroblast reprogramming. Early induction of EMT, either by sequential application of reprogramming factors or transcription factors that mentioned before related with EMT process improves reprogramming efficiency [53]. These contradictory results show us that MET and EMT processes are important in reprogramming and differentiation processes, but their dynamics is still a controversial issue in the literature.

2.5. System Biology

The term “systems biology” has been put forward to define the cross-disciplinary research in biology. This term became popular in the science world in 2000s with completion of the Human Genome Project and emergence of the omics [10, 54]. System biology for some remains to be limited to the genomic field, but for many researchers, it is a tool to understand vast amounts of biological data and investigate the complexity of the living things [55].

The main aim of system biology is to study the genes, transcripts of genes and their protein products together with the interactions between them. System biology is an interdisciplinary field in biology, therefore, include different types of omics technologies such as genomics, proteomics, transcriptomics, etc. In addition, various types of new "omics" technologies are defined recently [56]. Most system biology approaches include computational and mathematical modelling. In addition, the development and maintenance of software tools for system biology has a significant role in development of this field. Thanks to system biology, researchers can build a bridge between computational and experimental biology [57].

2.6. Omics Technologies

Genetics only examines the individual characteristics and specific variants of genes [58]. If "-omics" suffix is added to a molecular term, emerging new term includes a comprehensive assessment of a set of molecules. For example, unlike ‘genetics’ science,

'genomics' focuses on the study of the whole genome, genomics is the first omics technology used in systems biology field.

In recent years, biological data generation rate rapidly increased as a result of new high throughput appliances and omics technologies which become crucial for scientific advances [59]. Via the high throughput techniques, omics data such as epigenomes, transcriptomes and genomes can be generated. Especially next-generation sequencing (NGS) technologies affected the medical and scientific research landscape dramatically. For example, NGS can be used to sequence exomes or genomes rapidly and relatively at low cost [11, 60].

As it was mentioned above, if "omics" term is added to a molecular term, it refers to the comprehensive assessment therefore "transcriptomics" refers to all RNA transcripts in a biological sample. These RNAs include non-coding RNA such as ribosomal RNA (rRNA), protein-coding messenger RNA (mRNA) and small RNAs [61]. All cells in an organism have got the same genome but being exposed to cell-specific conditions affect their gene expression levels. The main aim of transcriptomics is to quantify the changes in expression levels of genes under different conditions such as time, disease or mutations [62]. Comparison of conditions are performed with the help of statistics science. Finding significantly changed genes according to their expression levels, gives crucial insights about the given conditions and significantly changed genes, so various results can be revealed about the condition-specific gene expression [61, 62].

2.6.1. Microarrays

Microarrays are cost-effective, high-throughput tools that are widely used in numerous applications of biology and medicine. mRNA microarrays contain a large number of cDNA fragments which are called probes and mRNAs converted into cDNA molecules can hybridize to these probes designed to be complementary to 3' ends of the transcripts. Light emitted by the hybridized probe sets indicate the expression levels of genes [63]. The aim of this method is to examine the expression profile of genes in the target tissues. There are different types of array technologies such as one-color and two-colour. Especially two-colour array methods are generally used for comparison of the test

and control samples. RNA samples are tagged with different colours of dyes, this facilitates comparison of gene expression profiles of different conditions [64]. General workflow of microarray technique can be seen in Figure 2.8 [65].

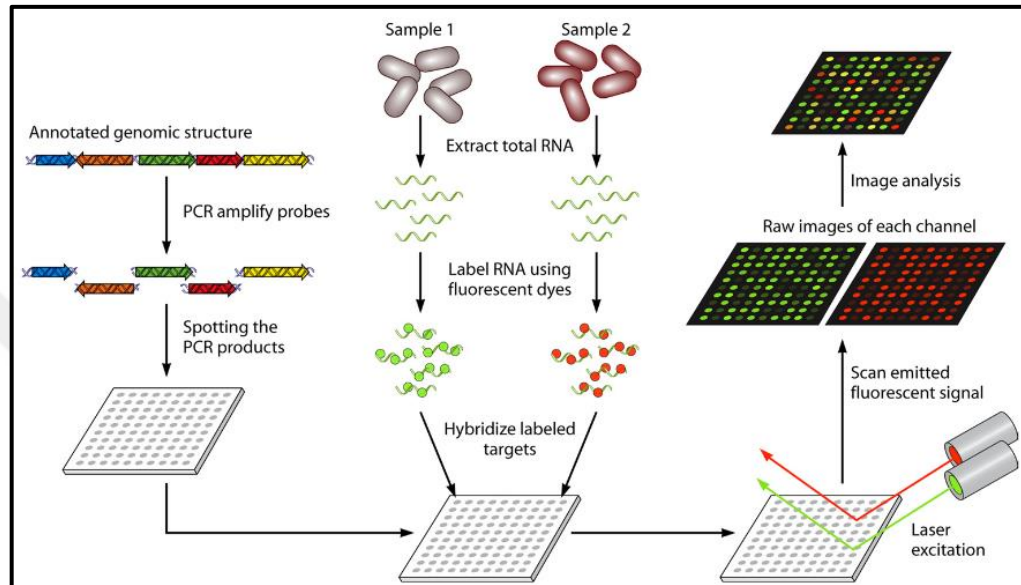


Figure 2.8: Workflow of microarray.

Nowadays, microarray technology is not preferred when compared with RNA-Seq since output of this hybridization-based technology is relative but not absolute. In spite of this downside, microarray technology is still widely used because it is cost-efficient and has established data analysis methods [63].

2.6.2. RNA-Sequencing

There are different types of high-throughput transcriptome sequencing methods like whole transcriptome shotgun sequencing (also known as RNA-Seq). Many tedious protocols such as Sanger sequencing-based and hybridization-based microarrays approaches are replaced by these new technological developments [66, 67].

Workflow of the classical RNA-Seq can be summarized as follows: RNA molecules are isolated from the biological samples, target type of RNA are enriched in the sample, converted to their complementary DNA (cDNA) and sequencing library is prepared

(Figure 2.9.) [66]. To perform a successful experiment, quality and amount of RNA needs to be sufficient to generate the sequencing library. After RNA isolation and poly-A tail selection, the fragmentation of RNA is necessary to achieve a certain size range because many sequencer devices such as Illumina have size limitations on the fragments they can sequence. After fragmentation, RNA-Seq library preparation is performed. In this step, RNA is reverse-transcribed (RT) to cDNA because almost all sequencer devices require DNA libraries. Specific adaptor sequences are added to cDNA molecule via ligation or RT/polymerase chain reaction (PCR) [62, 66].

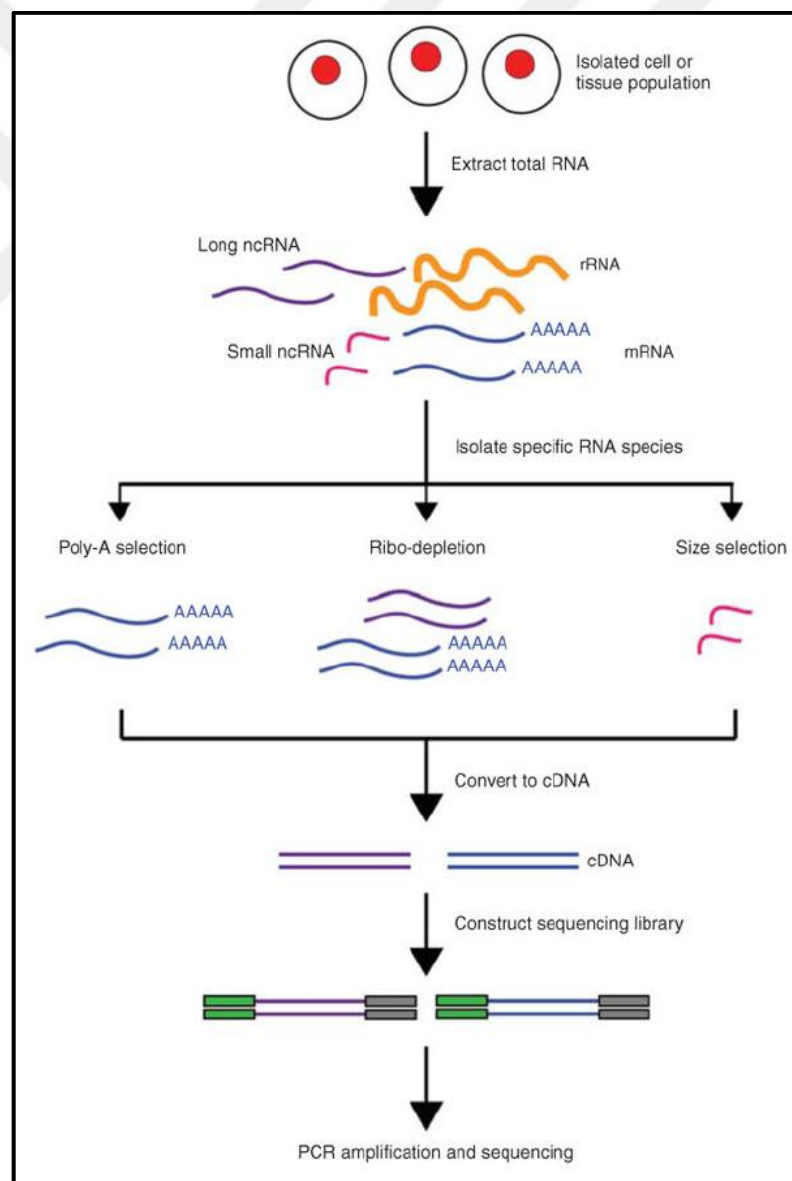


Figure 2.9: Overview of the RNA-Seq workflow.

The final step is called the "sequencing step". There are different types of groups of sequencers according to their sequencing capacity. As mentioned above, size of final fragment is important for sequencing quality (generally less than 500 base-pair). In addition, sequencing type can be single-end (SE) or paired-end (PE) reads. SE reads are sufficient for examination of expression levels under a certain condition for well-known organisms, whereas PE reads are more useful for discovering not well-defined transcriptomes [68]. FASTQ- file formats are produced by the sequencer. These files contain the nucleotides read from each fragment and quality scores of the reads. Quality control of raw data is crucial for RNA-Seq data analysis because the raw data contains adaptors, duplicated reads, sequencing errors and low-quality reads. There are many software tools developed to detect and remove these unwanted reads. For example, FastQC and Trimmomatic, respectively perform quality control of raw data and removal of low-quality reads and adaptors from raw data [68, 69].

Mapping of reads to a reference genome is the essential step of the RNA-Seq data analysis. There are different types of tools and software for this step, such as HISAT2, Bowtie2 or STAR. Aligned reads are used to calculate the counts and find locations of reads thus estimation of expression levels can be performed according to read counts. Tools such as HTSeq-count or FeatureCounts are useful in converting raw data into counts. Finally, counts matrix is produced and differential gene expression analysis can be performed by statistical hypothesis testing [69].

2.6.3. Networks in Biological Systems

Many different biological events can be modelled and interpreted via networks in biological systems. For instance, atomic interactions in a protein can be given as an example [70] or changes in expression levels of a group of genes between two specific condition can be given as an another example. To visualize the relationships of all elements sequentially and to interpret the relations between cases/phenomena, use of biological networks is one of the efficient ways. The meaning of a network is “interconnected system of things”. In biological sciences, to investigate biological events

in wide perspective, researchers need to reconstruct biological system networks. Networks include different components such as nodes and edges.

Signaling pathways are one of the important fields in biology and they can also be given as an example of biological system networks. In this case each node represents gene/transcription factors and the edges between nodes may represents the expression level of these genes and also edges can take different features such as colour or shape or both of them to express types of relationships between nodes.



3. MATERIAL AND METHODS

3.1. Datasets

In this study, to answer our biological questions we analysed time-course microarray, RNA-Seq and Chip-Seq datasets from cell differentiation/reprogramming experiments (Table 3.1)

Table 3.1: Datasets used in this thesis.

GSE ID	Organism	Process	Direction	Dataset Source	Dataset Type
GSE117204	Mouse	Differentiation	iPSC→Fibroblast	Thakurela et al 2019	RNA-Seq
GSE117203	Mouse	Differentiation	iPSC→Fibroblast	Thakurela et al 2019	ChIP-Seq
GSE137001	Mouse	Reprogramming	Fibroblast→iPSC	Velychko et al 2019	RNA-Seq
GSE67462	Mouse	Reprogramming	Fibroblast→iPSC	Chen et al 2016	Microarray
GSE67520	Mouse	Reprogramming	Fibroblast→iPSC	Chen et al 2016	ChIP-Seq
GSE111831	Human	Differentiation	iPSC→Astrocyte	Tekin et al 2018	RNA-Seq
GSE50206	Human	Reprogramming	Astrocyte→iPSC	Takahashi et al 2018	Microarray

3.1.1. Datasets 1-2 / GSE117204 – GSE117203

In the study of Thakurela et al., stem cells from mice are differentiated into fibroblasts, the aim of the study was to compare the reprogramming processes from different starting points of cell states. Samples from different time points were collected to examine the differences between time intervals, such as 0, 24, 36, 48, 60, 72, 96 hours and transcriptome data was produced using RNA-Seq. In the second part of the project, cells are reprogrammed, starting from the intermediate differentiation time points as mentioned above. Thus, researchers can compare the reprogramming processes based on the initial state of the process. In addition, researchers also collected ChIP-Seq data from 0, 48, 96 hours [71]. In this thesis, both RNA-Seq and ChIP-Seq datasets were re-analysed.

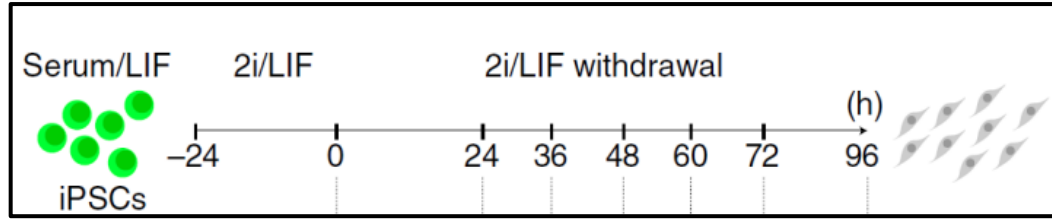


Figure 3.1: iPSC to MEF differentiation experiment design by Thakurela et al.

3.1.2. Dataset 3 / GSE137001

The aim of the study produced this dataset was to investigate importance of the OSKM factors, especially the existence and necessity of Oct4 in reprogramming processes. To examine Oct4 necessity, Velychko et al., reprogrammed fibroblast cells into stem cells in different conditions such as induction of SKM and OSKM expression which excludes Oct4 induction, OSKM induction was used as the control group. Samples from different time points such as 0, 2, 4, 6, 9, 12 days are collected to examine transcriptomics of datasets [72]. In this thesis, only control group dataset was used for examining the reprogramming process.

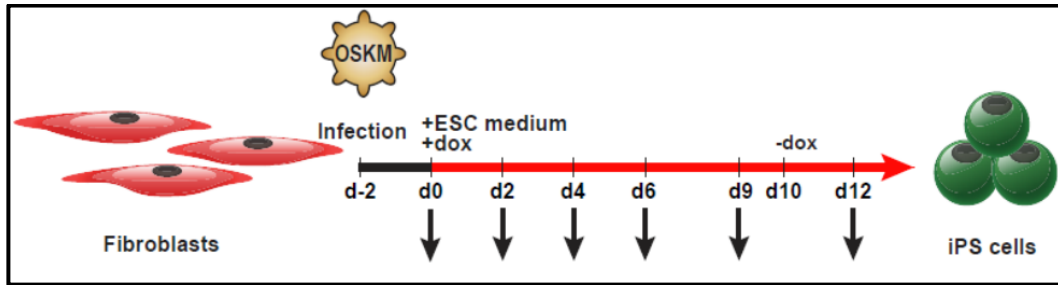


Figure 3.2: MEF to iPSC reprogramming experiment design by Velychko et al.

3.1.3. Datasets 4-5 / GSE67462 – GSE67520

The main aim of the study which produced these datasets was to examine the DNA binding profile of Oct4 transcription factor during the reprogramming. In this study, Chen et al., collected samples from mouse derived fibroblast cells during reprogramming in different time points such as 0, 1, 3, 5, 7, 9, 11, 15 to apply ChIP-Seq and microarray

analysis [73]. In this thesis, both datasets were re-analysed as reprogramming process datasets.

3.1.4. Dataset 6 / GSE111831

The purpose of the study which produced this dataset was to compare effects of different culture conditions on the differentiation of human embryonic stem cells into astrocytic cells. Tekin et al., collected samples from different time points such as 0, 2, 5, 15, 30, 67, 117 days to apply RNA-Seq analyses [74]. In this thesis, mentioned RNA-Seq dataset was used for investigation of differentiation process.

3.1.5. Dataset 7 / GSE50206

The purpose of the study which produced these datasets was to question the OSKM factors and suggest new transcription factors to enhance the efficiency of reprogramming processes. To reach their aim, Chen et al., reprogrammed fibroblast cells into stem cells and collected samples from different time points such as 0, 1, 3, 5, 7, 9, 11, 15. Days to apply microarray and ChIP-Seq analysis [75]. In the scope of this thesis, ChIP and microarray datasets were used for examination of differentiation process.

3.2. Pre-analysis inspection with PCA

Principle Component Analysis (PCA) is a data dimension reduction technique. In order to interpret big datasets, techniques are required to radically reduce their dimensionality in an illustratable way, while preserving most of the information in the data. PCA is one of the oldest and most widely used methods for this purpose [76]. The fundamental idea behind PCA is defining a lower dimensional coordinate systems to represent each data point by its projection on the new 2D or 3D coordinate system.

In this study, the PCA technique was applied to datasets for the initial inspection of sample profiles. As an example, PCA results from the Thakurela et al 2019 and Velychko

et al 2019 datasets can be seen in Figure 3.3. Thakurela dataset represents differentiation processes and when the PCA chart is examined, it can be seen that the correct order of the time points is represented by their scores on PC1 and the replicates for each time point are close to each other. The Velychko et al dataset was also examined similarly. The MEF time point appears to be a single point because only one sample raw data for this time point is available. Assuming that the MEF sample, which was the initial sample of the experiment, was less heterogeneous than the following time points, this data point was used as the initial point of both replicates. It is possible to tell by looking at the PCA chart that the 12th day and the initial sample are indeed different from each other, but that the other time points are very close to each other, which means that gene expression changed drastically at the beginning of the experiment and also at the end of the experiment. This simple inspection gives us a perspective on the general structure of the data.

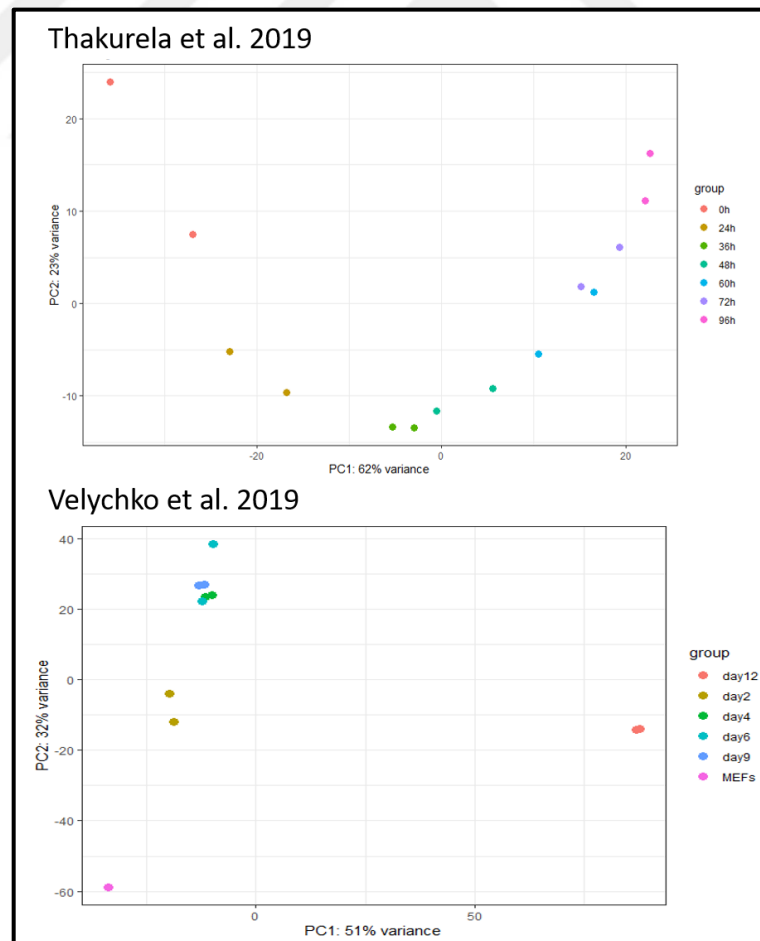


Figure 3.3: PCA charts of the Velychko et al 2019 and Thakurela et al 2019 datasets.

3.3. Software Tools for Transcriptomics Data Analysis

3.3.1. RNA-Seq Data Analysis

RNA-Seq data analysis can be split into two different parts. The first one is pre-processing of the raw data, the second one is extracting biological meaning from the count data using statistics. In this thesis study, the output of the pre-processing pipeline is used to identify differences between reprogramming and differentiation process via comparing the expression level of genes.

3.3.1.1. RNA-Seq Data Pre-processing

RNA-Seq data analysis is started from FASTQ files as the input of the pre-processing step. FASTQ files were downloaded from EBI SRA. The name of the file obtained at the end of the pre-process step is the count matrix. Count matrix includes Ensembl Stable IDs and counts of genes in each condition.

There are several steps in the analysis which follow each other. Generally, output of the one software/tool is input of the other one, therefore the collection of these steps constructs a “pipeline”. RNA-Seq pipeline includes quality control, trimming, mapping, sorting, and quantification steps. For each step there are different open sources tools/software. Each software has its strengths and weaknesses. In this thesis, software that have many citations and popular ones are used.

3.3.1.1.1. FastQC

At the beginning of the analysis, it is crucial to make sure that the FASTQ file quality is above the threshold levels for a number of criteria. Researchers face the several technical problems in analysis of high-throughput datasets. For example, one of the problems is adapters, sequences are often contaminated with adapters ligated to their 5’ or 3’ ends. The other problem is wide variation in base read quality within and between

reads [77]. These problems can affect the analysis results therefore the quality control step can be considered as the most important part of the pre-processing steps.

FastQC is a software that can report quality of the FASTQ files. It is developed by Babraham Bioinformatics [78]. FastQC tool provides command based and also GUI based applications for Linux users. It is one of the most popular tools because of availability of its open-source codes and ease of use. It produces a report for each FASTQ file in HTML format. This report includes different types of statistical information about reads. An example of FastQC reports can be seen in Figure 3.4.

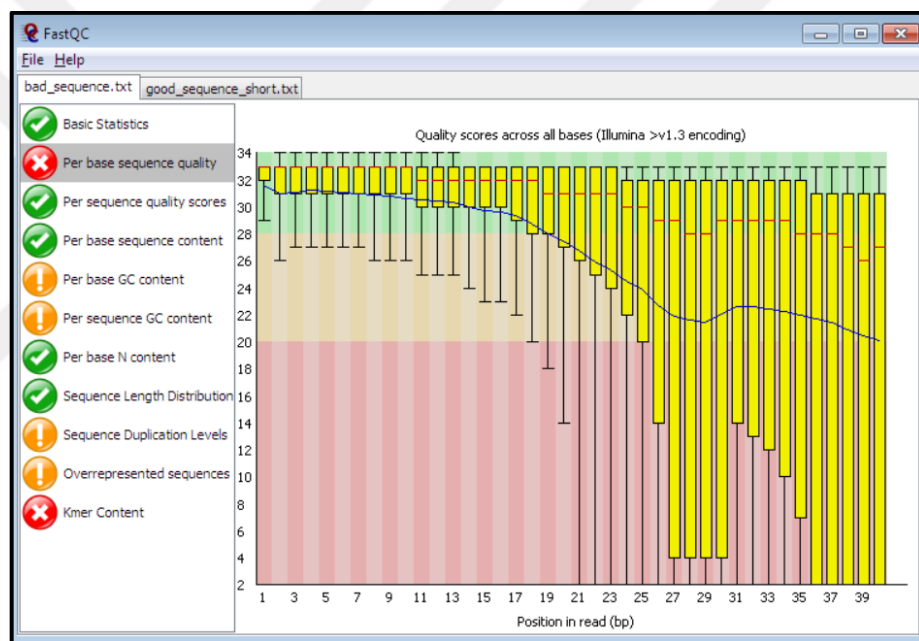


Figure 3.4: A quality report generated by FastQC.

3.3.1.1.2. Trim Galore

As mentioned before, reads need to be filtered based on their quality and need to be trimmed to remove the adapters from the reads. To overcome this problem after the quality control checking, the reads are trimmed from the FASTQ files with software such as Trim Galore [79].

Trim Galore is a tool developed by Felix Krueger at the Babraham Institute. Trim Galore works with another tool, Cutadapt [80]. Cutadapt finds the adapters and removes

their sequences from the reads. If there are not any adaptor sequences supplied by the user, Trim Galore can detect the adapters by analysing 1 million sequences of the first part of the file and find specific adapters that have 12-13 base-pairs. Users can also do manual adapter detection using parameters such as `--illumina`, `--nextera`.

