University of Algarve

ANALYSIS OF THE TRANSCRIPTIONAL REGULATORY NETWORK UNDERLYING HEART DEVELOPMENT

Rui Sotero Rodrigues Machado

Dissertation to obtain

Master Degree in Biomedical Sciences

Work performed under the supervision of:

PhD Matthias E. Futschik

PhD José Bragança

2013

University of Algarve

ANALYSIS OF THE TRANSCRIPTIONAL REGULATORY NETWORK UNDERLYING HEART DEVELOPMENT

Rui Sotero Rodrigues Machado

Dissertation to obtain

Master Degree in Biomedical Sciences

Work performed under the supervision of:

PhD Matthias E. Futschik

PhD José Bragança

2013

Analysis of the transcriptional regulatory network underlying heart development

ACKNOWLEDGMENTS

First of all, I want to acknowledge my supervisors, PhD Matthias E. Futschik and PhD José Bragança. Without them, taking this step to finish the Master's degree, and specially finish a very complex theme for the master thesis as heart development, would not be possible. I appreciate all the support given regarding the review of the bioinformatic and molecular biology by PhD Matthias Futschik and the molecular biology and gene functionality insight by PhD José Bragança. Finally a special thanks to PhD Matthias Futschik for spending a great amount of time teaching me, from scratch, all the bioinformatic tools I used.

I also want to acknowledge the Sysbiolab team (greatest team ever), for all the help during the master thesis. Especially to Miguel for helping me with some things in R, heatmaps and clusters and to Ravi that was of great help with UniHI. I also want to thanks Dulce, a former member of Sysbiolab, for all the afternoons discussing the better way to analyse/interpret the obtained results.

Quero também agradecer à minha familia por todo o apoio, mas quero agradecer em especial aos meus pais por todo o apoio e incentivo para continuar a estudar e a trabalhar naquilo que gosto. Todo o seu carinho e sustento foi tremendo, pois neste momento de crise, apoiaram-me psicológica e monetariamente quando mais precisei para acabar mais uma etapa da minha vida! Fico, por isso, eternamente grato. Agradeço ainda ao puto, meu irmão, pelos momentos de *relax* durante esta fase atarefada da minha vida.

Por fim, mas não menos importante, quero agradecer à Patrícia que esteve sempre ao meu lado todos estes meses em que estive empenhado na tese, apoiando-me nos bons e maus momentos, altos e baixos que esta montanha-russa, chamada tese de mestrado, proporcionou.

Todas estas peças juntas formam o colossal puzzle que é a experiência de passar pelo mestrado e terminar a tese de mestrado. / All these pieces came together to create this colossal puzzle that is the experience of going through the master and the master's thesis.

Obrigado por tudo/ Thanks for everything.

ABSTRACT

Analysis of the transcriptional regulatory network underlying heart development

Heart development is a highly complex process with a series of precisely spatially and temporally ordered events on molecular level. To understand how these events are controlled and coordinated, it is necessary to study the underlying gene expression and its regulation. While many studies have been carried out in the examination of single genes and their expression patterns, comprehensive analyses of genome-wide expression profiles associated with cardiomyogenesis (i.e. the differentiation of stem cells into cardiomyocytes) are still rare. In fact, no study exists to date which compares and consolidates the publicly available genome-wide measurement for cardiomyogenesis. Such endeavour however is important, as it is well known that individual microarray studies can be seriously compromised by artefacts. In contrast, the combination of various expression studies, which was performed in my study, can lead to more reliable results and help elucidate the different aspects of heart development and repair. Furthermore, a brief study was performed regarding the potential risk of originating cancer or teratomas from stem cell therapy. Finally, I carried out a network-based analysis, to identify regulatory actions between genes, based on published interaction data. This type of analysis can also help to identify novel genes with a role in heart development and provide new valuable targets to future experimental laboratorial analysis. The combination of the multiple dataset is thus an important approach to gain better insights of the different heart development processes as well as regenerative medicine applied to the heart.

Keywords: Cardiomyocytes, Gata4, heart development, induced cardiomyocytes, Meis1, microarray, networks, Smyd1

RESUMO

Análise da rede reguladora de transcrição subjacente ao desenvolvimento cardíaco

O desenvolvimento cardíaco é um processo extremamente complexo com uma série de eventos espácio-temporais precisos ao nível molecular. Para