Trim Galore is a Linux based software and it has different pre-defined parameters for trimming. For example, user needs to indicate whether the data is a paired-end or a single-end. In single-end sequencing, fragments are read only one end to the other but in paired-end sequencing, fragments are sequenced from both ends, thus high-quality sequencing datasets are generated [81]. Reliability and accuracy of the paired-end technique is higher than the single-end but paired-end sequencing is time consuming and more expensive than the single-end sequencing. While trimming FASTQ file, ‘sequencing technique’ parameter should be used to define sequencing methodology to trimming software.

3.3.1.1.3. HISAT2

Trimmed FASTQ files generated by Trim Galore, are then mapped to a reference genome. In this step, mapping algorithm matches the sequences that are found both in the FASTQ file and the reference genome. As a result of this matching, BAM (Binary Alignment Map) files are generated by HISAT2 [82]. HISAT2 is a software that is funded by National Human Genome Research Institute.

BAM files contain two sections, one of them is header and the other one is alignment section. Header section contains information about the file, such as name, length, alignment method. Alignment section contains the technical information such as read name, read sequence, quality of each base in the read. Start coordinates and chromosome information of the read on the reference genome are also included at the alignment section.

HISAT2 maps the reference genome and FASTQ files from the samples, hence this Linux-based software requires the reference genome as an input. HISAT2 tools can generate another file format, SAM (Sequence Alignment MAP). If the user wants to generate BAM file, which is a binary file and have smaller size than SAM, should add

another defining parameter about output of this software. In this thesis, SAM files were generated by using HISAT2 software.

3.3.1.1.4. Samtools

After the mapping step, SAM files were converted to BAM files and sorted according to ascending order of chromosomal locations of the reads. Samtools can not only process SAM files, it can also process BAM and CRAM file formats [83]. Samtools software includes more than ten commands but, in this thesis, only -view and -sort commands were used, sorting step helps the quantification to be faster. In addition, it can split, merge and index files.

3.3.1.1.5. FeatureCounts

Quantification step is the last pre-processing step for RNA-Sequencing. In this thesis study, FeatureCounts software was used for quantification of different samples that came from different conditions. The output of this software is generally a tab separated file that includes Gene ID, chromosome, start, end, strand, length and counts in the samples.

Transcriptome studies generally are done to reveal the differences or similarities between samples that originate from different conditions. To detect these differences or similarities, researchers use statistical tests. Here, the raw data was generated by featureCounts tool to detect significantly changed genes using statistical tests. The files generated by featureCounts are counts “read counts” or “count matrix” because the file includes genes and their expression levels as integers. At the downstream analysis, these integers are used as input for statistical tests [84].

3.3.1.2. RNA-Seq Data Downstream Analysis

After pre-processing step, count matrix is obtained to extract biological meaning with statistical analysis from the raw count data. To apply statistical analysis to data, data may be classified, split, grouped according to aim of the study.

3.3.1.2.1. DESeq2

After the count matrix has been generated in the pre-processing step, differential expression analysis is performed. DESeq2 [85] is an R package that has many user-friendly commands for statistical analysis of count matrix. While using DESeq2 package, one of the most important things is the statistical test design. DESeqDataSetFromMatrix command is used to give the count matrix as an input to the algorithm, and statistical test design is one of the parameters of that command. In this part of the analysis, user should specify the number of replicates and the experimental conditions. LRT (Likelihood Ratio Test) [86] was used in this thesis to detect significantly changed genes with respect to time. LRT test is performed comparing the fit of one model to the fit of the other, by removing one independent variable from the model, in the datasets analysed here, the only independent variable was the time of sampling. Hence, LRT test was used to compare the time intervals within themselves, which can be considered as Wald testing performed between each time interval. For instance, if there were three time points that are 0, 48, 96 hours, LRT test was used to detect significantly changed genes between 0-48 and 48-96 hours.

3.3.1.2.2. K-means clustering

Clustering algorithms are used to group the data according to their profiles. There are many algorithms that cluster the datasets with different mechanisms. In this thesis, K-means clustering was used.

K-means algorithm is one of the most popular unsupervised machine learning clustering algorithms. In K-means clustering algorithms, the number of groups is

represented by K , a hyper parameter that needs to be specified by the user. In this method, algorithm finds the similarities between samples or variables that included in the data. In this case, K-means algorithm finds the similar genes and groups them., J-express software was used for applying k-means clustering to datasets. J-express [87] is a java-based tool that has functionalities such as clustering the datasets, apply PCA (principal component analysis) or draw heatmaps.

3.3.1.2.3. Cytoscape

Cytoscape is a Java based software that can visualize molecular interaction networks and pathways. In this study, correlations between genes are visualized by using Cytoscape software. A screenshot from Cytoscape version 3.8.2 GUI is shown in Figure 3.5 [88]

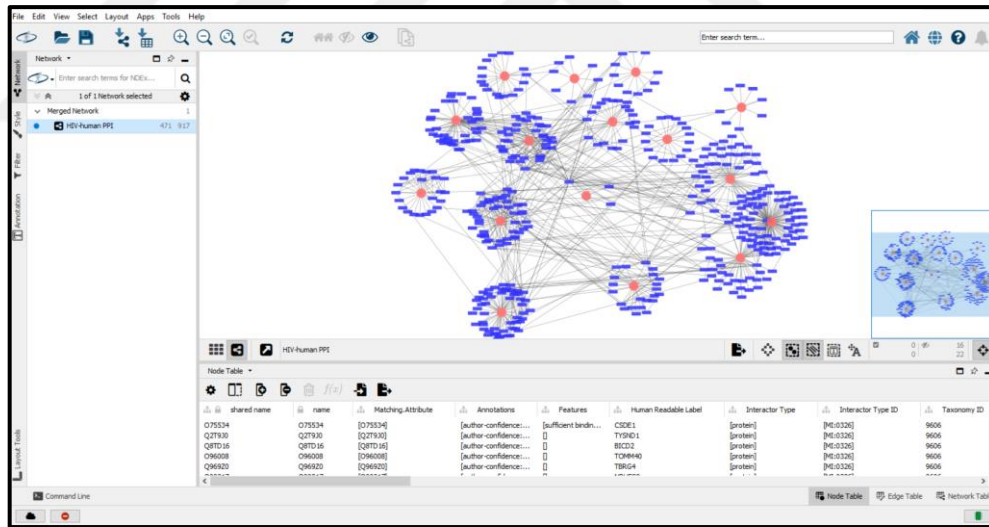


Figure 3.5: Cytoscape interface.

3.3.1.2.4. Dynamic Regulatory Events Miner (DREM)

The Dynamic Regulatory Events Miner (DREM) [89] is a software that is used for analysis and visualization of transcriptional gene regulation dynamics. DREM software takes time series gene expression datasets and transcription factor-gene interaction network as input, it produces a dynamic regulatory network as an output. In this study,

time series datasets were analysed with DREM to find transcription factors that regulate the gene expression. To generate TF-gene interaction list, many databases were used such as, TRRUST [90], Harmonizome [91], RegNetwork [92]. The user interface of DREM2 can be seen in Figure 3.6.

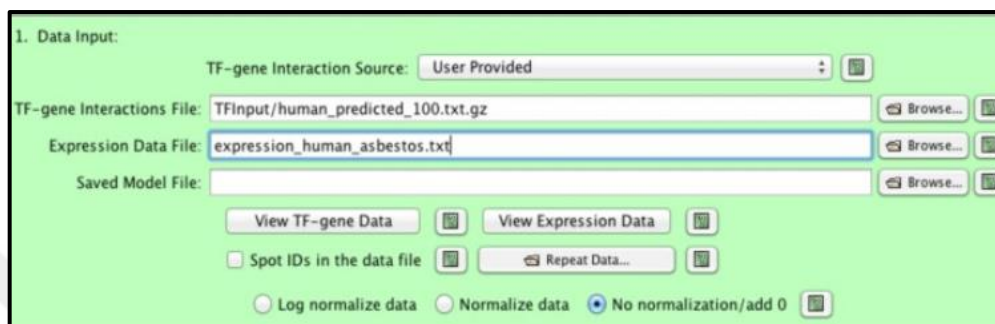


Figure 3.6: DREM2 interface.

3.3.1.2.5. KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database that provides information about genomes and their relationships to biological systems. There are different kind of KEGG databases such as KEGG GENES, KEGG LIGAND, KEGG PATHWAY. In this thesis, KEGG PATHWAY is used for matching the significantly changed genes to their signaling pathways.

KEGG PATHWAY database contains manually curated diagrams of genes and compounds for different signaling pathways such as Notch and WNT. The list of signaling pathways can be found at KEGG PATHWAY website. Genes involved in signaling pathways are shown in boxes on the diagram, genes on the map can be highlighted by the user interactively. Input genes are marked with different colours when they match with genes in the relevant signaling pathway.

3.3.2. Microarray Data Analysis

Microarray is hybridization-based technique as previously described at Background Aspects section. Light emitted by the hybridized probe sets indicate the expression level

of genes[63]. First, hybridized probes need to be normalized after the experimental step, to eliminate technical variation in the arrays. In this way, a normalized matrix is obtained that includes genes and their expression levels. Then the statistical tests are used for obtaining biological meaning from the data.

3.3.2.1. Limma

Limma is a R package that can perform differential expression analysis of data from microarray experiments. Time-course data is data generated by extracting RNA from the sample at several different time points. Limma provides p-value for each gene, and genes with most significant differential expression can be detected according to p-value and/or fold change thresholds [93]. In this thesis, Limma package is used to obtain significantly change genes only from microarray data. Rest of the analyses were performed using instruments that analyse RNA-Seq data.

4. RESULTS

This section first summarizes the general analysis of each data set and presents the results. Following subsections describe integration of these results and discusses their biological interpretation.

4.1. iPSC to MEF Differentiation in Mice

As mentioned in Material and Methods section, to investigate iPSC to MEF process, Thakurela et al. data (GSE117204) was used. After passing through our raw data analysis workflow, 1762 significantly changed genes (adjusted p-value < 0.01, all p-values presented in this thesis were corrected using the Benjamini-Hochberg method), obtained with the DESeq2 tool, were first clustered by using k-means method (Figure 4.1) then GO Biological Process (GO BP) enrichment analysis was performed on genes that showed either a continuous increase or continuous decrease trend (Figure 4.2). Then selected significantly changed genes were analysed to find transcription factors that regulate them by using DREM2 tool (Figure 4.3).

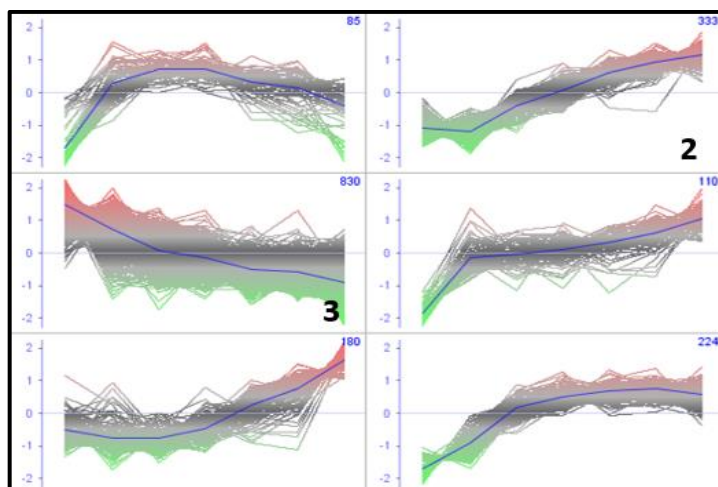


Figure 4.1: Clustering result of Thakurela et al dataset. On the x-axis time is ordered 0, 24, 36, 48, 60, 72, 96 hours, on the y-axis is the level of gene expression, blue curves show the cluster mean. The gene numbers of the clusters are located in the upper right corner of each cluster. The numbers of the clusters analysed by GO are shown above the graph.

In analysis of this dataset and the other datasets given in the following sections, two clusters representative of the gene expression dynamics were chosen. The increase or decrease is not linear in most cases, but faster at the beginning or end of the process. Still, a monotonic trend is seen in most of them. Among the clusters which show monotonic (continuous) increase or decrease in expression levels, the ones that contain markers of the starting or end point cell type were chosen, assuming that the genes in these clusters were driving the differentiation or reprogramming processes, under the regulation of few master regulators (TFs), which potentially have a similar expression pattern themselves. In addition, GO biological process term enrichment results were considered when selecting clusters with increasing and decreasing gene expression profiles.

For this data set, where iPS cells were differentiated into MEF cells, iPS markers are expected to be in one of the decreasing clusters, while markers of MEF are expected to be in one of the increasing clusters. According to the clustering result shown in Figure 4.1, cluster-2 was chosen to represent genes with increased expression, and cluster-3 to represent genes with decreased expression level. Then as mentioned above, GO enrichment analysis was performed on these two clusters. Cluster-2 GO BP result can be seen in Figure 4.2.A and cluster-3 GO BP result can be seen in Figure 4.2.B.

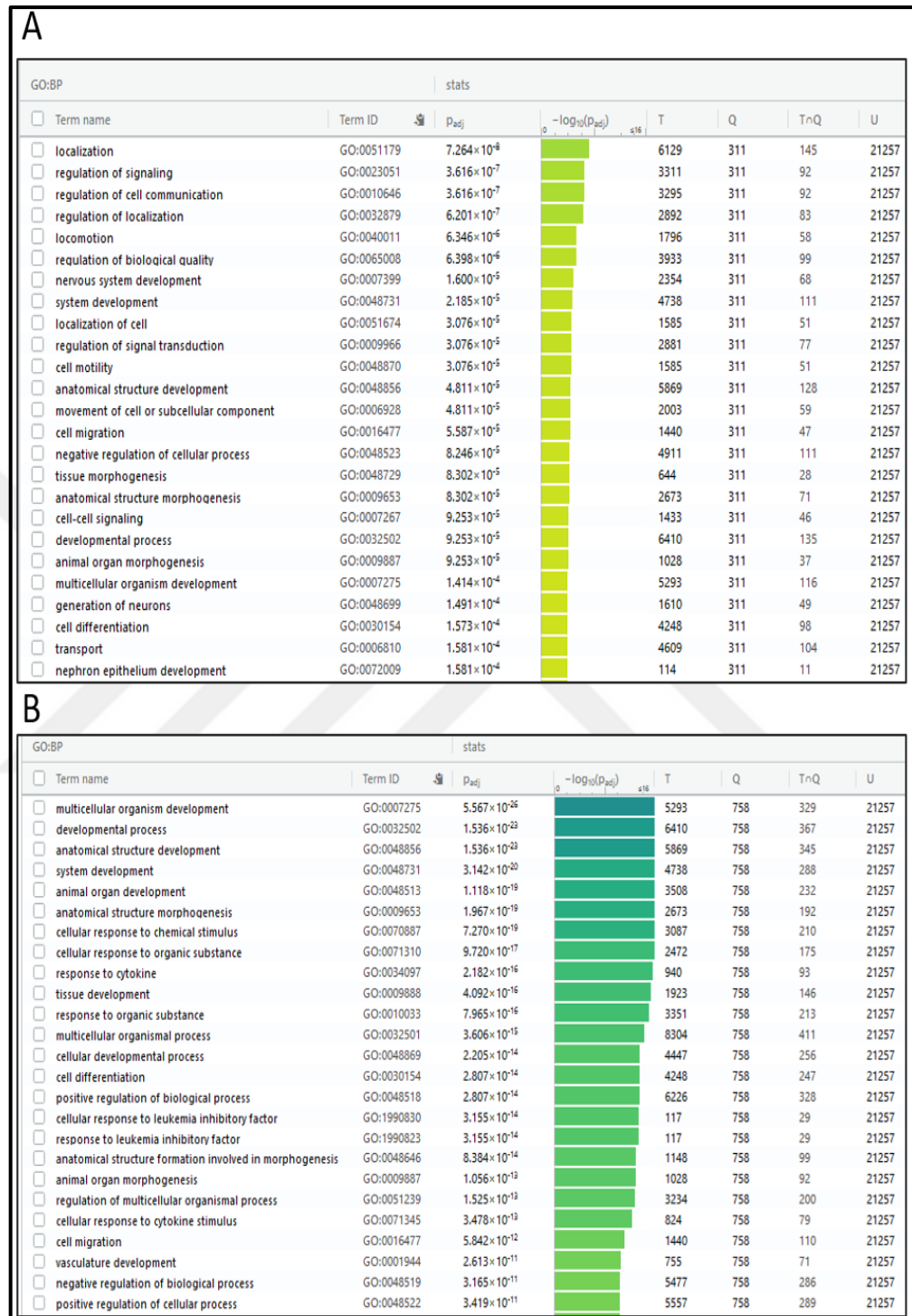


Figure 4.2: GO BP analysis results of selected gene clusters of iPSC to MEF process. While A represents genes that have increased expression level, B represents genes that have decreased expression level.

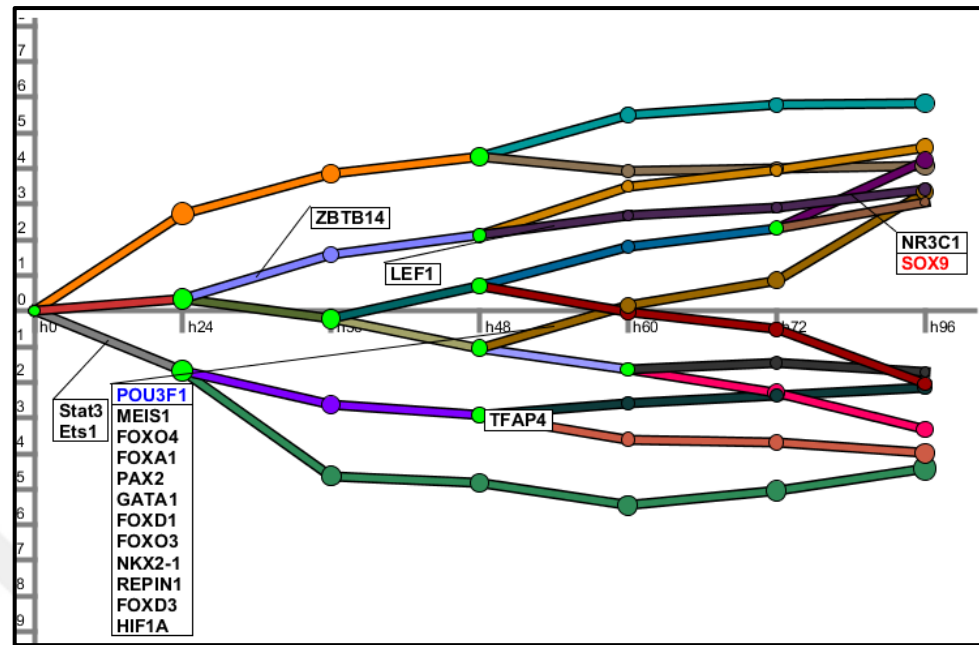


Figure 4.3: The result of co-regulation analysis of significantly changed genes of Thakurela et al dataset (iPSC -> MEF).

Since the observed GO biological terms are very general terms, these terms only provide an overview of genes with increased or decreased expression levels, but do not provide detailed information. In order to understand the activity of pluripotency signals during the differentiation of cells, the “pluripotency pathway” in the KEGG pathway can be examined. To survey the relationship of significantly changing genes with each other in this context, the pairwise correlations of the genes in this pathway were calculated. After constructing a network of all the interacting genes on the map, the correlations between the interacting genes were mapped on the pathway. The correlations between “pluripotency pathway” genes can be seen in Figure 4.4. The correlation figure will be discussed in the following sections.

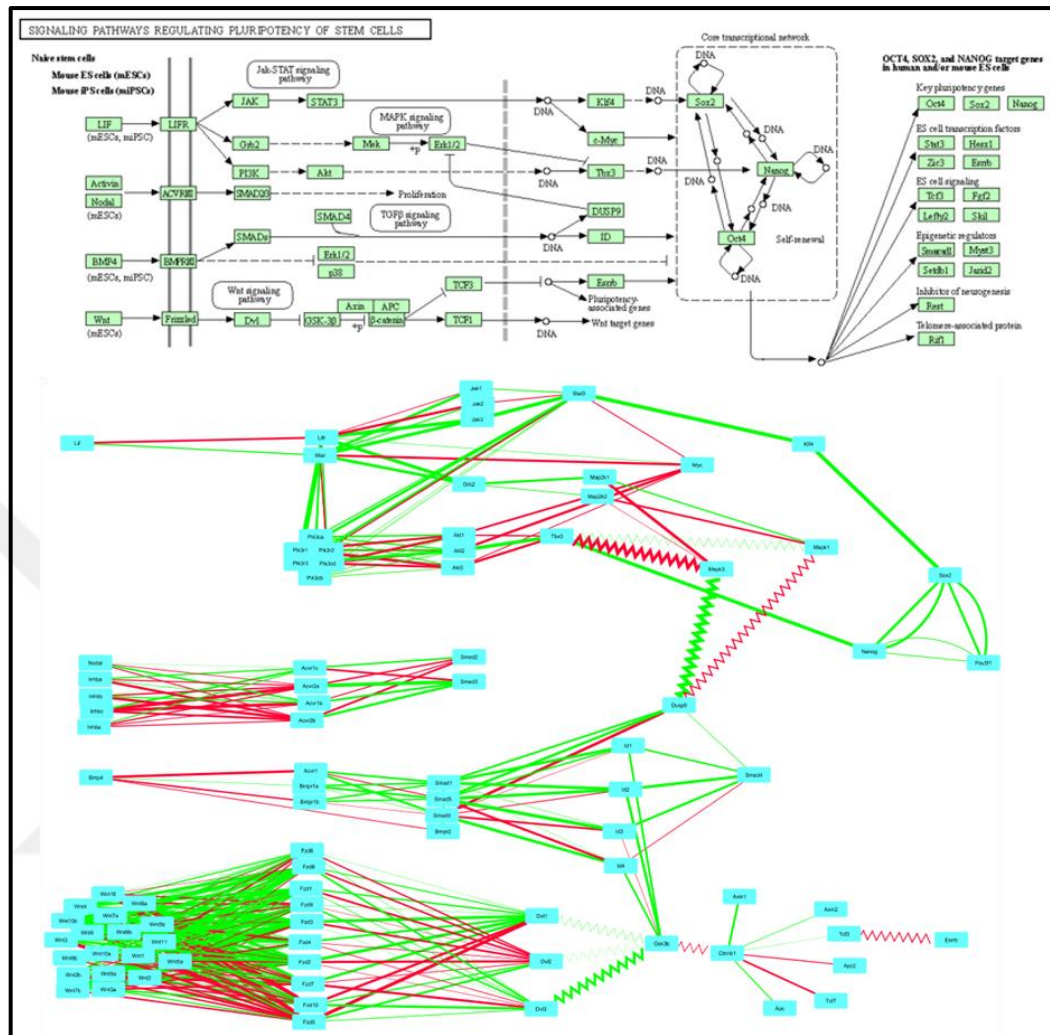


Figure 4.4: Genes that correlate in the iPSC to MEF process and interact in Pluripotency signaling pathways. Red lines indicate negative correlations, green lines indicate positive correlations, and the thickness of the line indicates the strongness of the correlation. The zigzag lines show interactions known in the literature as suppression.

4.2. MEF to iPSC Reprogramming in Mice

An experiment to investigate reprogramming of MEF cells into iPSC cells, the process in the opposite direction to the process described in the above section, was performed by Velychko et al (GSE137001). In order to initiate the reprogramming process, OSKM transcription factors were induced as mentioned in the Material and Methods section. OSKM factors are encoded by Oct4, Sox2, Klf4 and Myc genes.

Same workflow was performed in this dataset and 5428 significantly changed genes were found. Significantly changed genes were clustered by using the k-means method (Figure 4.5), then GO Biological Process (GO BP) enrichment analysis was performed on genes that showed either a continuous increase or continuous decrease trend (Figure 4.6). Then selected significantly changed genes were analysed to find transcription factors that regulated them by using DREM2 tool (Figure 4.7).

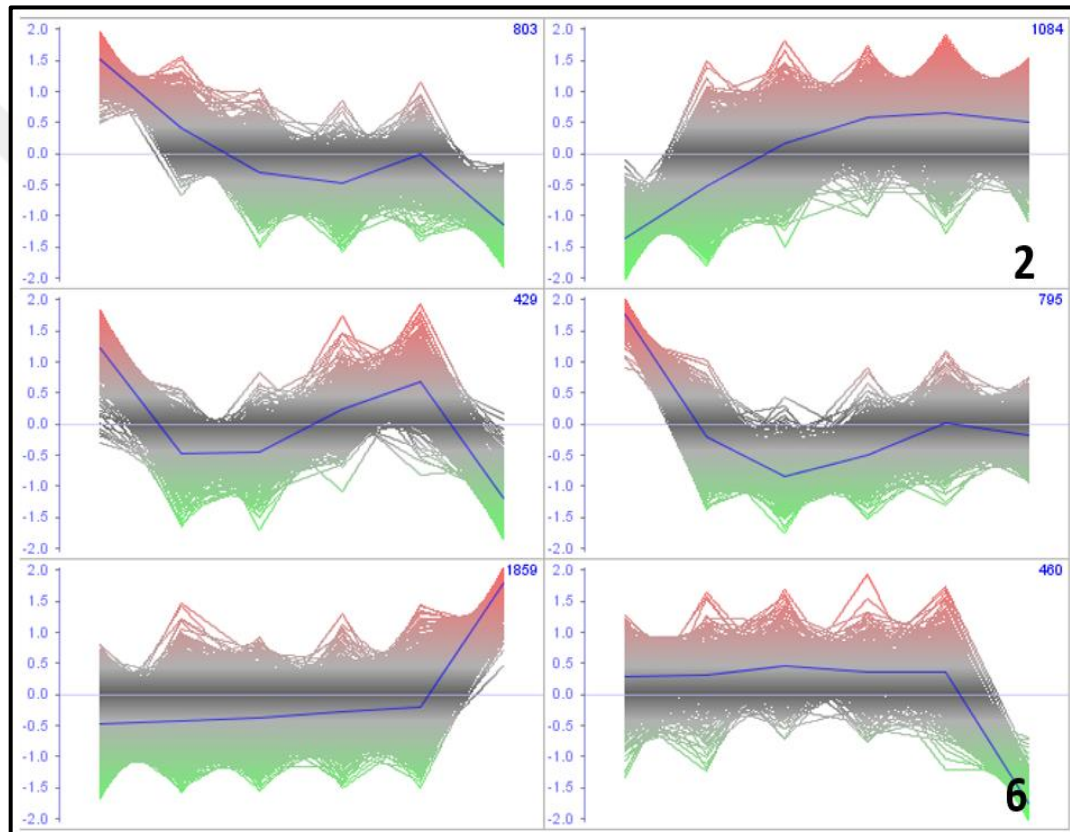


Figure 4.5: Clustering result of the Velychko et al dataset. On the x-axis time is ordered 0, 2, 4, 6, 9, 12. day, on the y-axis is the level of gene expression, blue curves show the cluster mean. The gene numbers of the clusters are located in the upper right corner of each cluster. The numbers of the clusters analysed by GO are shown above the graph.

According to the clustering results, Cluster 2 and 6 were chosen for further analysis. While cluster-2 represents genes that have increased expression profile, cluster-6 represents genes that have decreased expression profile. Then, as mentioned above, GO analysis was performed on these two clusters. Cluster-2 GO BP result can be seen in Figure 4.6.A and cluster-6 GO BP result can be seen at Figure 4.6.B.

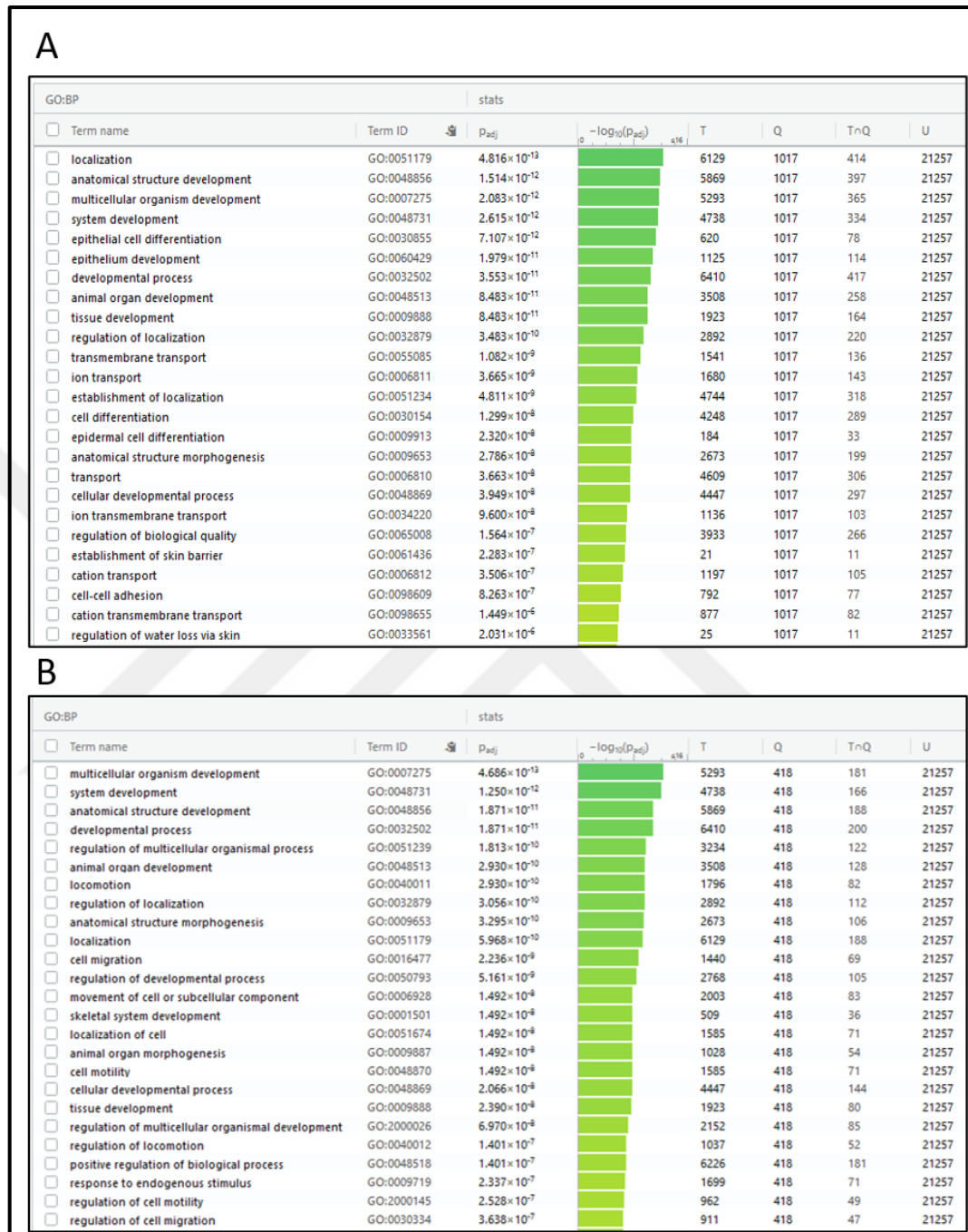


Figure 4.6: GO BP analysis results of selected gene clusters of MEF to iPSC process. While A represents genes that have increased expression level, B represents genes that have decreased expression level.

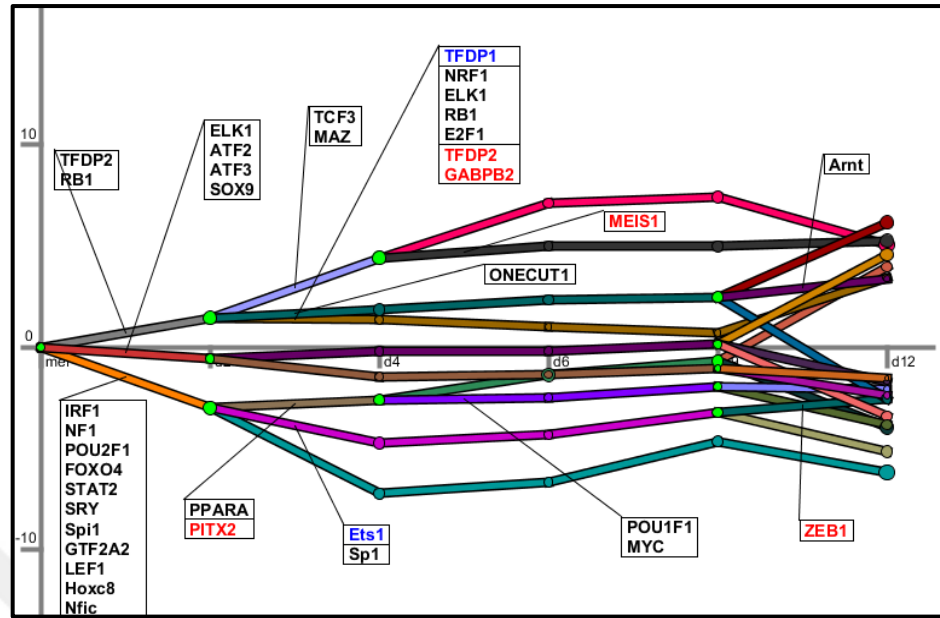


Figure 4.7: The result of co-regulatory analysis of significantly changed genes of Velychko et al dataset.

4.3. iPSC to Astrocyte Differentiation in Human

In order to compare the reprogramming and differentiation processes of induced pluripotent stem cells in a second biological system, the differentiation of induced pluripotent stem cells (iPSC) into human astrocyte cells was examined and time-dependent activation maps of signaling pathways were obtained. As mentioned in the Materials and Methods section, to investigate the iPSC to astrocyte differentiation process, Tekin et al. data (GSE111831) was used.

Same workflow was performed in this dataset and 3794 significantly changed genes were found. Then, significantly change genes were clustered by using k-means method, then GO Biological Process (GO BP) enrichment analysis was performed on genes that showed either a continuous increase or continuous decrease trend. Then selected significantly changed genes were analysed to find the transcription factors that regulated them by using DREM2 tool. Clustering results can be seen in Figure 4.8.

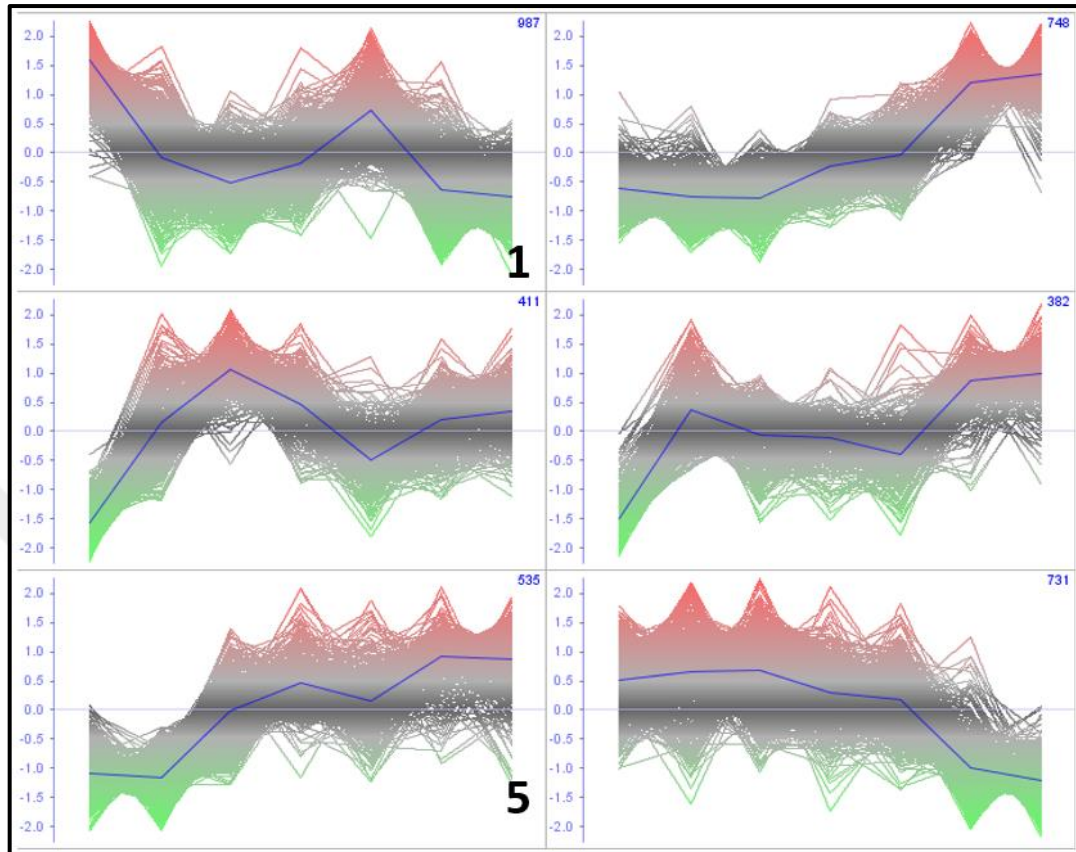


Figure 4.8: Clustering result of Tekin et al dataset. On the x-axis time is ordered 0, 2, 5, 15, 30, 67, 114. day, on the y-axis is the level of gene expression, blue curves show the cluster mean. The gene numbers of the clusters are located in the upper right corner of each cluster. The numbers of the clusters analysed by GO are shown above the graph.

Based on Figure 4.8, cluster-1 was chosen to represent genes with decreased expression, and cluster-5 to represent genes with increased expression level. Then as mentioned above, GO analysis was performed on these two clusters. Cluster-5 GO BP result can be seen in Figure 4.9.A and cluster-1 GO BP result can be seen in Figure 4.9.B.



Figure 4.9: GO BP analysis results of selected gene clusters of the iPSC to astrocyte process. While A represents genes that have increased expression level, B represents genes that have decreased expression level.

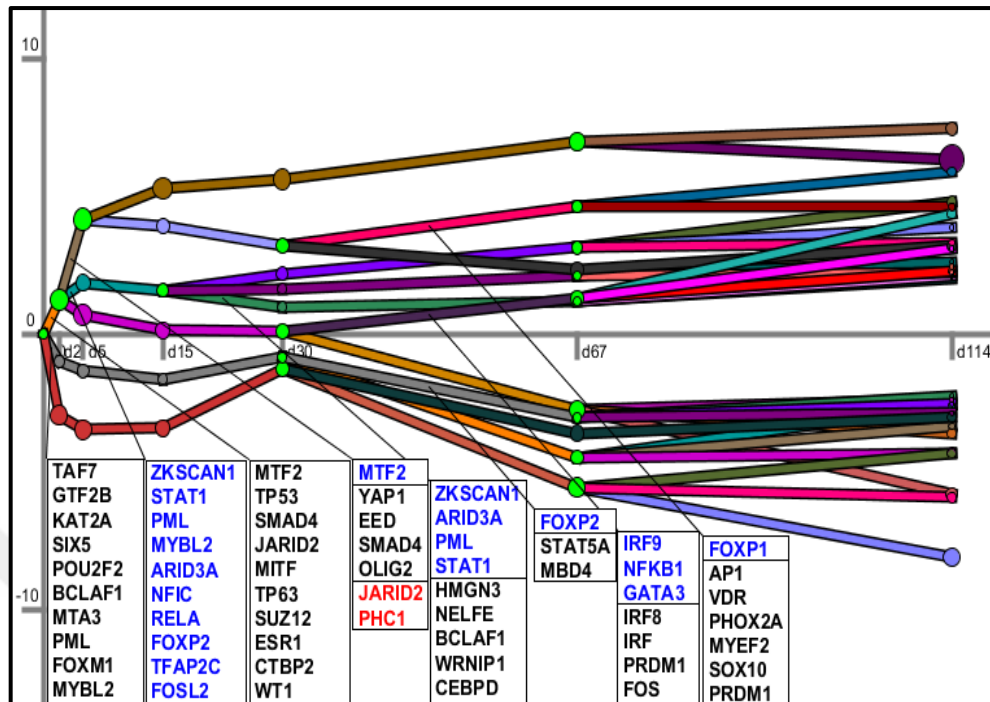


Figure 4.10: The result of co-regulatory analysis of significantly changed genes of Tekin et al dataset.

When the GO enrichment analysis results of the cluster-5, which has an increasing profile in terms of the expression levels, are examined, it is seen that the terms related to morphogenesis are at the top of the list with very low adjusted p-values. These results indicate that the cells are going through major morphologic changes in iPSC to astrocyte differentiation.

iPS cells differentiating into astrocyte cells are expected to downregulate their growth-related genes and also pluripotency or Leukemia Inhibitory Factor (LIF) pathway related genes. Looking at the GO enrichment results of cluster-1, which is a cluster of genes with decreasing profiles (Figure 4.9.B), the terms related to translation and metabolism in general appear, but interestingly, no term associated with the LIF is seen. However, this cluster was chosen because it contains more terms that can be associated with cell division than other clusters, and pluripotency genes which are major targets of LIF pathway (NANOG, POU5F1) are included in this cluster.

4.4. Astrocyte to iPSC Reprogramming in Human

To investigate the reversibility of the astrocyte-iPSC differentiation/reprogramming process, the Takahashi et al. (GSE50206) dataset was analysed as the reverse process of the Tekin et al. dataset. Detailed information about Takahashi et al. dataset can be found in the Material and Methods section.

This dataset is produced using microarrays, hence, to obtain differentially expressed genes; Limma R package was used. According to 0.01 adjusted p-value threshold, 9225 significantly changed genes were detected. The further steps in the analysis are the same as the methodology used in the analysis of other datasets. Clustering result can be seen in Figure 4.11.

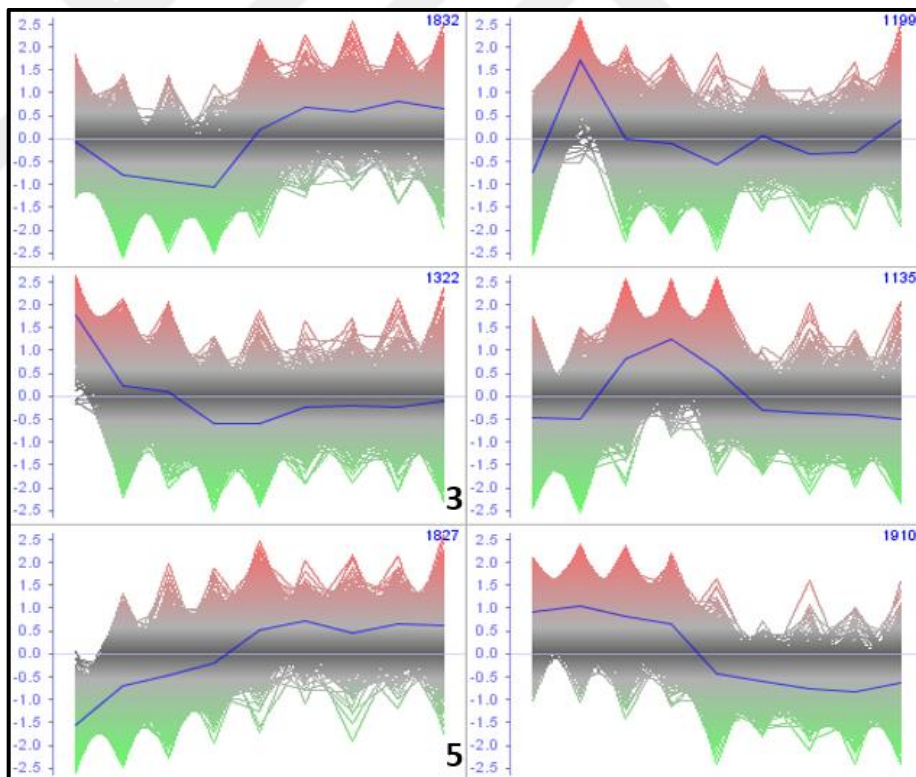


Figure 4.11: Clustering result Takahashi et al dataset. On the x-axis time is ordered 0, 3, 7, 11, 15, 20, 28, 35, 42 days, on the y-axis is the level of gene expression, blue curves show the cluster mean. The gene numbers of the clusters are located in the upper right corner of each cluster. The numbers of the clusters analysed by GO are shown above the graph.

Out of the clustering results, cluster-3 was chosen to represent genes with decreased expression, and cluster-5 to represent genes with increased expression levels. Then as mentioned above, GO analysis was performed on these two clusters. Cluster-5 GO BP result can be seen in Figure 4.12.A and cluster-3 GO BP result can be seen in Figure 4.12.B.

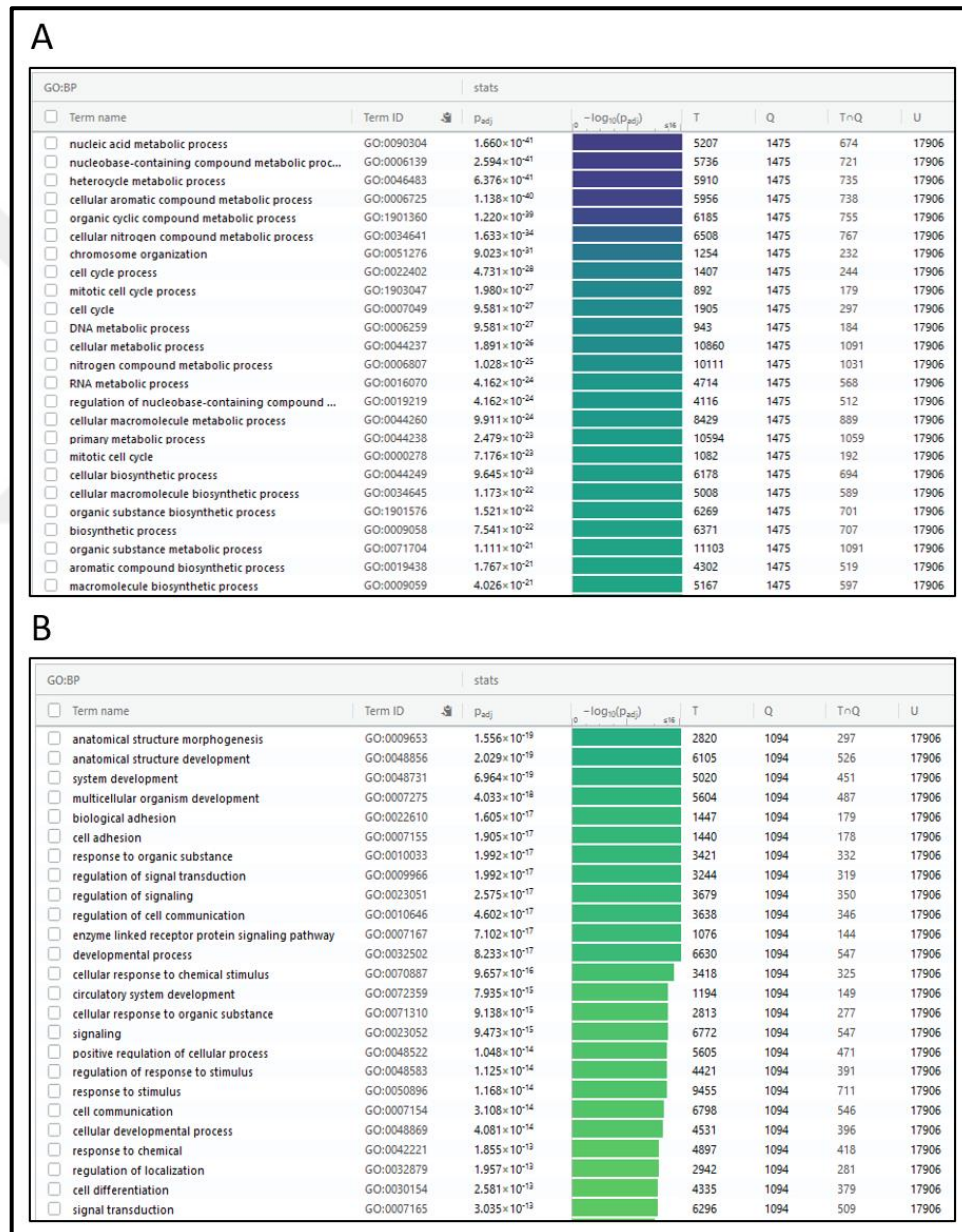


Figure 4.12: GO BP analysis results of selected gene clusters of astrocytes to iPSC process. While A represents genes that have increased expression level, B represents genes that have decreased expression level.

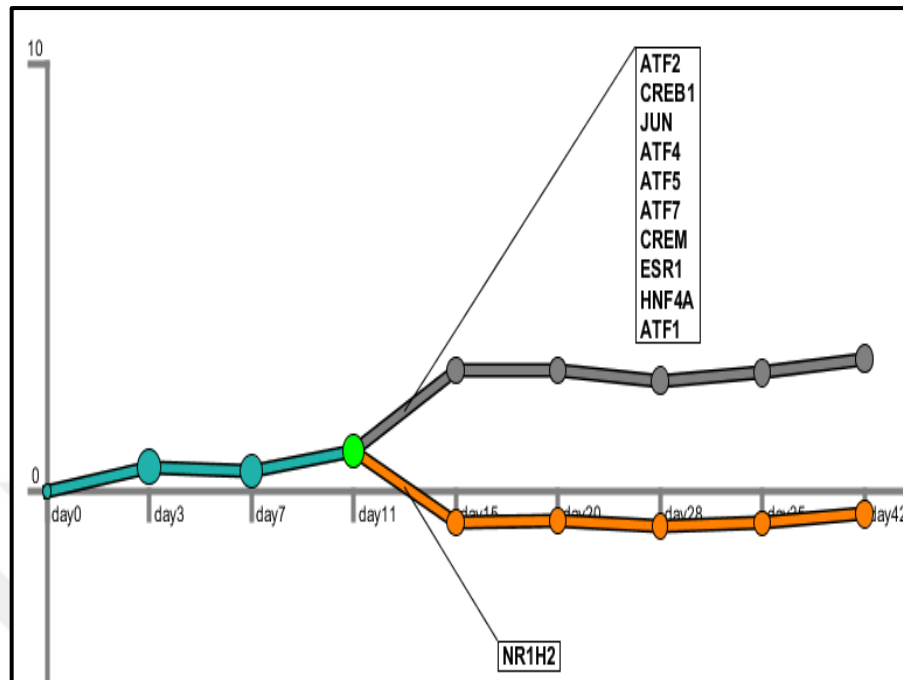


Figure 4.13: The result of co-regulatory analysis of significantly changed genes of the Takahashi et al dataset.

4.5. Reversibility of Reprogramming and Differentiation Processes in Pluripotency Signaling Pathway

At the first step of these analyses, significantly changed gene and their co-regulation were analysed. Then these genes were clustered according to their expression profiles. The genes in the selected clusters were mapped into the signaling pathways that play important role in the pluripotency processes. Signaling pathways regulating pluripotency of stem cell map is provided by KEGG PATHWAY, this pathway allows a global visualization of signaling pathways related to pluripotency, differentiation and reprogramming processes. Hence, in this thesis, pluripotency signaling map was used as the key pathway to understand the reversibility of examined biological processes (reprogramming and differentiation). This step is important for understanding the reversibility of differentiation and reprogramming biological process because, according to the highlighted genes/component of the signaling pathways, new approaches can be improved related with reversibility and efficiency of these processes

4.5.1. Reversibility of Pluripotency Between Reprogramming and Differentiation Process in Mice

KEGG pathway presents the pluripotency pathway for human and mice on the same map. Hence, to interpret the pathway reversibility in mice, one should focus on the top panel of the map. Pluripotency pathway reversibility mapping with increasing gene profiles in mice can be seen in Figure 4.14. Decreasing expression profiles and their roles on signaling pathway can be seen in Figure 4.15.

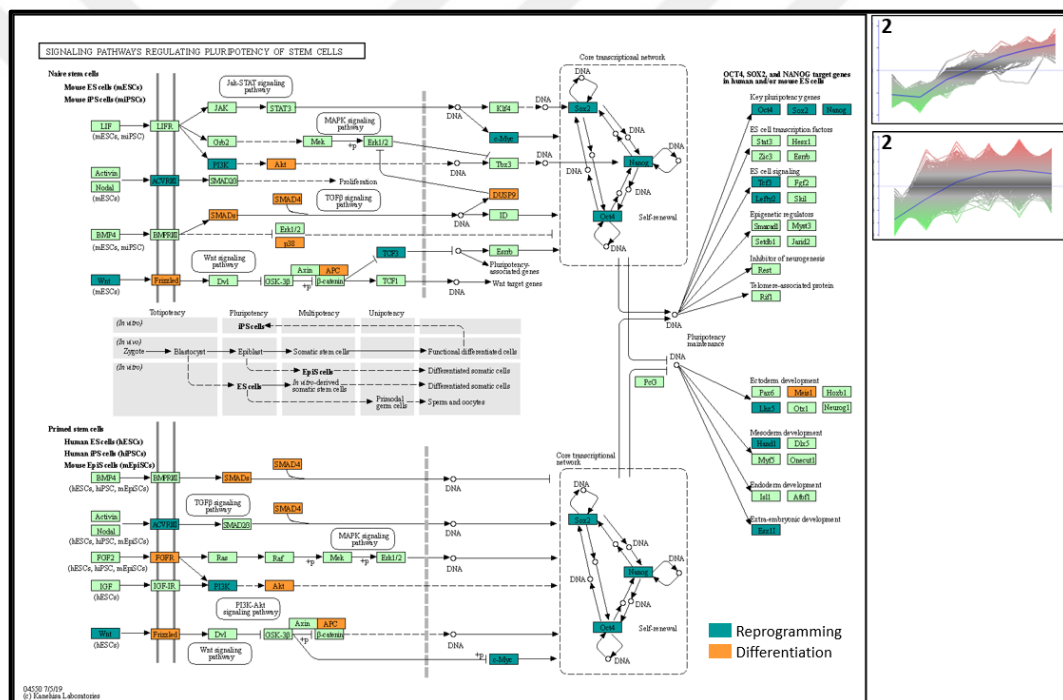


Figure 4.14: Pluripotency and related signaling pathways were mapped with significantly changed genes with increased expression level in mice. While Velychko et al. 2019 (GSE137001) dataset was used to represent the reprogramming process, Thakurela et al. 2019 (GSE117204) dataset was used to represent the differentiation process.

Figure 4.14 explain the gene activity of these two opposite processes because both of these clusters show increasing expression level profiles also when examine the core transcriptional network, only reprogramming process genes are highlighted.

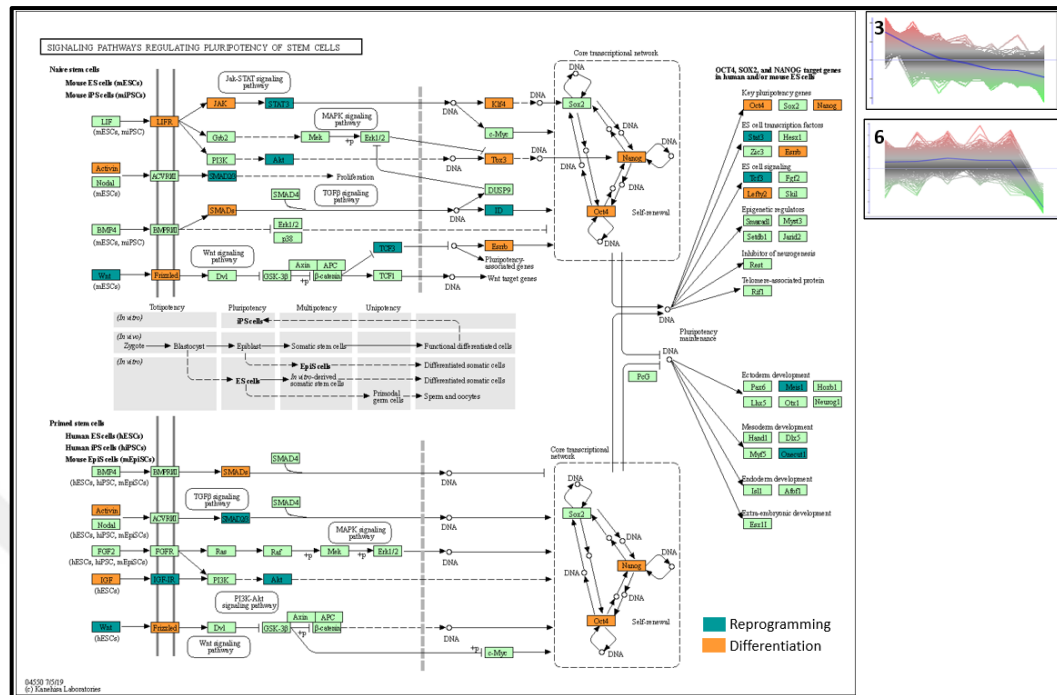


Figure 4.15: Pluripotency and related signaling pathways were mapped with significantly changed genes with decreased expression level in mice. While Velychko et al. 2019 (GSE137001) dataset was used to represent the reprogramming process, Thakurela et al. 2019 (GSE117204) dataset was used to represent the differentiation process.

Taken together, Figure 4.14-15 explain the reversibility of these two opposite processes, the genes increased in reprogramming (Fig 4.14, green boxes) are expected to decrease in differentiation (Fig 4.15, orange boxes), whereas the genes increased in differentiation in Fig. 4.14 (orange boxes) are expected to decrease in differentiation in Fig 4.15 (green boxes). Therefore, the boxes are expected to switch colors in Fig 4.14 and 14.5 in case of two processes being the reverse of each other in terms of gene expression. The expected switch in the colors is found only in some of the target genes of the pathway (Oct4, Nanog, Lefty), while genes that encode the signalling proteins mostly do not follow the reversibility profile. Sox2, one of the major targets of LIF pathway, was not highlighted in these charts as it was not in the selected clusters.

4.5.2. Reversibility of Pluripotency Between Reprogramming and Differentiation Process in Human

KEGG pathway presents the pluripotency pathways in two panels as human and mice. Hence, when examining the reversibility in humans, one should focus on the bottom panel of the figure. Pluripotency pathway reversibility mapping with increasing gene profiles in human can be seen in Figure 4.16 and decreasing gene profiles are given in Figure 4.17.

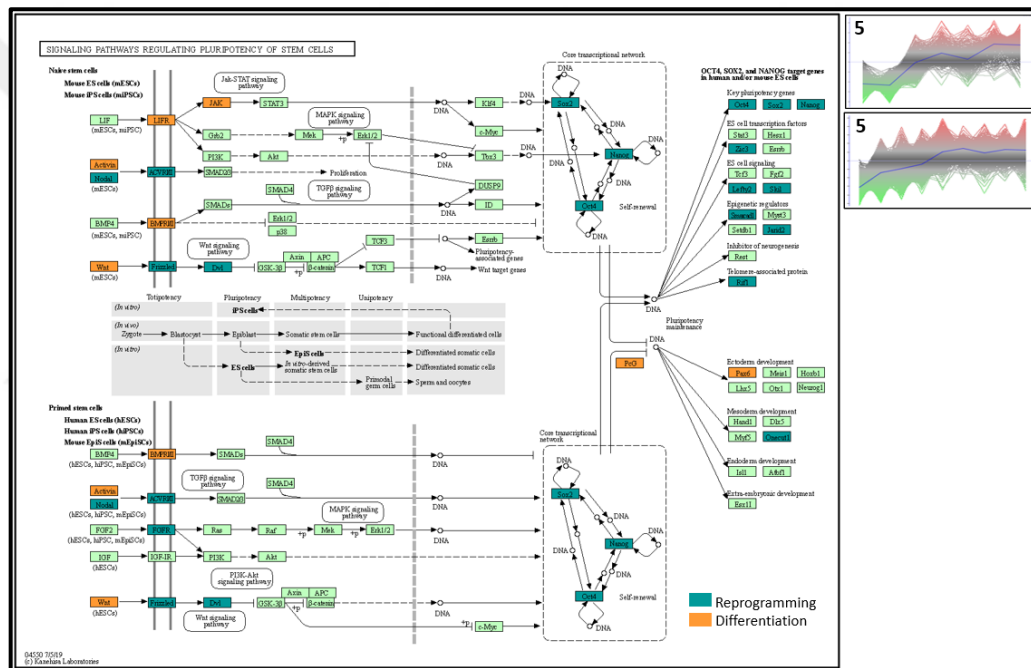


Figure 4.16: Pluripotency and related signaling pathways were mapped with significantly changed genes with increased expression level in human. While Takahashi et al. (GSE50206) dataset was used to represent the reprogramming process, Tekin et al. (GSE111831) dataset was used to represent the differentiation process.

Interpretation of Figure 4.16-17 is similar to Figs 4.14-15, a switch in color is expected for the genes that have reversible dynamics in two opposite processes between human iPSC and astrocytes. Many target genes have the colour switch, demonstrating their reversible behaviour. Few genes encoding receptors, ligands or signalling molecules also switched their colors, demonstrating that the human processes compared here may have a more robust reversibility, or astrocytes being more mature cells when compared to

MEF cells may have increased the magnitude of the changes in gene expression, leading to more significantly changed genes.

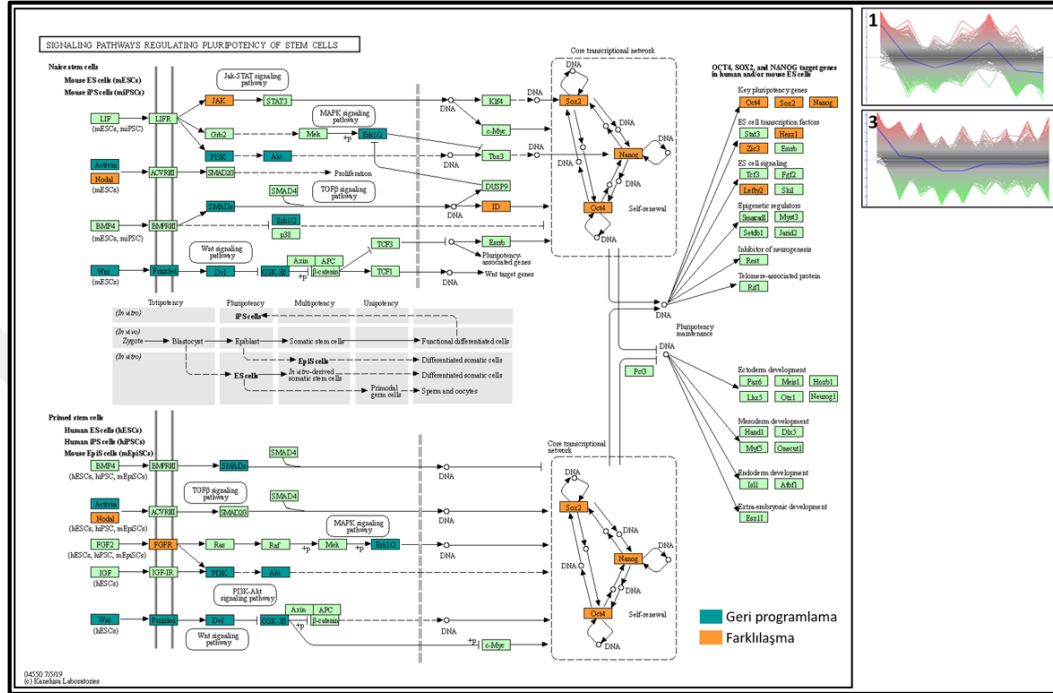


Figure 4.17: Pluripotency and related signaling pathways were mapped with significantly changed genes with decreased expression level in human. While Takahashi et al. (GSE50206) dataset was used to represent the reprogramming process, Tekin et al. (GSE111831) dataset was used to represent the differentiation process.

4.6. Reversibility of the Epithelial-Mesenchymal Transition in Differentiation and Reprogramming

As mentioned, in Background Aspects section, EMT and MET are important phenomena in differentiation and reprogramming of cells. Epithelial-mesenchymal (EMT) and mesenchymal-epithelial (MET) transitions are one of the main changes that occur in differentiation and reprogramming processes. These two processes are opposite of each other in terms of phenotypical changes in cells, and EMT refers to the transition from epithelium to mesenchymal and MET refers to the transition from mesenchymal to epithelium. Although the two processes are phenotypically inverse of each other, it is unclear whether they are inversely related also in gene regulation because the regulatory

networks governing the two processes appear to be different. For instance, while *Crb3*, *Epcam*, *Esrp1* genes are considered as MET genes, *Thy1*, *Snai1*, *Snai2*, *Zeb1* genes are considered as fibroblast genes. We expect that gene expression profile of *Thy1* and its group should have a decreasing profile during reprogramming because *Thy1* and its group play role in cell-cell communication and adhesion. As expected, *Epcam* gene and genes in its group have expression profiles that are different from the fibroblast gene group, but these differences are not distinct, hence, the pattern between these two groups is not clear. Gene expression profiles of the mice datasets are shown in Figure 4.18 and 4.19.

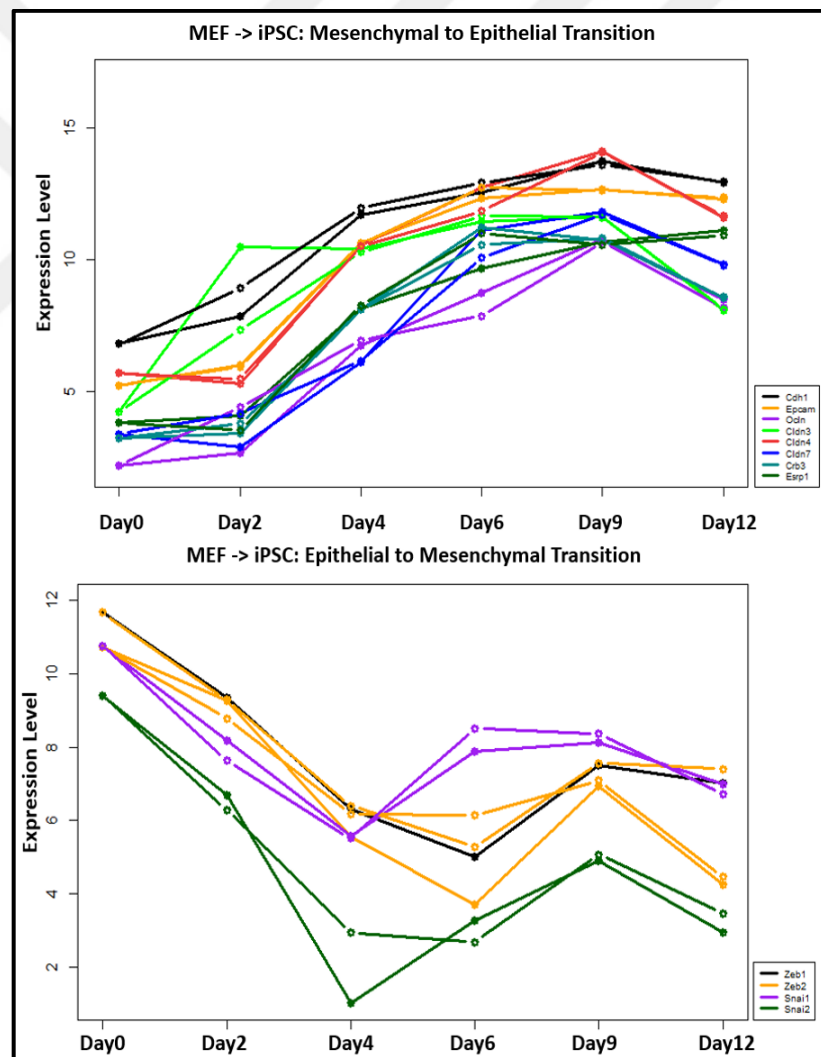


Figure 4.18: Expression profiles of MET and EMT genes in MEF – iPSC processes (replicate data were drawn separately thus there are two lines for each gene).

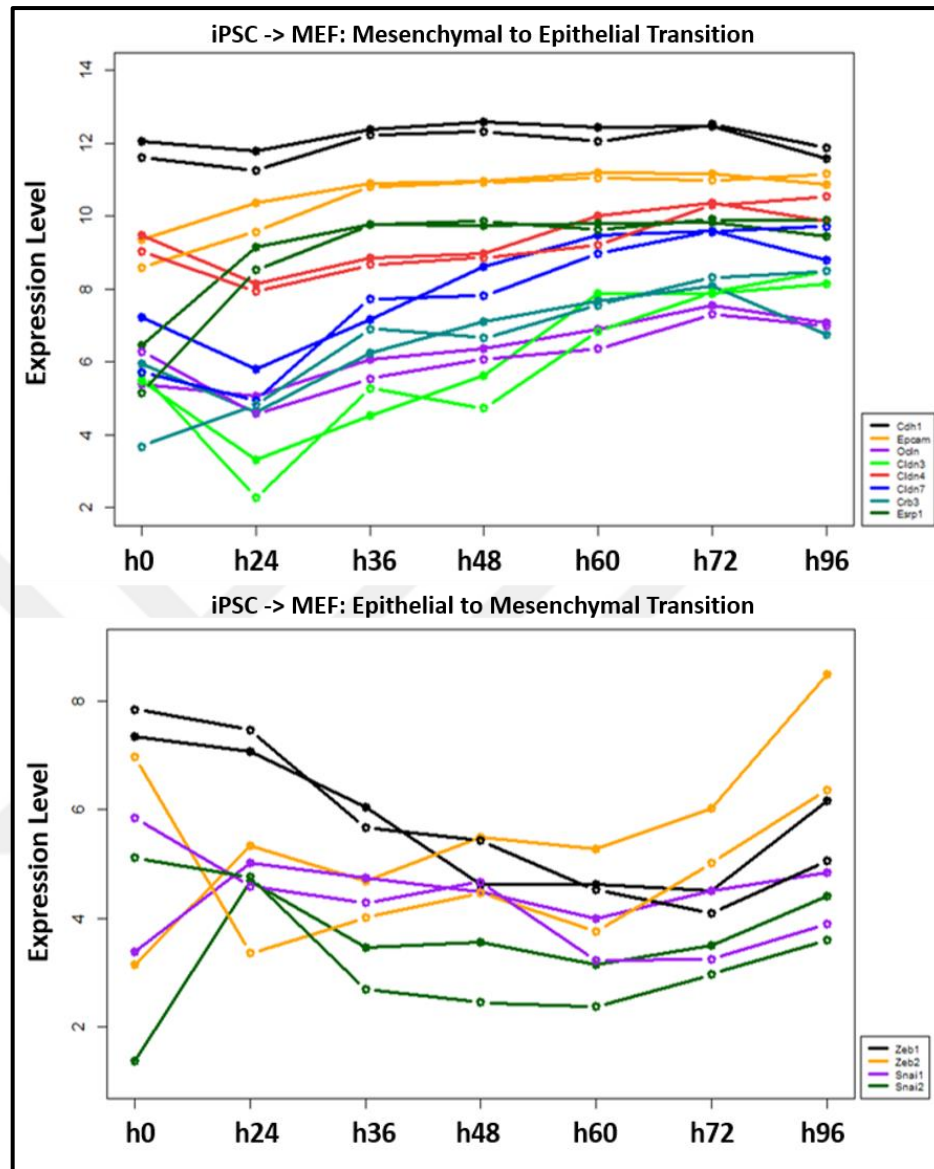


Figure 4.19: Expression profiles of MET and EMT genes in iPSC - MEF processes (replicate data were drawn separately thus there are two lines for each gene).

In the reprogramming processes, mesenchymal to epithelial gene expression is increasing at the beginning of the process (Figure 4.18), whereas, epithelial to mesenchymal gene expression level is affected at the beginning part of this process and their level is decreasing until day4. These genes become more stable after the 6th day. This shows that MET marker genes increase in the early stages during the reprogramming process in mice and remain stable afterwards, however, EMT marker genes, which is a process that progresses the opposite direction, decrease in the early stages and become

stable later on. This indicates that other processes preceding or proceeding EMT and MET do not take place in the same order.

Expression of the MET gene markers should be decreased during the differentiation process. In Figure 4.19, during differentiation from iPSC to fibroblast, interestingly there is almost no change in expression levels of MET genes except for some genes like *Cldn3*, *Espr1*. Expression levels of these genes are increasing between h60-h72. This situation can be interpreted as indication of noise in the data or effect of some other parameters such as epigenetic factors which we cannot directly observe in RNA-Seq data. In Figure 4.19, cells that differentiate from stem cells to fibroblasts are expected to have an increase in EMT marker genes, while some marker genes decrease in the first time period but then all EMT gene expression profiles show the increasing trend especially in the h60-h96 time period. These indicators are quite interesting because they show that EMT and MET during differentiation does not occur smoothly according to transcriptomics data.

The same processes were also investigated in human-sourced datasets and similar results are found. MET and EMT gene markers expression profiles are shown in Figure 4.20 and 4.21.

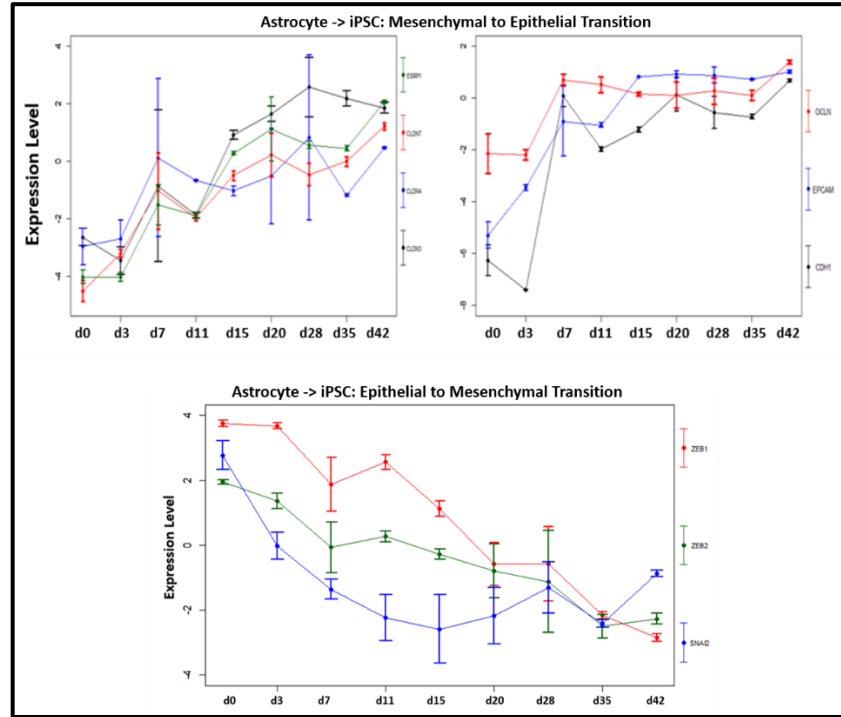


Figure 4.20: EMT and MET genes in the process of reprogramming from human astrocyte cells to stem cells (Takahashi et al. 2014) (Error bars show the standard deviation of 3 replicates).

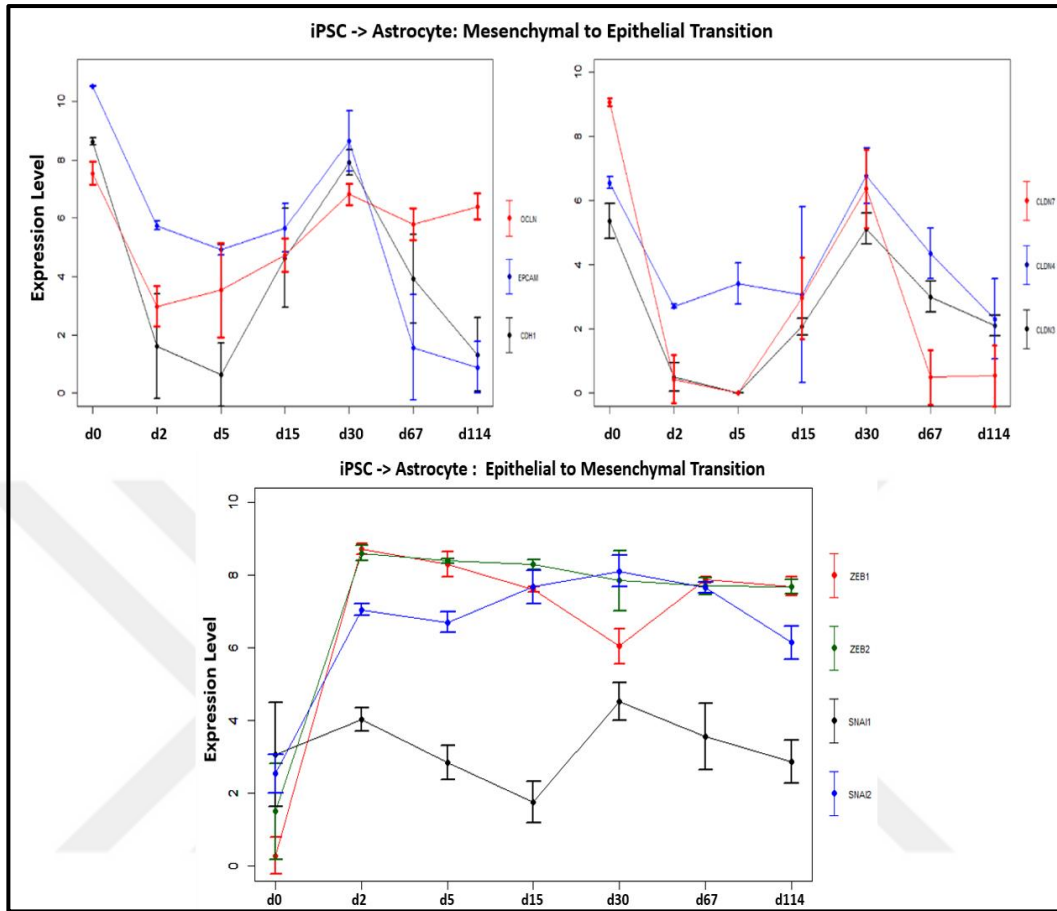


Figure 4.21: MET and EMT marker genes in the process of differentiation from stem cells to human astrocytes (Tekin et al. 2016) (Error bars show the standard deviation of 3 replicates).

As seen in the figures above, while expression of MET genes decrease in differentiation, EMT gene expression increases. Similarly, while expression of MET genes increasing in reprogramming, EMT gene expressions decreases.

To sum up the results of this subsection, it has been revealed that these processes are reversible with each other in terms of transcriptome, but another point to be emphasized here is that these processes operate differently in humans and mice in terms of timing within the process. In humans, EMT occurs in the early stages of differentiation, while in mice it occurs in the late stage. These findings will be discussed in detail in the Discussion section.

4.7. Comparison of the Signaling Pathways in Differentiation and Reprogramming Processes in Mice

Primarily the pluripotency signaling pathway was surveyed in this thesis but actually, differentiation and reprogramming processes are very complex phenomena, hence a wide range of aspects need to be considered when analysing these processes. For this reason, genes with increased expression level in the differentiation dataset in mice and genes with decreased expression level in the reprogramming dataset were mapped to WNT, TGF- β , and Notch KEGG signaling pathways. Based on all these results, it will be possible to obtain a general view of the concept of reversibility in gene expression, which will be examined in the next sections, all relevant signal pathways. WNT and TGF- β signaling pathways can be seen in Figure 4.22.

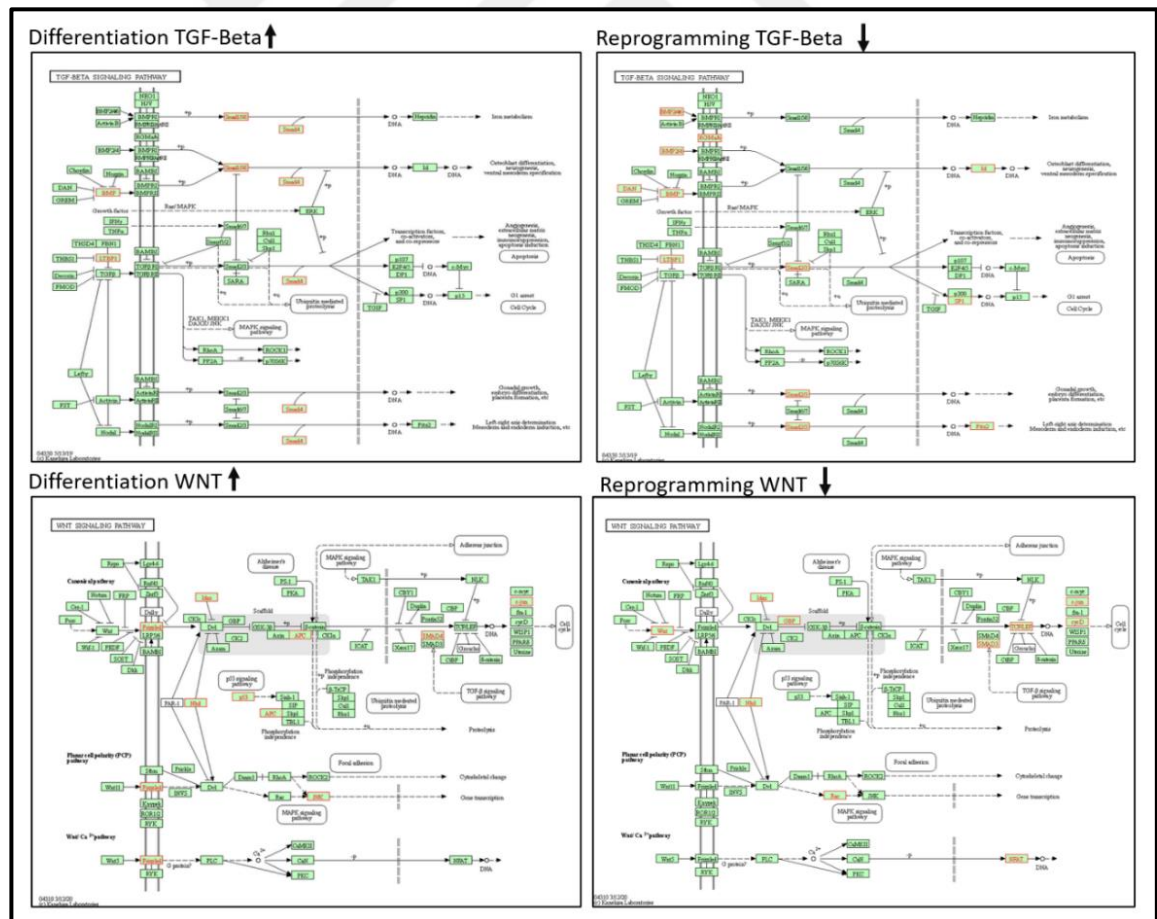


Figure 4.22: WNT and TGF- β KEGG signaling pathways.

According to Figure 4.22, genes with increased expression in differentiation show the activity of WNT and TGF- β signaling pathways, while genes with decreased expression levels in reprogramming show the activity of WNT and TGF- β signaling pathways. The same approach applies for the Notch signaling pathway, the Notch signaling pathway can be seen in Figure 4.23, where genes with increasing expression level in the differentiation dataset and decreasing expression level in the reprogramming dataset are mapped.

The maps of the processes in the opposite direction to Figures 22 and 23 are given in the attached Appendix section.

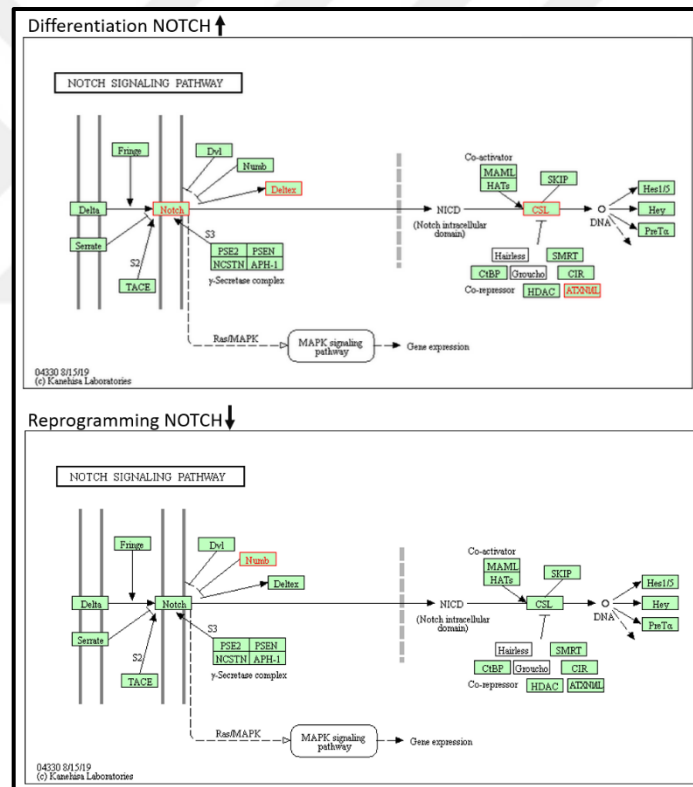


Figure 4.23: Notch signaling pathway.

Lack of genes mapped on Figure 4.23 indicated that the Notch signaling pathway is interestingly stable as a result of the mapping of the genes with decreased expression level to the Notch signaling pathway in the reprogramming dataset, and a relatively active Notch signaling pathway is observed when mapped with the genes with increased expression level in the differentiation data.

4.8. Reversibility of the Signaling Pathways in Differentiation and Reprogramming Processes

Genes targeted by the signalling pathways that significantly changed between pairs of experimental time points ($p\text{-adj} < 0.05$) were found and mapped onto the relevant signaling pathways. As a result of this study, a map was obtained for each pair of consecutive time points and the activity of signal pathways was evaluated for each time interval. A target-genes scoring system was developed for evaluation of activity. In order to apply this scoring system, a subset of genes from each signaling pathway were selected as markers for both organisms as shown in Tables 4.1 and 4.2.

Table 4.1: Marker genes selected from relevant signaling pathways in human.

Human											
Pluripotency	SOX2	NANOG	POU5F1	ERK1/2	ESRRB	ZIC3	STAT3	HESX1	FGF2	RAS	
WNT	WNT	DVL	B-CATENIN	TCF/LEF	MYC	JUN	FRA1	CYCD	WISP1	PPAR5	UTERINE
Notch	DELTA	NOTCH	CSL	HES1/5	HEY	PTCRA					
TGF-β	BMP2/4	SMAD1/5/8	ID	P15	SMAD2/3	TGF- β					

Table 4.2: Marker genes selected from relevant signaling pathways in mouse.

Mouse											
Pluripotency	Sox2	Nanog	Pou5f1	LIF	Esrrb	Klf4	JAK	Zic3	Stat3	Hesx1	
WNT	WNT	DVL	β -catenin	TCF/LEF	Myc	Jun	Fra1	CycD	Wisp1	Ppar5	Uterine
Notch	Delta	Notch	CSL	Hes1/5	Hey	Ptcra					
TGF-β	Bmp2/4	Smad1/5/8	Id	p15	Smad2/3	TGF- β					

In the scoring system, the maximum activity level of each signaling pathway was given 100 points. The number of marker genes that significantly changed between time pairs was divided by the total number of marker genes assigned to the selected signaling pathway, then this ratio was multiplied by 100 and assigned as the activation (or inactivation) score of the signaling pathway at the corresponding time interval. Thus, each

signaling pathway was scored according to the extent it was activated or inactivated in the time interval. In order to visualize the obtained scores, they were assigned to the font size of the name of the signal pathway as shown in Figures 4.24. Thus, the font size of the pathway name is an indication of fraction of pathway genes upregulated or downregulated in that time interval.

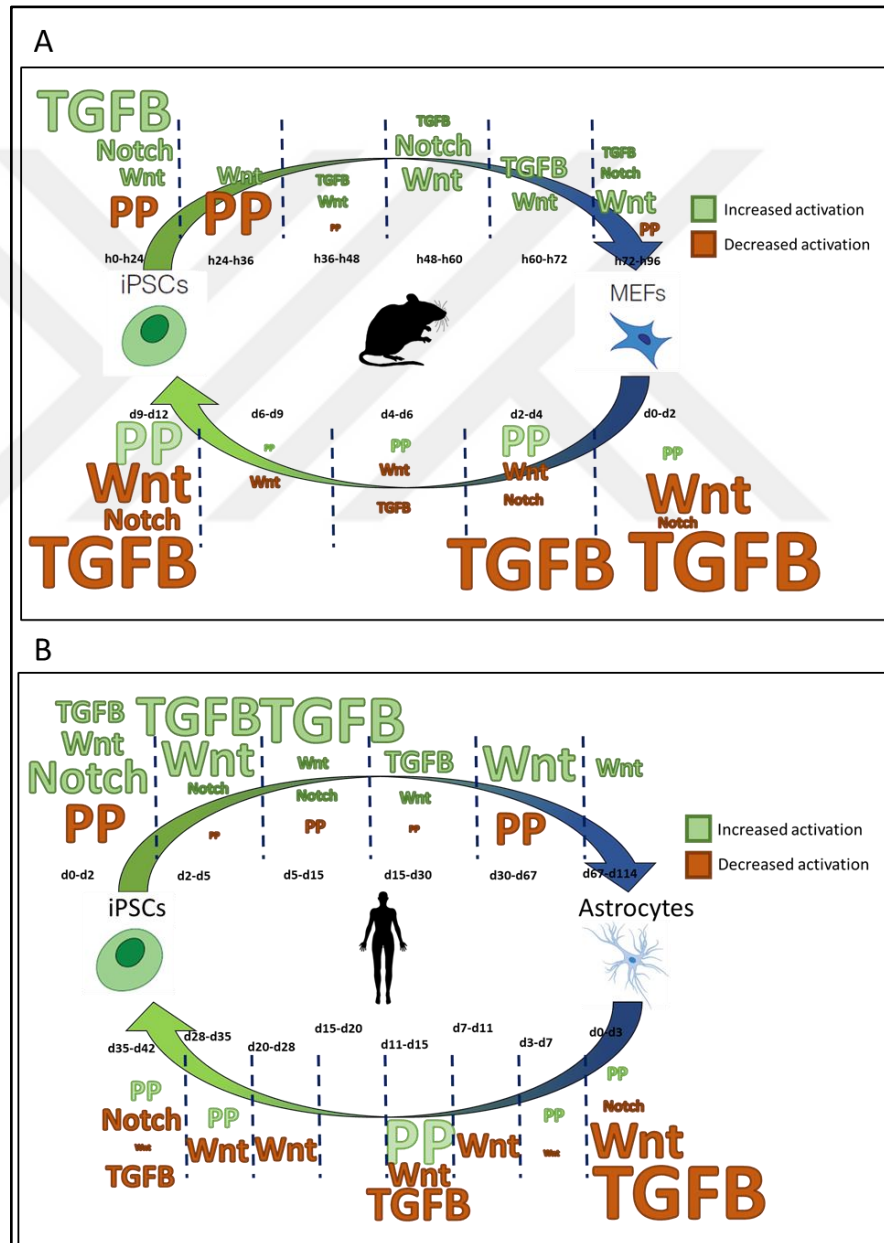


Figure 4.24: Activation and suppression of signaling pathways at time intervals during differentiation and reprogramming processes in both organisms, human and mouse.

In Figure 4.24.A, strong activation of the TGF- β signaling pathway appears to be the first major change in the differentiation of iPS cells to MEF, followed by a decrease in activation in the pluripotency (PP) pathway, differentiation was completed within 96 hours, the signaling pathway changes seen after 60 hours were not as strong as the changes in the previous hours. On the other hand, in Figure 4.24.B, in human, appears to be the first major change in the differentiation of iPSC to astrocytes is the similar activation increase in TGF- β , WNT and NOTCH but in time, TGF- β activation rate accelerated more than the others. So, according to these results, one can say that the TGF- β signaling pathway is reversible and it is key signaling pathway between differentiation and reprogramming processes. Detailed discussion will be done in the Discussion section.

4.9. Effects of Epigenetic Factors in MEF-iPSC Reprogramming Processes

In order to examine the reversibility of the differentiation and reprogramming processes, transcriptome data were analysed as explained above, and the dynamics of the relevant signaling pathways throughout the process were examined based on the expression levels of the relevant genes. Additionally, epigenetic regulations also play a very important role in determining of the cell fate. For this reason, in addition to the transcriptome data, the datasets obtained by using the chromatin immunoprecipitation (ChIP)-sequencing method were also analysed within the scope of this thesis, and the epigenetic changes in the differentiation processes and the effects of these changes on the transcriptome were studied. The pre-processing of ChIP-Seq data was done by PhD candidate Gülben AVŞAR. I want to say thank you to her for her help in the pre-processing steps.

In addition to the microarray and ChIP datasets produced by Chen et al., ChIP dataset produced by Thakurela et al. was also used in this study, whose RNA-sequencing dataset were analysed in the above sections.

Before the analysis of the chromatin immunoprecipitation dataset, the analysis of the microarray dataset (Chen et al. 2016) was performed. The reprogramming dataset from mouse embryonic fibroblast cells to stem cells is a dataset produced by using microarray

method and the limma package was used to find the significantly changed genes in the downstream analysis, 12545 significantly changed genes were obtained using the cut-off $0.01 < p\text{-adj}$. Then, these significantly changed genes were clustered into 9 different clusters, as shown in Figure 4.25, using the K-means clustering algorithm. The biological functions of the clusters were examined by performing GO enrichment analysis on the clusters, and the clusters containing the biological process most related to our biological question were selected for further analysis

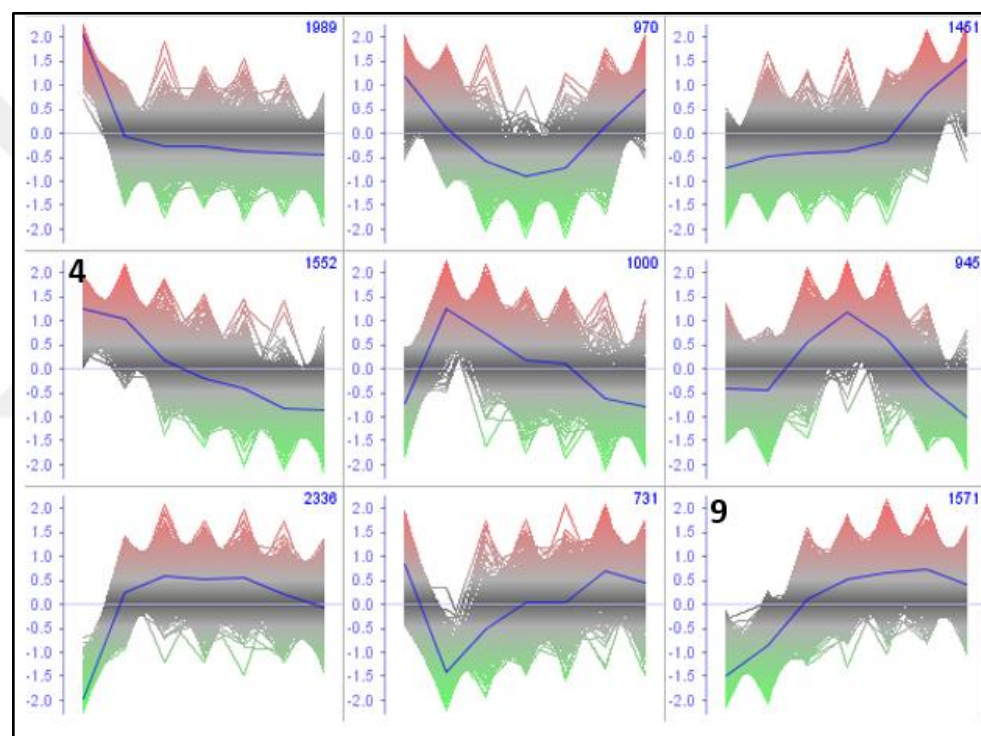


Figure 4.25: Clustering result of Chen et al dataset. On the x-axis time is ordered 0, 1, 3, 5, 7, 11, 15 days, on the y-axis is the level of gene expression, blue curves show the cluster mean. The gene numbers of the clusters are located in the upper right corner of each cluster. The numbers of the clusters analysed by GO are shown above the graph.

Cluster-4 was chosen as genes with decreasing profile, and cluster-9 was chosen as a cluster of genes with increased expression profile. GO biological process enrichment results of the selected clusters are shown in Figures 4.26.

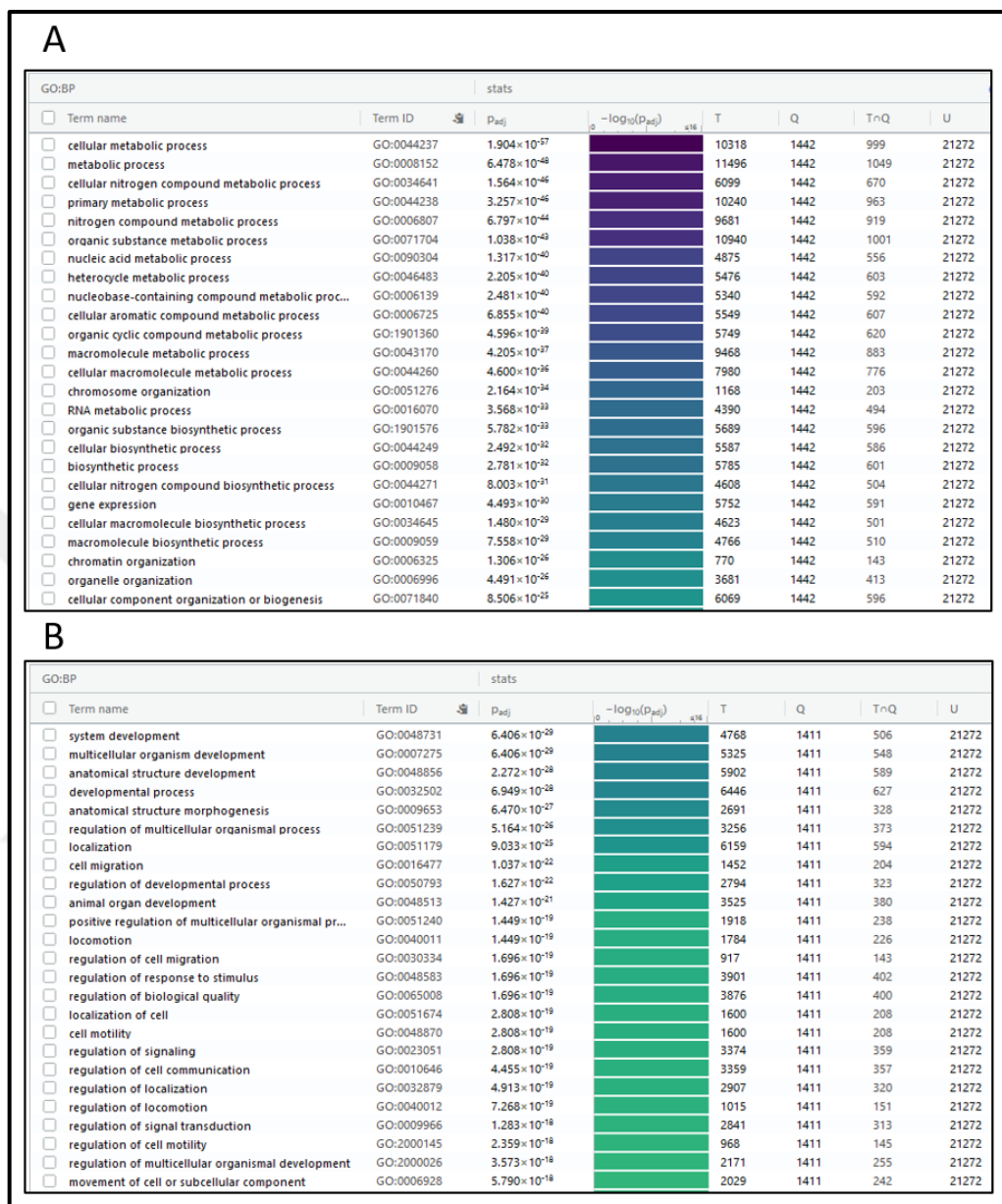


Figure 4.26: GO BP analysis results of selected gene clusters. While A represents genes that have increased expression level, B represents genes that have decreased expression level.

When the GO enrichment results of cluster-9 elements with an increased expression profile were examined in Figure 4.26.A, general terms related to metabolism among the first 25 terms can be seen. Interestingly, the same situation was observed in the biological process GO enrichment results of other datasets. The fact that the terms related to metabolism are seen at the top leads to the conclusion that the differentiation and

reprogramming processes of the cells mostly affect the genes related to metabolism; the fact that terms such as cell cycle and chromatin organization are seen again with very low adjusted p-values, but just below the terms of metabolism, leads to the conclusion that the cell cycle is highly affected in cell differentiation and reprogramming processes, and that the rate changes in the cell cycle may also be closely related to metabolism.

In Figure 4.26.B, results of the GO enrichment analysis of cluster-4 is given, in which the genes showing decreasing profile are clustered, it is seen that the term with the lowest adjusted p-value is system development. The following terms are developmental process, anatomical structure morphogenesis, regulation of developmental process and animal organ development, which can also be related to system development.

Relevant transcription factors were obtained by applying the DREM2 algorithm to the significantly altered genes, as shown in Figure 4.27.

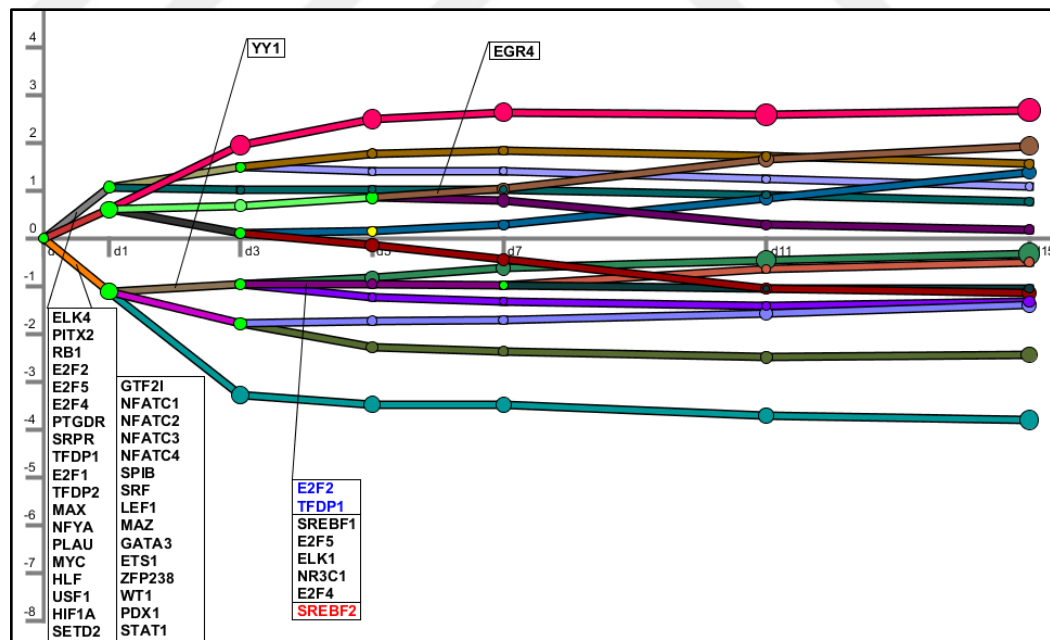


Figure 4.27: The result of co-regulatory analysis of significantly changed genes of Chen et al dataset.

In Chen et al. (2016) ChIP dataset, the profile of H3K27ac marks, i.e., acetylation of lysine amino acids at position 27 of histone tails, was investigated during the reprogramming process. This epigenetic mark, in addition to marking enhancer regions, also marks active promoter regions. The peaks at the promoter regions of 11888 genes in the data were standardized within themselves and clustered into 9 different clusters using the k-means clustering algorithm (Figure 4.28). GO enrichment analyses of selected clusters can be seen in Figure 4.29.

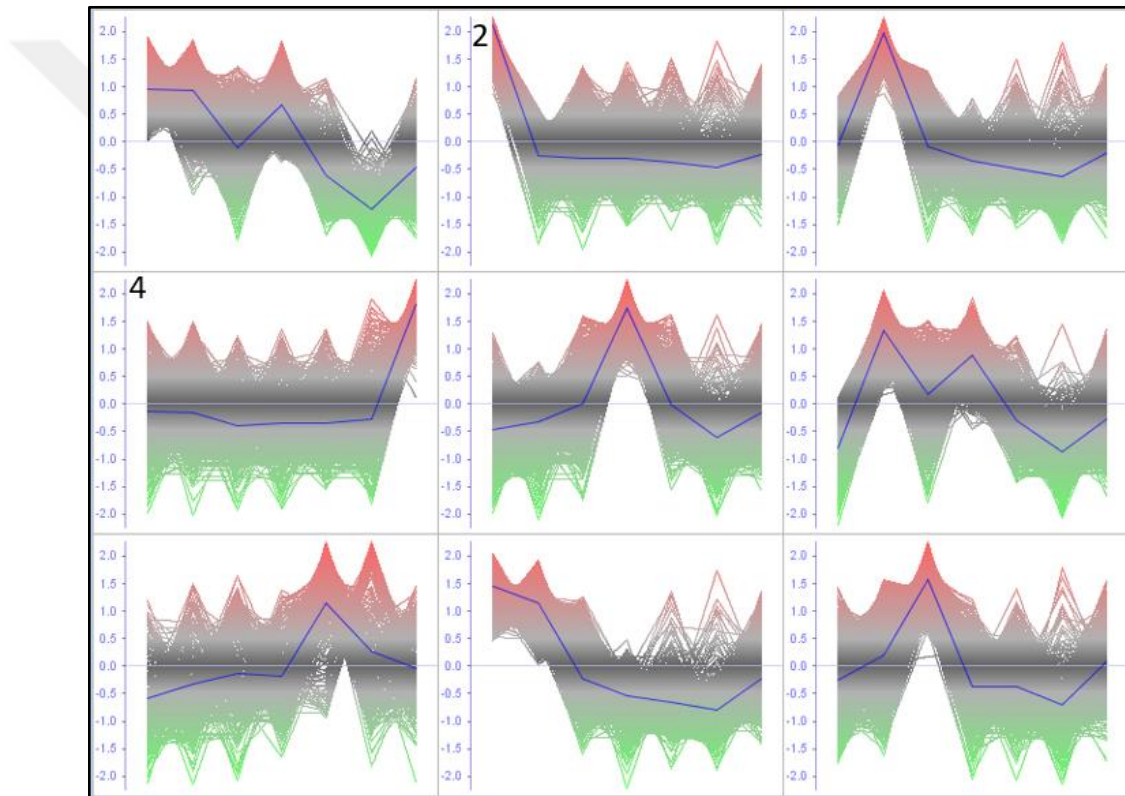


Figure 4.28: Chen et al. ChIP-Seq dataset clustering results. Process of induced stem cell generation from mouse fibroblast cells (data 0, 1, 3, 5, 7, 11, 15 days on x-axis), y-axis given the binding levels of acetyl group to promoter regions of genes, blue curves, shows the cluster average.

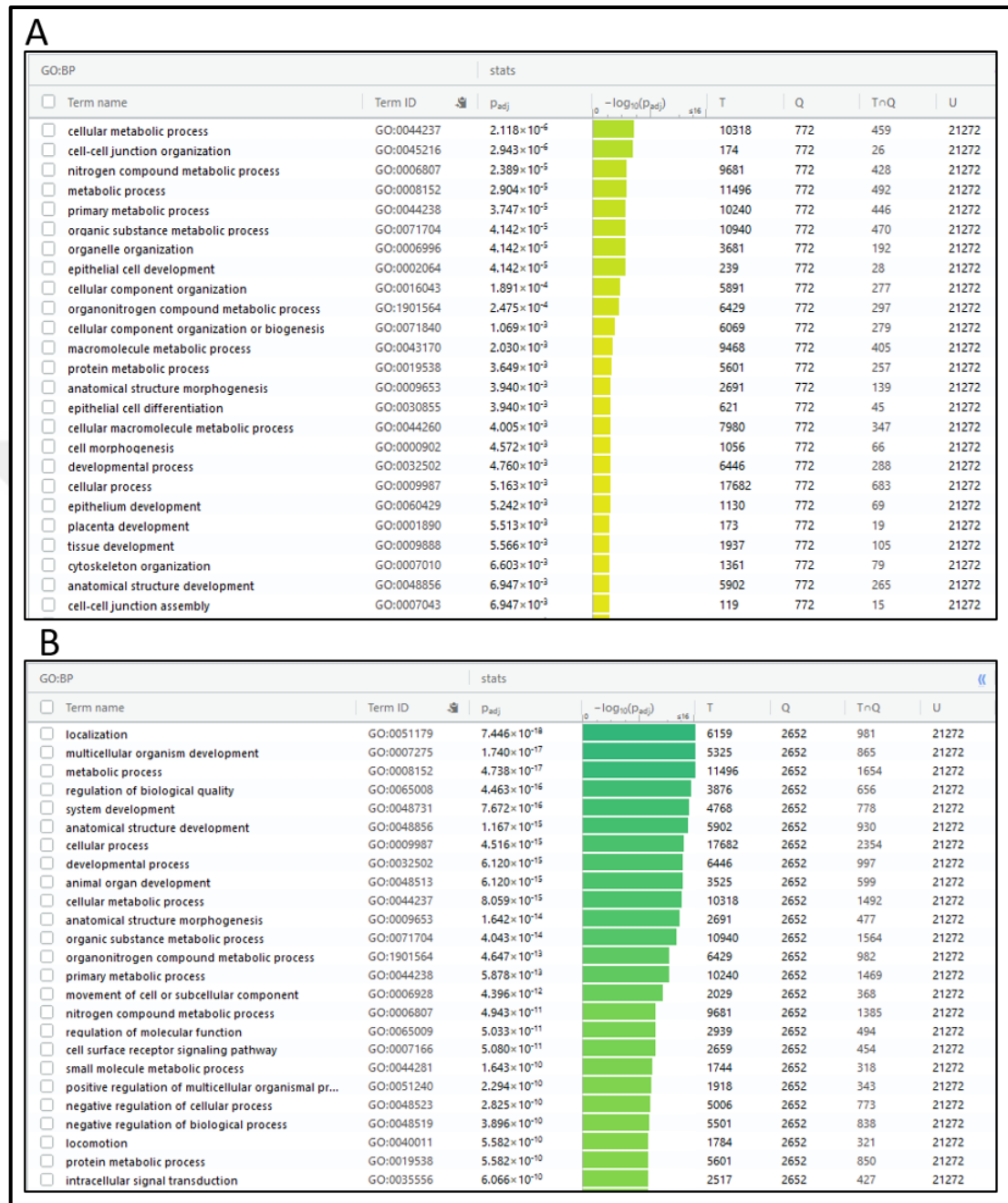


Figure 4.29: GO BP analysis results of selected gene clusters of Chen et al ChIP-seq dataset. While A represents genes that have increased expression level, B represents genes that have decreased expression level.

The GO Biological Process enrichment analysis results of genes with the increased expression profile in Figure 4.29.A indicates that, interestingly, terms related to development stand out. Contrary to many of the previously reported GO analysis results, the adjusted p-values in these analysis results are generally higher and their degree of

significance is low. However, the appearance of the terms like cell-cell junction indicated that a number of genes that provide cellular communication have activate promoters as a result of epigenetic regulation.

The GO Biological Process enrichment analysis results of genes with the decreased expression profile in Figure 4.29.B, indicates that the biological process GO terms with the lowest adjusted p-value are organism development, system development, anatomical structure development and similar terms. As expected, these developmental terms are related to genes whose expression levels are reduced due to the removal of the acetyl groups from the highly acetylated regions during cell reprogramming leading to gene silencing.

4.10. Effects of Epigenetic Factors on Signaling Pathways in MEF-iPSC Reprogramming Processes

In order to understand how the previously discussed signaling pathways are affected by epigenetic factors during reprogramming, heatmaps were drawn using specific time intervals and markers selected from the relevant signaling pathways (Figure 4.30).

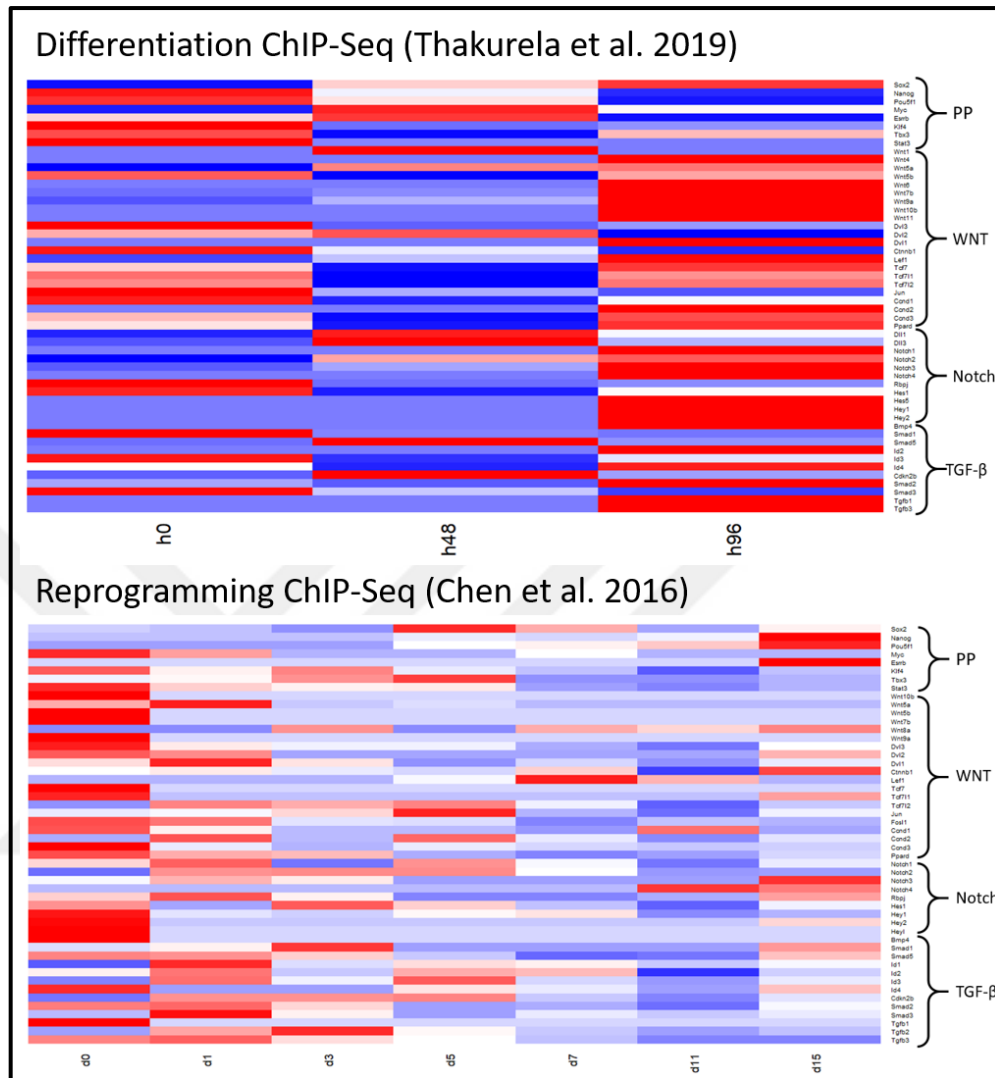


Figure 4.30: Heatmap showing the amount of acetylation in the process of genes involved in Pluripotency, WNT, Notch, TGF- β signaling pathways in the reprogramming and differentiation processes. It is coloured from red to blue, and while the red areas show high acetylation, the blue areas show low acetylation.

ChIP dataset that was generated by Thakurela et al 2019 includes only 3 time points: h0, h48, and h96. When we look at the top heatmap given in Figure 4.30, contrary to the bottom heatmap in Figure 4.30, a general decrease in the amount of acetylation of the pluripotency promoter genes was observed, especially at the last time point. Only Sox2 acted contrary to other pluripotency components. We see a similar situation in the transcript results of the same dataset, the expression level of the Sox2 gene does not decrease, indicating that Sox2 gene is regulated with a different mechanism when compared to other pluripotency TFs.

It is noticeable at the bottom heatmap is that elements in the same signal pathway generally show similar profiles. In addition, it is understood that the amount of acetylation of genes such as Pou5f1 and Nanog, which play a role in the pluripotency signaling pathway and are known to support pluripotency in cells, is low at first, but increases visibly, especially on the 15th day. However, it is observed that the amount of acetylation of genes that play a role in the differentiation of cells, such as WNT, Notch, TGF- β , in signal pathways, decreases over time, unlike pluripotency genes. To compare this case with the dataset that have the reverse direction, same heatmap was drawn with using ChIP dataset that generated by Thakurela et al. 2019.

4.11. Comparison of the Epigenome and Transcriptome Data from the Same Process on Signaling Pathways

It is known that acetylation status has a positive effect on the expression levels of genes. Therefore, in the same data set, in order to test whether there is a similarity between the acetylation status showing an increasing trend and the transcriptome data showing an increasing trend, and also to what extent this similarity is reflected in the pluripotency signaling pathway, genes with increasing and decreasing acetylation levels were mapped to the pluripotency pathway (Figure 4.31).

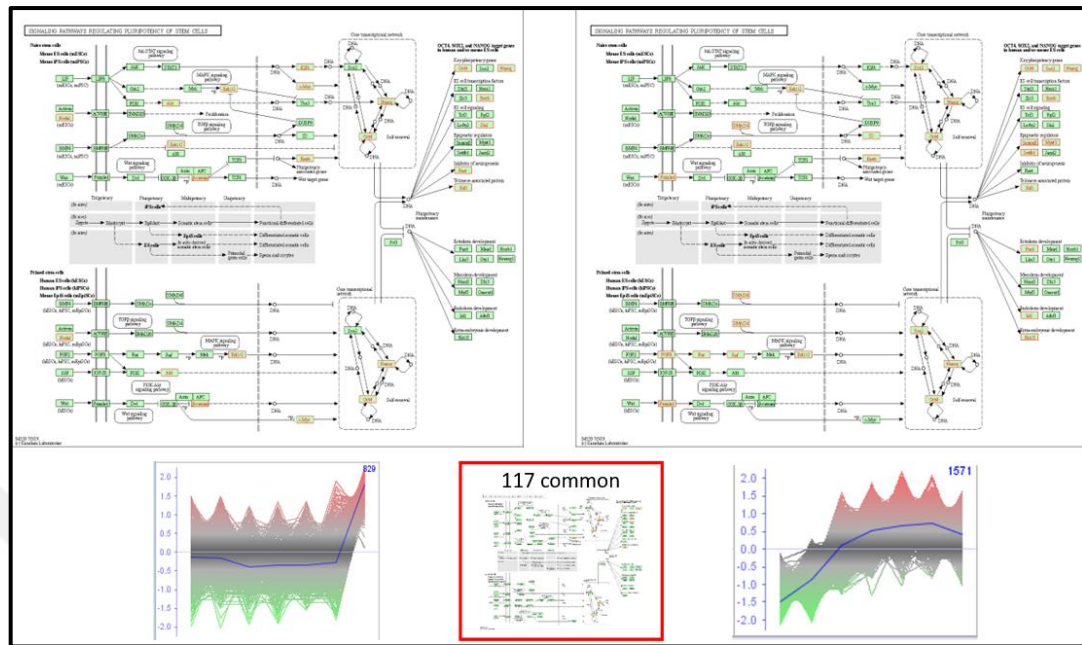


Figure 4.31: Signaling pathway mapped with significantly changed genes with increased expression levels (right). Signaling pathway mapped with genes that show an increasing profile over time with acetylation (left). The datasets generated by Chen et al. 2016 were used as both ChIP and transcriptome data in the figure.

As seen in Figure 4.31, the transcriptome data showing an increasing profile of expression levels indicate that the pluripotency signaling pathway is activated over time, as mentioned before. In addition, when we examine the epigenetic regulations, when we look at the figure on the left, it is seen that the genes with increased expression levels are acetylated.

The same analysis was also performed for genes with decreased expression level and decreased acetylation level, and the results are as shown in Figure 4.32.

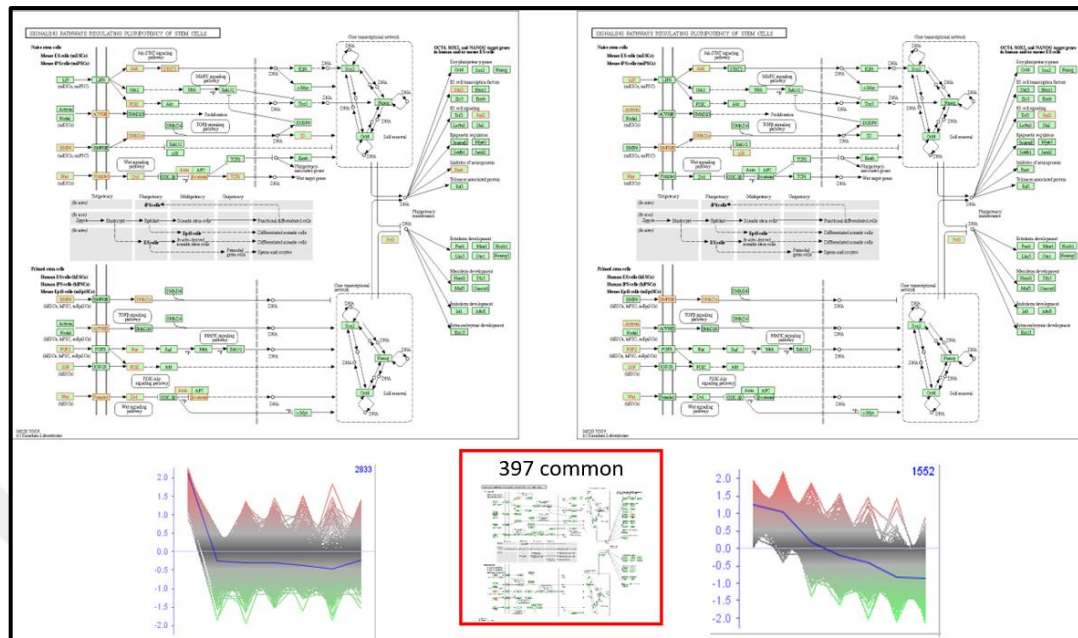


Figure 4.32: Signaling pathway mapped with significantly changed genes with decreased expression levels (right). Signaling pathway mapped with genes that show a decreasing profile over time with acetylation (left). The datasets generated by Chen et al. 2016 were used as both ChIP and transcriptome data in the figure.

At Figure 4.32, it can be seen that the effects of acetylation and transcriptome data decreasing profiles on the signaling pathway are almost exactly the same as expected.

The same analyses were also performed for the transcriptome and ChIP data produced by Thakurela et al. 2019. The result of this study, which was carried out to see the effect of epigenetics on signaling pathways and transcriptome in differentiation processes, is shown in Figure 4.33.

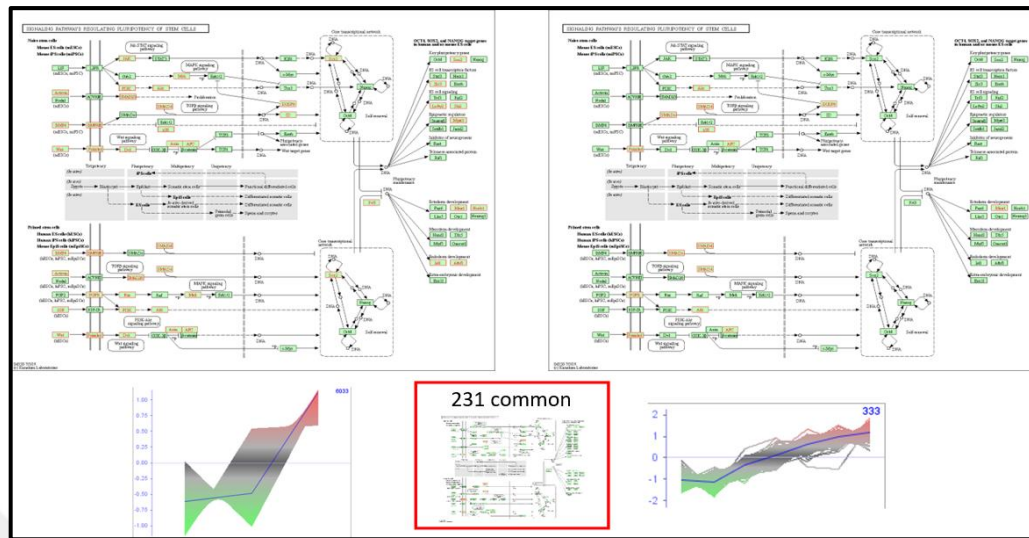


Figure 4.33: Signaling pathway mapped with significantly changed genes with increased expression levels (right). Signaling pathway mapped with genes that show an increasing profile over time with acetylation (left). The datasets generated by Thakurela et al. 2019 were used as both ChIP and transcriptome data in the figure.

It is seen that the Sox2 gene, which is not expected to be acetylated during the differentiation process, is acetylated, but when the expression level is examined, it is not included in the cluster showing an increased profile. It was observed that the same gene behaved abnormally in Figure 4.31, and the expression level increased even though it was not acetylated.

The same analysis was also performed for the genes of the same datasets showing a decreasing profile over time, and the results are given in Figure 4.34.

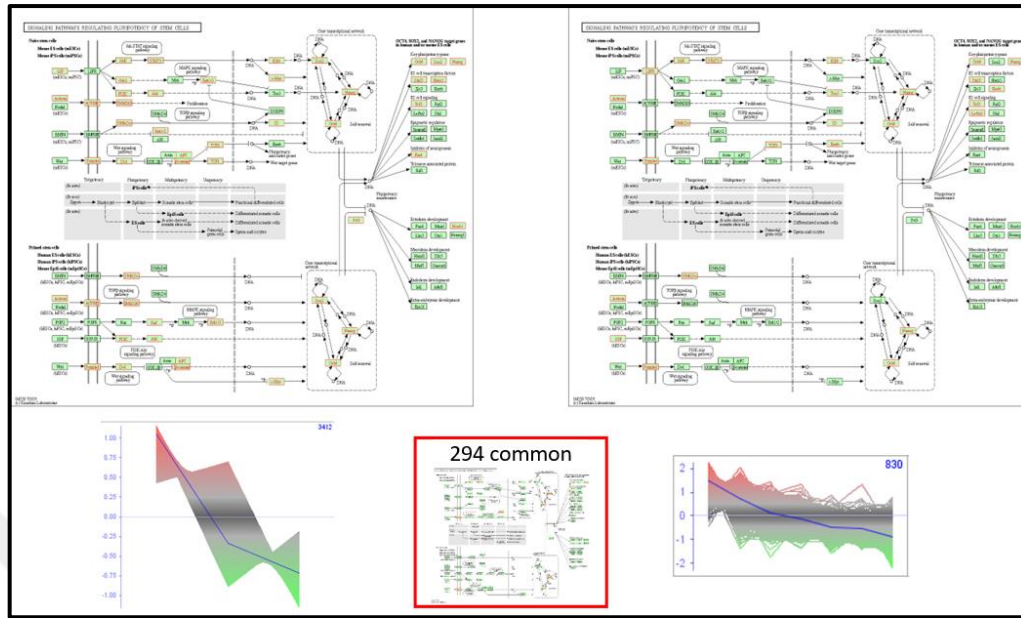


Figure 4.34: Signaling pathway mapped with significantly changed genes with decreased expression levels (right). Signaling pathway mapped with genes that show a decreasing profile over time with acetylation (left). The datasets generated by Thakurela et al. 2019 were used as both ChIP and transcriptome data in the figure.

At Figure 4.34, it is seen that many genes with decreased acetylation level and expression level are common. In addition to the fact that there are 294 common genes in total, in the dataset showing the differentiation process, as expected, a decrease in the acetylation and expression levels of the genes known to support pluripotency was observed in general.

4.12. Comparison of Reversibility by Transcriptome and Epigenome Data

In order to test the reversibility in ChIP and transcriptome data, a reversibility chart was created between Thakurela et al 2019 and Chen et al 2016 datasets using the scoring system mentioned earlier. The transcriptome reversibility scheme is shown in Figure 4.35.A, while the ChIP reversibility scheme is shown in Figure 4.35.B.

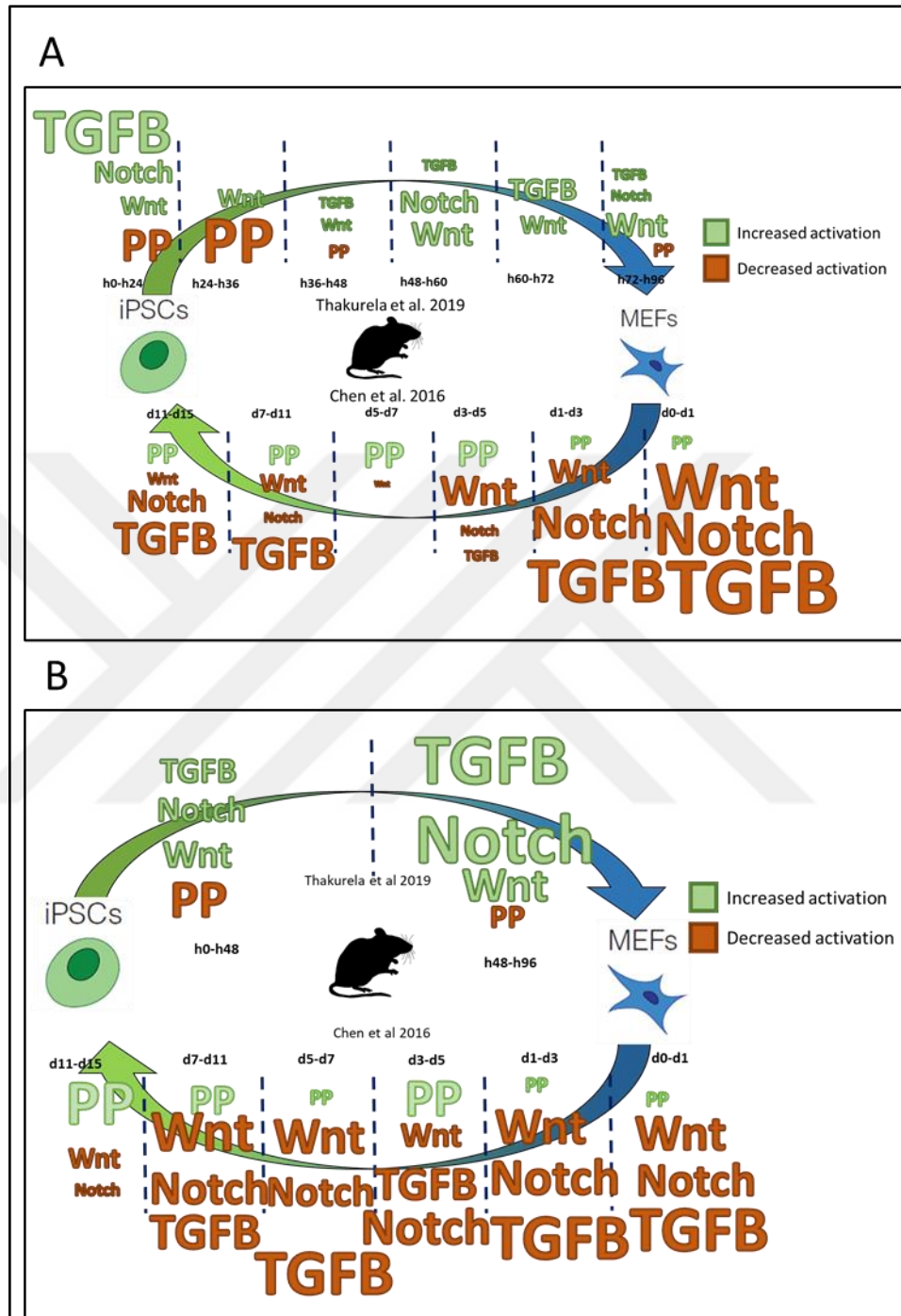


Figure 4.35: Visualization of activation and inactivation at time intervals of signaling pathways in the mouse organism during differentiation and reprogramming processes. A) The figure was prepared using the transcriptome datasets B) The figure was prepared using ChIP-Seq dataset generated by Thakurela et al 2019 and Chen et al. 2016.

Based on Figure 4.35, it will not be possible to make a healthy comparison due to insufficient ChIP data in the differentiation process. However, in general, it can be said

that acetylation of signaling pathways that play a role in differentiation processes such as TGF- β , Notch, WNT increases more sharply, especially at the last time point, as in clustering analysis. On the other hand, it can be said that epigenome data change more gradually than transcriptome data. When we look at Figure 4.35.B, it is observed that the signaling pathways are stably inactivated in reprogramming but gradually activated in differentiation, whereas the transcriptome data shown in Figure 4.35.A shows a more complex profile.

During the cell reprogramming process, acetylation on histone tails leads to increased availability of the DNA region for TF binding and this process may mediate the increase of the expression of the genes. The experimental and bioinformatic procedures for quantification of gene expression (RNA-Seq) and histone tail acetylation (Chip-Seq) differ considerably, however similarity of the GO terms, genes and signalling pathways found in the two datasets validate that epigenetic status of promoters and level of gene expression is tightly coupled, in addition, validate that the data produced independently by these two experimental and bioinformatic procedures reflect the dynamics of the cell reprogramming accurately.

5. DISCUSSION

5.1. Reprogramming and Differentiation of MEF

As mentioned in the Methods and Results section, at the first step of downstream analysis, significantly changed genes were found and then clustered according to their expression trends. Then using the gProfiler web browser, their GO term enrichment results were found. GO results of genes significantly changed in reprogramming and differentiation processes indicate a clear difference between enrichment results.

In the GO enrichment analysis of genes whose expression level increases during the differentiation process in mouse cells (Fig.4.2), we see that the terms with the lowest p-adj values are related to developmental processes such as anatomical structure development and cell differentiation. These results tell us that in the differentiation process, the genes that are related to developmental terms have an increasing trend as expected. In the expression level decreases during the differentiation process, we see the terms response to Leukemia Inhibitory Factor-LIF and cellular response to Leukemia Inhibitory Factor. These two terms are very important in reprogramming process [94]. Seeing LIF related terms are expected in the expression level decreases during the differentiation process.

Interestingly in the GO enrichment analysis of genes whose expression level increases during the reprogramming process, we see that the terms with lowest p-adj values are related to development process. During the reprogramming process, we do not expect development related terms in GO enrichment results, on the other hand the cluster that has increasing expression level trend include Nanog and Pou5f1 (Oct4) genes that are core transcription factors in reprogramming process. Nevertheless, their expression level decreases during the differentiation process, the developmental terms have the lowest p-adj values as expected.

The results of the DREM2 analysis aiming to find common regulators of the significantly changed genes are given in Figure 4.3 (differentiation) and 4.7 (reprogramming). When the TF lists in Figure 4.3, which are seen to be significantly effective in gene expression, are examined, TFs such as Ptf1a, Gata1, Pax4, which support

cellular differentiation and development, are seen among TFs regulating the genes with increasing profile. When the groups showing decreased profile were examined, OSKM factors were not found, but the Stat3 factor, which is the regulator of all these factors and is directly stimulated by LIF, was found. On the other hand, in Figure 4.7 that shows the reprogramming process common regulators and the regulators that have increasing trend are Myc, Sox2, Pou5f1 and Nanog that support reprogramming process (they are not seen in the figure because of the limited figure space, the lists of regulators are accessible in the supplementary file of this thesis.) On the other hand, regulators that have decreasing trend are Foxa1, Gata1, Pax4 that have role in development process. Hence it can be said that in the differentiation and reprogramming process, there is consistency between biological processes, regulatory factors, and gene expression levels. The two biological processes act in opposition to each other in mice.

5.2. Reprogramming and Differentiation of Astrocytes

In the GO enrichment analysis of genes whose expression level increases during the differentiation process in human cells, the terms with the lowest p-adj values are related to developmental processes such as tissue development, animal organ development and anatomical structure development as expected. However, interestingly the genes whose expression level decreases during the differentiation process, the terms are not related with LIF or cell cycle processes. As seen in Figure 4.8 which shows the clustering results of the differentiation process, there are 6 different clusters and there are multiple clusters that have increasing trends, cluster-5 was selected based on development related GO terms enriched among the genes in the cluster. On the other hand, Pou5f1 and Nanog genes are located in the cluster-1 that has decreasing profile during the differentiation process. In system biology, the inferences are done according to the general structure of the data, these general analyses and noise in the data may lead to unexpected results for some of the genes.

In the GO enrichment analysis of reprogramming data, for genes whose expression level increases in human cells, the terms with the lowest p-adj values are related to cell cycle and chromosome organization. Astrocytes are slow dividing cells

while iPSCs are fast dividing cells. For this reason, it is an expected result to see terms related to the cell cycle during reprogramming. In addition, results of the GO enrichment that has decreasing trend are related with developmental process as expected. In this case, we can say that in the reprogramming process increasing trend there are inconsistencies between human and mouse cells. In human, cell cycle and mitosis division can be seen to be affected while in mice they were not affected. One possible reason may be the cell types being studied, astrocytes are more mature cells when compared to MEF, therefore cell cycle in MEF may take place in similar rate to that in iPSC, but in much faster rate than the cell cycle in astrocytes. In addition, the reason for this inconsistency may be the noise of the data or the quality of the clustering.

The results of the DREM2 analysis aiming to find common regulators of the significantly changed genes are given in Figure 4.10 (differentiation) and Figure 4.13 (reprogramming). the TFs that have decreasing trend in differentiation are POU5F1, KLF4, NANOG and SOX2 (they cannot be seen in the figure because of the limited space in the figure, detailed information can be accessed in the supplementary data of this thesis). In addition, SOX1, SOX3, POU3F2 have increasing trend in Figure 4.10 in the differentiation process. According the literature these TFs have different role in differentiation of different cell types such as zebrafish ectoderm, neural cells in human [95–97]. This study reveals that these TFs are also involved in the differentiation of human astrocyte cells. On the other hand, Figure 4.13 shows the common reprogramming regulators and the regulators that have increasing trend are ATF1-2-4-5-7, ESR1 that support reprogramming process. There are some studies demonstrating that overexpression of ATF1 suppresses differentiation of neuro ectoderm [98]. This study also supports that result. However, at the decreasing branch of the chart in Figure 4.13, NR1H2 is seen. No studies on NR1H2 and cell differentiation have been published before. In this study, however, we see that NR1H2 shows a decreasing trend during reprogramming, which reveals a new function for NR1H2.

5.3. Effects of Epigenetic Factors to Signaling Pathway in Reprogramming and Differentiation

As mentioned in the Results section, promoter acetylation positively affects the gene expression, Figures from 4.31 to 4.34 demonstrate that there is consistency between ChIP and RNA-Seq results for the signaling pathways. Figure 4.31 shows the reprogramming process using the data from Chen et al 2016 dataset, both transcriptome data and ChIP-Seq data show profiles with increasing expression levels, indicating that pluripotency signaling pathway is highly activated as expected. In Figure 4.31, Pou5f1, Nanog and Esrrb genes are highlighted in both graphs, therefore, we can say that there is consistency between RNA-Seq and ChIP-Seq data analysis results. The same situation is true for Figure 4.34, Pou5f1 and Nanog genes are highlighted in decreased expression profile for Thakurela dataset. As a result, epigenetic factors affect the gene expression profiles and gene expression profiles effect signaling pathways strongly.

5.4. Outputs of this Thesis That Can Contribute to the Literature

In this study, using different technologies such as microarray, RNA-Seq, ChIP-Seq, datasets from 4 different types of time-series experiments were re-analysed, and results were presented comparatively. While some results confirm the literature, some of them illuminate the points that have not been emphasized much in the literature. In this section, the points that have not been emphasized much in the literature were given under separate subtitles.

5.4.1. Correlation Between Pluripotency Signaling Pathway Gene Expressions

Figure 4.4 in the Results section illustrates a more complex network than it appears on in the pluripotency signaling KEGG pathway because in the KEGG pathway there is only one box for each gene family. For instance, for WNT, there is only one WNT gene box but actually that box refers more than one member of WNT gene such as WNT2,

WNT3, WNT5 and etc. The network given in Figure 4.4 shows each gene individually together with its correlation to its interacting partners on the signalling pathway. It is understood that Jak3, a member of the JAK family, and Stat3 regulated by it show a high correlation. Similarly, it is observed that Pik3r3 from Pi3k gene family and Akt2 genes from AKT gene family are highly correlated with each other. These regulations led to the loss of expression levels of the Sox2, Oct4 and Nanog genes, which fulfill the task of establishing pluripotency. On the other hand, it is also seen that members of the same gene family have a negative correlation with each other. For example, Jak2, which has a negative correlation with the Stat3 gene, may cause a different interpretation and there is no specific information about this issue in the literature.

The other important issue in this Figure 4.4 is inhibition events. For instance, it is known in the literature that Dusp9 inhibits Mapk protein activity, and this interaction has been shown as a 'zigzag' on the network to indicate inhibition, but there is a positive correlation between some gene pairs, although biological inhibition between their protein products is known. There is a zigzag between Dusp9-Mapk1 and Dusp9-Mapk3 but one of them is red (negatively correlated) and another one is green (positively correlated). In literature there is no deep examination between this phenomenon however, within the scope of this thesis, these relationships were examined and revealed from the perspective of transcriptomics.

Protein products of the genes on these pathways regulate each other by phosphorylation in the signalling cascade, hence regulation at gene expression level is not expected. However, the analysis presented here shows complex gene expression regulation patterns among the genes from the same family. Selective positive feedback at the gene expression level, i.e, targets of the signalling pathway activating the specific members of the signalling pathway (such as selective upregulation of Jak3 and downregulation of Jak2 and similarly Mapk1 and Mapk3 being regulated in opposite directions) is an interesting phenomenon which may be used to develop strategies to optimize the protocols for such experiments.

5.4.2. Reversibility of MET-EMT in Differentiation and Reprogramming

MET and EMT are two processes that lead to alternative phenotypes to each other. The somatic cells reprogramming to pluripotent stem cells lose their differentiated identity and assume an embryonic gene expression pattern and growth behaviour. This reprogramming is accompanied by morphological changes as the cells transition from a single layer of adherent cells to become multi-layered epithelial cells, this conversion is called a mesenchymal-to-epithelial transition (MET).

In this study MET-EMT relationship was examined separately in mouse and human. In the mouse, the marker genes of MET are *Esrp1*, *Cldn7*, *Epcam*, *Cdh1* and Figure 4.18 shows that the expression levels of MET marker genes have increased as expected. Especially in the early stages of reprogramming, a sharp increase is noticeable, on the other hand, it is observed that EMT marker gene expression profiles have the opposite profile of MET. Figure 4.18 is created with reprogramming data (Velychko et al. 2019). Contrary to expectations, more complex graphics emerge when the same marker genes are examined with differentiation data (Thakurela et al. 2019). While we expect to see a decreasing profile in Figure 4.19 MET (upper panel), we see a stagnant profile and the other hand at Figure 4.19 EMT (lower panel), we see a decreasing profile at the beginning of the process than this situation changes with the increasing profile interestingly. These profiles demonstrate that these genes are not strictly regulated at the gene expression level.

When the same marker genes were examined in human data, it was observed that the overall profiles were not much different, only the differentiation and reprogramming processes of MET and EMT processes were more active at different time points. When the time-dependent profiles of the expression levels of EMT marker genes are examined, it is observed that the expression levels of the marker genes increase in the early stages of the differentiation in humans, while an increase is observed in the late stages of differentiation in mice. On the other hand, it is possible to see similar profiles in MET marker genes in both organisms. Figure 5.1 summarizes these findings.

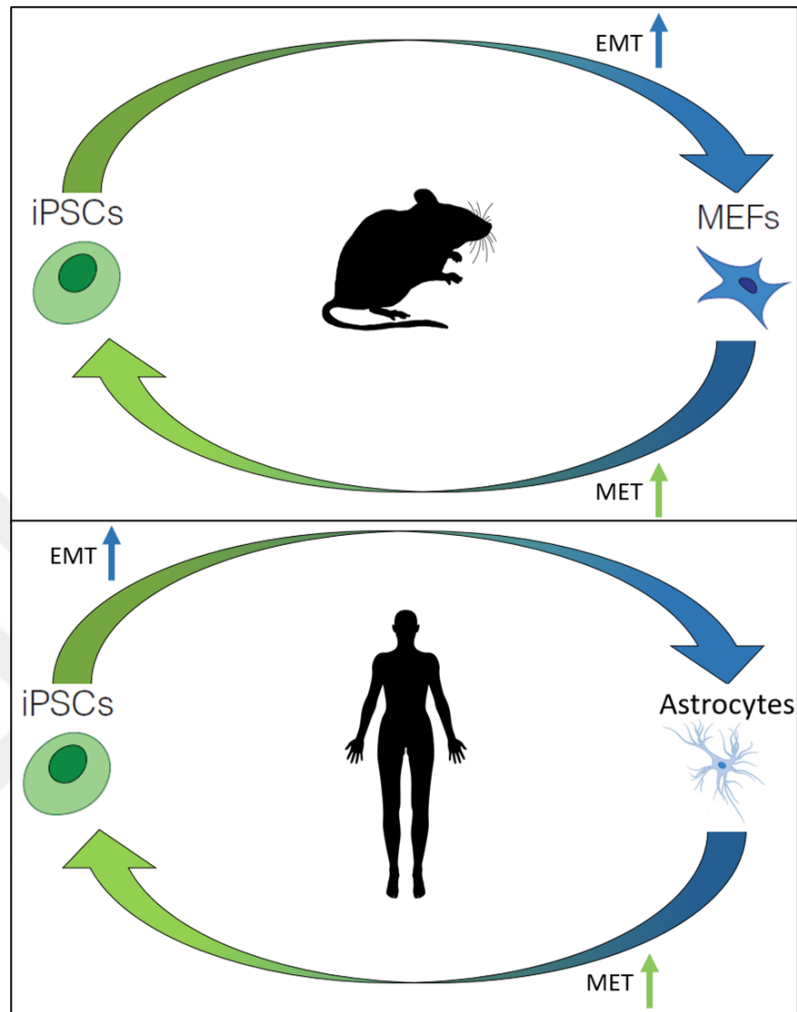


Figure 5.1: The visualization of the MET-EMT processes in mice and human.

5.4.3. Reversibility of Signaling Pathways

Reversibility is one of the main topics of this thesis. While examining the reversibility of reprogramming and differentiation, we should focus on the processes for each organism separately as the signaling pathway dynamics can be different for each organism and cell type.

Figure 4.24 presents the reversibility of gene expression in mice, TGF- β is the first prominent signaling pathway. While fibroblasts are being reprogrammed into stem cells, TGF- β signaling pathway activation seems to decrease sharply, and while stem cells are being differentiated into fibroblasts, activation of TGF- β signaling pathway seems to increase. In both cases, TGF- β activation status is affected in the first step of these two

processes that reprogramming and differentiation. On the other hand, it seems that the WNT signaling pathway always increases its activation in the differentiation process and maintains the same action in the opposite direction, although there are some exceptions in the reprogramming process. Likewise, the Notch pathway does not show sharp increases and decreases in either differentiation or reprogramming processes and moves in opposite directions with similar gradual rate in both processes.

In the mouse induced reversibility scheme, the most important signaling pathway to consider is the pluripotency crossover signaling pathway. In the differentiation process, the activation of the pluripotency pathway decreased especially in the early stages of the process, while in the reprogramming, it increased in the first and early times and maintained its increase with very low scores in the middle of the process. This situation can be interpreted as an interesting result in terms of reversibility.

Another point that should be mentioned here is that these increases and decreases indicate the activation speed. That is, for example, a low score in the direction of increased pluripotency signaling does not mean that the activation of this signaling pathway is decreased, but only that the rate of increase in activation is slow. When examined in this way, we can infer that they are reversible in differentiation and reprogramming processes for all signaling pathways, only slow in terms of increase and decrease rates.

In Figure 4.24, when reversibility is examined in humans, more drastic upregulation and downregulation events are observed in many signal pathways, unlike in mice. In particular, the TGF- β signaling pathway in the reprogramming process and the WNT signaling pathway in the differentiation process increase and decrease at high accelerations. When we look at reprogramming in humans, it seems that the TGF- β pathway decreases sharply, similar to that in the mouse. On the other hand, it is seen that the same signal pathway increases with high accelerations only at the beginning of the differentiation process, then maintains its increase with low scores or remains stable. The pluripotency signal pathway decreased its activity with a sharp acceleration at the beginning and then decreased it with a sharp acceleration during the differentiation process. On the other hand, during the reprogramming process, the activity of the pluripotency signal pathway started to increase slowly in the first time period and then accelerated. In later time periods, it has maintained its acceleration.

It should be noted that the interpretation of pathway activities are made on the basis of the previously mentioned scoring system. If the selected marker genes are changed, the scores and therefore interpretations will also change.

5.4.4. Reversibility of Epigenetic Factors for Reprogramming and Differentiation

For a better understanding of the dynamics of the signaling pathways and their reversibility, we should examine the system with multiple approaches and different perspectives because biology is complex, dynamic and difficult to predict without measurements. Looking at the reversibility of signaling pathways from an epigenetic point of view, a similar dynamic with transcriptome data was observed, and they affect the signaling pathway similarly. Especially for reprogramming, the transcriptomic signaling pathway behaviour is the same with the ChIP-Seq signaling pathway, demonstrating that regulation of gene expression is closely related to promoter acetylation profile. Although my analysis does not imply causality, targeted acetylation of promoter regions may be a useful strategy to accelerate the differentiation or reprogramming experiments.

5.5. Strategies for Future Studies

In the future, experimental reprogramming and differentiation protocols can be improved based on the results presented in this thesis. For example, one of the outputs of this thesis is the dynamics of the MET marker genes, which were increased during reprogramming. In the future, studies can be carried out to increase the efficiency of reprogramming by inducing the expression levels of these marker genes as well as the Yamanaka factors. From the opposite point of view, increasing the expression levels of EMT marker genes early in the stem cell to somatic cell differentiation process may increase efficiency in terms of differentiation rate.

In addition, within the scope of this thesis, it has been determined how much signal pathways such as Notch, WNT and TGF- β increase in which time intervals. In the future, the activation levels of these signaling pathways can be artificially controlled by providing

the corresponding ligands and increased or decreased, resulting in a positive effect on the differentiation and reprogramming efficiency of stem cells.

The cytokine LIF is often added to growth medium for reprogramming and maintenance of iPSCs. However, activity of this ligand may be limited by number of receptors available on the cell membrane. Therefore, the overexpression of LIF receptors may be a strategy to upregulate the activity of pluripotency signaling. LIF gene itself can also be induced for autocrine activation of LIF/Stat3 pathway. Moreover, often MAPK and GSK3b protein inhibitors (named 2i) are added in the growth medium for the same purpose (*i.e.* LIF2i medium), however my analyses revealed that both MAPK and GSK3b are upregulated in gene level in reprogramming. Therefore, either silencing these genes with selective methylation of their promoters or silencing of their master regulator TFs may lead to more sustainable inhibition effect on the culture. A similar analysis may reveal targets in other pathways both in reprogramming and the differentiation protocols.

Finally, gene clusters with significant differences in expression levels in differentiation and reprogramming steps, but which seem to be unrelated to these steps biologically, were also identified in this thesis. Although these findings are left out of interpretation as of today, they will be accepted as data worthy of evaluation in the future and will gain value in terms of the relationship of the differentiation process with other biological processes.

6. CONCLUSIONS

In this thesis, signaling pathways that are important for differentiation and reprogramming processes are investigated with a range of datasets from different technologies. To achieve this; microarray, RNA-Seq and ChIP-Seq datasets were used from different research groups. Reversibility of the signaling pathway is the main focus of this thesis but the other topics like relationship of the transcriptomics with epigenomics and differences between different organisms are also examined. To drive more general results, more datasets should be analysed but topic of this thesis is niche and availability of the datasets dealing with two reversible processes is limited.

Differentiation and reprogramming processes are important processes for different kinds of fields like regenerative medicine, tissue engineering and biomedical engineering. Defining the limitations of differentiation and reprogramming from the epigenomics and transcriptomics perspective will add value to pure and applied sciences.

The signaling pathways such as TGF- β , WNT and their related components were deeply surveyed, and their time-dependant activities were examined and visualized by subjecting them to a scoring system. As a conclusion, according to the data examined in this thesis, it has been revealed that the relevant signaling pathways are in a reversible relationship with each other and this reversibility was also confirmed by ChIP-Seq technology. The effects of epigenetic factors such as histone acetylation on the transcriptome were investigated, and the reflection of these effects on cell differentiation and reprogramming were evaluated. It was observed that the differentiation and reprogramming phenomena were reversible in two different organisms, human and mouse, and the differences between them were inspected in detail in terms of expression levels of genes. In addition, the correlation of genes with each other during differentiation was revealed, negative correlations of some inhibition pairs were confirmed, positive correlations of some inhibition gene pairs were revealed, and the limitations of evaluation based on gene expression levels alone was revealed. The contrariety of the MET-EMT relationship, especially in terms of phenotype, were also discussed in terms of transcriptome, and it was shown that they were opposite to each other at gene expression level. While the expression levels of MET marker genes increased in differentiation, it

was observed that the expression levels of EMT marker genes increased in reprogramming, but it was revealed that the increase and decrease profiles occurred in different time periods in humans and mice.

I hope that the molecular profiles revealed in this study are used in future studies to generate induced stem cells and somatic cells cost and time-effectively, leading to increased efficiency and applicability of regenerative medicine.



REFERENCES

- [1] De Los Angeles A., Ferrari F., Xi R., Fujiwara Y., Benvenisty N., Deng H., Hochedlinger K., Jaenisch R., Lee S., Leitch H. G., Lensch M. W., Lujan E., Pei D., Rossant J., Wernig M., Park P. J., Daley G. Q., (2015), "Hallmarks of pluripotency", *Nature*, 525(7570), 469–478.
- [2] Singh V. K., Saini A., Kalsan M., Kumar N., Chandra R., (2016), "Describing the stem cell potency: The various methods of functional assessment and in silico diagnostics", *Front Cell Dev Biol*, 4(NOV).
- [3] Van Pham P., (2016), "Clinical trials for stem cell transplantation: When are they needed?", *Stem Cell Res Ther*, 7(1), 16–18.
- [4] Ilic D., Devito L., Miere C., Codognotto S., (2015), "Human embryonic and induced pluripotent stem cells in clinical trials", *Br Med Bull*, 116(1), 19–27.
- [5] Bozdağ S. C., Yüksel M. K., Demirel T., (2018), "Adult stem cells and medicine", *Adv Exp Med Biol*, 1079, 17–36.
- [6] 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.
- [7] Yoshida Y., Yamanaka S., (2017), "Induced Pluripotent Stem Cells 10 Years Later", *Circ Res*, 120(12), 1958–1968.
- [8] Gomes K. M. S., Costa I. C., Santos J. F. dos, Dourado P. M. M., Forni M. F., Ferreira J. C. B., (2017), "Induced pluripotent stem cells reprogramming: Epigenetics and applications in the regenerative medicine", *Rev Assoc Med Bras*, 63(2), 180–189.
- [9] Mohammad H. P., Baylin S. B., (2010), "Linking cell signaling and the epigenetic machinery", *Nat Biotechnol*, 28(10), 1033–1038.
- [10] Kitano H., (2002), "Systems biology: a brief overview", *Science*, 295(5560), 1662–1664.
- [11] Tebani A., Afonso C., Marret S., Bekri S., (2016), "Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations", *Int J Mol Sci*, 17(9).
- [12] Baedke J., (2013), "The epigenetic landscape in the course of time: Conrad Hal

Waddington's methodological impact on the life sciences", *Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci*, 44(4), 756–773.

- [13] Takahashi K., Yamanaka S., (2015), "A developmental framework for induced pluripotency", *Dev*, 142(19), 3274–3285.
- [14] Mertens J., Marchetto M. C., Bardy C., Gage F. H., (2016), "Evaluating cell reprogramming, differentiation and conversion technologies in neuroscience", *Nat Rev Neurosci*, 17(7), 424–437.
- [15] Han J., Sidhu K., (2011), "Embryonic stem cell extracts: use in differentiation and reprogramming", *Regen Med*, 6(2), 215–227.
- [16] Kim H. D., Lee E. A., Choi Y. H., An Y. H., Koh R. H., Kim S. L., Hwang N. S., (2016), "High throughput approaches for controlled stem cell differentiation", *Acta Biomater*, 34, 21–29.
- [17] Albert Basson M., (2012), "Signaling in cell differentiation and morphogenesis", *Cold Spring Harb Perspect Biol*, 4(6), 1–21.
- [18] Guo X., Wang X. F., (2009), "Signaling cross-talk between TGF- β /BMP and other pathways", *Cell Res*, 19(1), 71–88.
- [19] Takahashi K., (2014), "Cellular reprogramming", *Cold Spring Harb Perspect Biol*, 6(2), 1–2.
- [20] GURDON J. B., ELSDALE T. R., FISCHBERG M., (1958), "Sexually Mature Individuals of *Xenopus laevis* from the Transplantation of Single Somatic Nuclei", *Nature*, 182(4627), 64–65.
- [21] Davis R. L., Weintraub H., Lassar A. B., (1987), "Expression of a single transfected cDNA converts fibroblasts to myoblasts", *Cell*, 51(6), 987–1000.
- [22] Wilmut I., Schnieke A. E., McWhir J., Kind A. J., Campbell K. H. S., (1997), "Viable offspring derived from fetal and adult mammalian cells", *Nature*, 385(6619), 810–813.
- [23] Huang P., Yan R., Zhang X., Wang L., Ke X., Qu Y., (2019), "Activating Wnt/ β -catenin signaling pathway for disease therapy: Challenges and opportunities", *Pharmacol Ther*, 196(1200), 79–90.
- [24] Nusse R., Varmus H. E., (1982), "Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome", *Cell*, 31(1), 99–109.
- [25] Purro S. A., Dickins E. M., Salinas P. C., (2012), "The secreted Wnt antagonist Dickkopf-1 is required for amyloid β -mediated synaptic loss", *J Neurosci*, 32(10),

3492–3498.

- [26] Rijsewijk F., Schuermann M., Wagenaar E., Parren P., Weigel D., Nusse R., (1987), "The *Drosophila* homolog of the mouse mammary oncogene *int-1* is identical to the segment polarity gene *wingless*", *Cell*, 50(4), 649–657.
- [27] Willert K., Nusse R., (1998), "Beta-catenin: a key mediator of Wnt signaling", *Curr Opin Genet Dev*, 8(1), 95–102.
- [28] Mo Y., Wang Y., Zhang L., Yang L., Zhou M., Li X., Li Y., Li G., Zeng Z., Xiong W., Xiong F., Guo C., (2019), "The role of Wnt signaling pathway in tumor metabolic reprogramming", *J Cancer*, 10(16), 3789–3797.
- [29] Jacques B. E., Montgomery W. H., Uribe P. M., Yatteau A., Asuncion J. D., Resendiz G., Matsui J. I., Dabdoub A., (2014), "The role of Wnt/ β -catenin signaling in proliferation and regeneration of the developing basilar papilla and lateral line", *Dev Neurobiol*, 74(4), 438–456.
- [30] Komiya Y., Habas R., (2008), "Wnt signal transduction pathways", *Organogenesis*, 4(2), 68–75.
- [31] Adams J. M., Jafar-Nejad H., (2019), "The roles of notch signaling in liver development and disease", *Biomolecules*, 9(10).
- [32] Luo Z., Shang X., Zhang H., Wang G., Massey P. A., Barton S. R., Kevil C. G., Dong Y., (2019), "Notch Signaling in Osteogenesis, Osteoclastogenesis, and Angiogenesis", *Am J Pathol*, 189(8), 1495–1500.
- [33] Reicher A., Foßelteder J., Kwong L. N., Pichler M., (2018), "Crosstalk between the Notch signaling pathway and long non-coding RNAs", *Cancer Lett*, 420, 91–96.
- [34] Liu Z. J., Li Y., Tan Y., Xiao M., Zhang J., Radtke F., Velazquez O. C., (2012), "Inhibition of fibroblast growth by Notch1 signaling is mediated by induction of WNT11-dependent WISP-1", *PLoS One*, 7(6).
- [35] Zhang J., Sumich A., Wang G. Y., (2017), "Acute effects of radiofrequency electromagnetic field emitted by mobile phone on brain function", *Bioelectromagnetics*, 38(5), 329–338.
- [36] Zi Z., (2019), "Molecular Engineering of the TGF- β Signaling Pathway", *J Mol Biol*, 431(15), 2644–2654.
- [37] Clark D. A., Coker R., (1998), "Transforming growth factor-beta (TGF-beta)", *Int J Biochem Cell Biol*, 30(3), 293–298.
- [38] Xu F., Liu C., Zhou D., Zhang L., (2016), "TGF- β /SMAD Pathway and Its Regulation in Hepatic Fibrosis", *J Histochem Cytochem*, 64(3), 157–167.

- [39] Hata A., Chen Y. G., (2016), "TGF- β signaling from receptors to smads", Cold Spring Harb Perspect Biol, 8(9).
- [40] Watabe T., Miyazono K., (2009), "Roles of TGF- β family signaling in stem cell renewal and differentiation", Cell Res, 19(1), 103–115.
- [41] Pelullo M., Zema S., Nardoza F., Checquolo S., Screpanti I., Bellavia D., (2019), "Wnt, Notch, and TGF- β pathways impinge on hedgehog signaling complexity: An open window on cancer", Front Genet, 10(JUL), 1–16.
- [42] Carballo G. B., Honorato J. R., De Lopes G. P. F., Spohr T. C. L. D. S. E., (2018), "A highlight on Sonic hedgehog pathway", Cell Commun Signal, 16(1), 1–15.
- [43] Walter K., Omura N., Hong S. M., Griffith M., Vincent A., Borges M., Goggins M., (2010), "Overexpression of smoothened activates the Sonic hedgehog signaling pathway in pancreatic cancer-associated fibroblasts", Clin Cancer Res, 16(6), 1781–1789.
- [44] Horn A., Palumbo K., Cordazzo C., Dees C., Akhmetshina A., Tomcik M., Zerr P., Avouac J., Gusinde J., Zwerina J., Roudaut H., Traiffort E., Ruat M., Distler O., Schett G., Distler J. H. W., (2012), "Hedgehog signaling controls fibroblast activation and tissue fibrosis in systemic sclerosis", Arthritis Rheum, 64(8), 2724–2733.
- [45] Pietrobono S., Gagliardi S., Stecca B., (2019), "Non-canonical hedgehog signaling pathway in cancer: Activation of GLI transcription factors beyond smoothened", Front Genet, 10(JUN), 1–20.
- [46] Nickel J., Mueller T. D., (2019), "Specification of BMP Signaling", Cells, 8(12).
- [47] Gomez-Puerto M. C., Iyengar P. V., García de Vinuesa A., ten Dijke P., Sanchez-Duffhues G., (2019), "Bone morphogenetic protein receptor signal transduction in human disease", J Pathol, 247(1), 9–20.
- [48] Katagiri T., Watabe T., (2016), "Bone morphogenetic proteins", Cold Spring Harb Perspect Biol, 8(6), 1–27.
- [49] Ebisawa T., Tada K., Kitajima I., Tojo K., Sampath T. K., Kawabata M., Miyazono K., Imamura T., (1999), "Characterization of bone morphogenetic protein-6 signaling pathways in osteoblast differentiation", J Cell Sci, 112(20), 3519 LP – 3527.
- [50] Derynck R., Zhang Y. E., (2003), "Smad-dependent and Smad-independent pathways in TGF- β family signalling", Nature, 425(6958), 577–584.
- [51] Ozkaynak E., Schnegelsberg P. N., Jin D. F., Clifford G. M., Warren F. D., Drier

- E. A., Oppermann H., (1992), "Osteogenic protein-2. A new member of the transforming growth factor-beta superfamily expressed early in embryogenesis", *J Biol Chem*, 267(35), 25220–25227.
- [52] Lim J., Thiery J. P., (2012), "Epithelial-mesenchymal transitions: Insights from development", *Dev*, 139(19), 3471–3486.
- [53] Wang H., Unternaehrer J. J., (2019), "Epithelial-mesenchymal Transition and Cancer Stem Cells: At the Crossroads of Differentiation and Dedifferentiation", *Dev Dyn*, 248(1), 10–20.
- [54] Aydin B., Arga K. Y., (2019), "Co-expression network analysis elucidated a core module in association with prognosis of nonfunctioning non-invasive human pituitary adenoma", *Front Endocrinol (Lausanne)*, 10(JUN), 1–15.
- [55] Stéphanou A., Fanchon E., Innominato P. F., Ballesta A., (2018), "Systems biology, systems medicine, systems pharmacology: The what and the why", *Acta Biotheor*, 66(4), 345–365.
- [56] Delles C., Holger H., (2017), "Systems biology approach in hypertension research", *Methods Mol Biol*, 1527, 69–79.
- [57] Likić V. A., McConville M. J., Lithgow T., Bacic A., (2010), "Systems biology: The next frontier for bioinformatics", *Adv Bioinformatics*, 2010.
- [58] Hasin Y., Seldin M., Lusi A., (2017), "Multi-omics approaches to disease", *Genome Biol*, 18(1), 1–15.
- [59] Tieri P., Fuente A. De, Termanini A., Franceschi C., (2011), "Bioinformatics for Omics Data", *Omi A J Integr Biol*, 719, 3–30.
- [60] Le Gallo M., Lozy F., Bell D. W., (2017), "Next-Generation Sequencing", *Adv Exp Med Biol*, 943, 119–148.
- [61] Qian X., Ba Y., Zhuang Q., Zhong G., (2014), "RNA-seq technology and its application in fish transcriptomics", *Omi A J Integr Biol*, 18(2), 98–110.
- [62] Wang X., Chang X., Facchinetti V., Zhuang Y., Su B., (2009), "Proliferation and Survival 1"1–22.
- [63] Agapito G., (2019), "Computer Tools to Analyze Microarray Data", *Methods Mol Biol*, 1986, 267–282.
- [64] Hung J. H., Weng Z., (2017), "Analysis of microarray and RNA-seq expression profiling data", *Cold Spring Harb Protoc*, 2017(3), 191–196.
- [65] Miller M. B., Tang Y. W., (2009), "Basic concepts of microarrays and potential

applications in clinical microbiology", *Clin Microbiol Rev*, 22(4), 611–633.

- [66] Kukurba K. R., Montgomery S. B., (2015), "RNA Sequencing and Analysis.", *Cold Spring Harb Protoc*, 2015(11), 951–969.
- [67] Zhang H., He L., Cai L., (2018), "Transcriptome sequencing: RNA-seq", *Methods Mol Biol*, 1754(1), 15–27.
- [68] Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szczesniak M. W., Gaffney D. J., Elo L. L., Zhang X., Mortazavi A., (2016), "A survey of best practices for RNA-seq data analysis", *Genome Biol*, 17(1), 1–19.
- [69] Finotello F., Di Camillo B., (2015), "Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis", *Brief Funct Genomics*, 14(2), 130–142.
- [70] Alm E., Arkin A. P., (2003), "Biological networks", *Curr Opin Struct Biol*, 13(2), 193–202.
- [71] Thakurela S., Sindhu C., Yurkovsky E., Riemenschneider C., Smith Z. D., Nachman I., Meissner A., (2019), "Differential regulation of OCT4 targets facilitates reacquisition of pluripotency", *Nat Commun*, 10(1), 1–11.
- [72] Velychko S., Adachi K., Kim K. P., Hou Y., MacCarthy C. M., Wu G., Schöler H. R., (2019), "Excluding Oct4 from Yamanaka Cocktail Unleashes the Developmental Potential of iPSCs", *Cell Stem Cell*, 25(6), 737-753.e4.
- [73] Chen J., Chen X., Li M., Liu X., Gao Y., Kou X., Zhao Y., Zheng W., Zhang X., Huo Y., Chen C., Wu Y., Wang H., Jiang C., Gao S., (2016), "Hierarchical Oct4 Binding in Concert with Primed Epigenetic Rearrangements during Somatic Cell Reprogramming", *Cell Rep*, 14(6), 1540–1554.
- [74] Tekin H., Simmons S., Cummings B., Gao L., Adiconis X., Hession C. C., Ghoshal A., Dionne D., Choudhury S. R., Yesilyurt V., Sanjana N. E., Shi X., Lu C., Heidenreich M., Pan J. Q., Levin J. Z., Zhang F., (2018), "Effects of 3D culturing conditions on the transcriptomic profile of stem-cell-derived neurons", *Nat Biomed Eng*, 2(7), 540–554.
- [75] Takahashi K., Tanabe K., Ohnuki M., Narita M., Sasaki A., Yamamoto M., Nakamura M., Sutou K., Osafune K., Yamanaka S., (2014), "Induction of pluripotency in human somatic cells via a transient state resembling primitive streak-like mesendoderm", *Nat Commun*, 5, 1–9.
- [76] Jolliffe I. T., Cadima J., (2016), "Principal component analysis: A review and recent developments", *Philos Trans R Soc A Math Phys Eng Sci*, 374(2065).
- [77] Davis M. P. A., van Dongen S., Abreu-Goodger C., Bartonicek N., Enright A. J.,

- (2013), "Kraken: A set of tools for quality control and analysis of high-throughput sequence data", *Methods*, 63(1), 41–49.
- [78] Andrews S., (2010), "FastQC: A quality control tool for high throughput sequence data.". In: [Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/](http://www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/).
- [79] Felix Krueger, "Babraham Bioinformatics - Trim Galore!". https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed 21 Mar 2021.
- [80] Martin M., (2011), "Cutadapt removes adapter sequences from high-throughput sequencing reads". *EMBnet.journal*; Vol 17, No 1 Next Gener Seq Data Anal. doi: 10.14806/ej.17.1.200.
- [81] Werner J. J., Zhou D., Caporaso J. G., Knight R., Angenent L. T., (2012), "Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys", *ISME J*, 6(7), 1273–1276.
- [82] Kim D., Langmead B., Salzberg S. L., (2015), "HISAT: A fast spliced aligner with low memory requirements", *Nat Methods*, 12(4), 357–360.
- [83] Danecek P., Bonfield J. K., Liddle J., Marshall J., Ohan V., Pollard M. O., Whitwham A., Keane T., McCarthy S. A., Davies R. M., Li H., (2021), "Twelve years of SAMtools and BCFtools.", *Gigascience*, 10(2), 1–4.
- [84] Liao Y., Smyth G. K., Shi W., (2014), "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features", *Bioinformatics*, 30(7), 923–930.
- [85] Love M. I., Huber W., Anders S., (2014), "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", *Genome Biol*, 15(12), 550.
- [86] Protassov R., Dyk D. A. van, Connors A., (2002), "STATISTICS, HANDLE WITH CARE: DETECTING MULTIPLE MODEL COMPONENTS WITH THE LIKELIHOOD RATIO TEST Rostislav Protassov and David A. van Dyk", *Astrophys J*, 571(1), 545–559.
- [87] Dysvik B., Jonassen I., (2001), "J-Express: exploring gene expression data using Java.", *Bioinformatics*, 17(4), 369–70.
- [88] Paul Shannon 1, Andrew Markiel 1, Owen Ozier, 2 Nitin S. Baliga, 1 Jonathan T. Wang, 2 Daniel Ramage 2, Nada Amin 2, Benno Schwikowski, 1, 5 and Trey Ideker2, 3, 4 5, 山本隆久, 豊田直平, 深瀬吉邦, 大森敏行, (2003), "Cytoscape: A Software Environment for Integrated Models", *Genome Res*, 13(22), 426.
- [89] Schulz M. H., Devanny W. E., Gitter A., Zhong S., Ernst J., Bar-Joseph Z., (2012),

"DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data", *BMC Syst Biol*, 6(1), 104.

- [90] Han H., Cho J.-W., Lee S., Yun A., Kim H., Bae D., Yang S., Kim C. Y., Lee M., Kim E., Lee S., Kang B., Jeong D., Kim Y., Jeon H.-N., Jung H., Nam S., Chung M., Kim J.-H., Lee I., (2018), "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions", *Nucleic Acids Res*, 46(D1), D380–D386.
- [91] Rouillard A. D., Gundersen G. W., Fernandez N. F., Wang Z., Monteiro C. D., McDermott M. G., Ma'ayan A., (2016), "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins", *Database*, 2016, baw100.
- [92] Liu Z. P., Wu C., Miao H., Wu H., (2015), "RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse", *Database*, 2015, 1–12.
- [93] Smyth G. K., (2005), "Limma: linear models for microarray data BT - Bioinformatics and Computational Biology Solutions Using R and Bioconductor", *Bioinforma Comput Biol Solut Using R Bioconductor*, 397–420.
- [94] Graf U., Casanova E. A., Cinelli P., (2011), "The role of the leukemia inhibitory factor (LIF) - Pathway in derivation and maintenance of murine pluripotent stem cells", *Genes (Basel)*, 2(1), 280–297.
- [95] Lin Y. M. J., Hsin I. L., Sun H. S., Lin S., Lai Y. L., Chen H. Y., Chen T. Y., Chen Y. P., Shen Y. T., Wu H. M., (2018), "NTF3 Is a Novel Target Gene of the Transcription Factor POU3F2 and Is Required for Neuronal Differentiation", *Mol Neurobiol*, 55(11), 8403–8413.
- [96] Lei X. X., Liu Y., Wang J. X., Cai Q., Yan M., He H. P., Liu Q., Long Z. J., Guan Z., (2020), "SOX1 promotes differentiation of nasopharyngeal carcinoma cells by activating retinoid metabolic pathway", *Cell Death Dis*, 11(5), 1–15.
- [97] Dee C. T., Hirst C. S., Shih Y. H., Tripathi V. B., Patient R. K., Scotting P. J., (2008), "Sox3 regulates both neural fate and differentiation in the zebrafish ectoderm", *Dev Biol*, 320(1), 289–301.
- [98] Yang S. C., Liu J. J., Wang C. K., Lin Y. T., Tsai S. Y., Chen W. J., Huang W. K., Tu P. W. A., Lin Y. C., Chang C. F., Cheng C. L., Lin H., Lai C. Y., Lin C. Y., Lee Y. H., Chiu Y. C., Hsu C. C., Hsu S. C., Hsiao M., Schuyler S. C., Lu F. L., Lu J., (2019), "Down-regulation of ATF1 leads to early neuroectoderm differentiation of human embryonic stem cells by increasing the expression level of SOX2", *FASEB J*, 33(9), 10577–10592.

BIOGRAPHY

Enes AK completed his undergraduate studies in the Department of Molecular Biology and Genetics at Yıldız Technical University between 2014 - 2018. In 2018, he started to study for his master's degree in Bioinformatics and System Biology Programme at Gebze Technical University. During his master's education, he participated in a stem cell research project funded by TUBITAK, project number 116S388. He has been working as Bioinformatician in Massive Bioinformatics R&D company since 2020, December.



APPENDICES

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

- Appendix A: WNT and TGF- β and NOTCH KEGG PATHWAY results.
- Appendix B: Thakurela et al. 2019 and Velychko et al 2019 datasets TF regulator lists.

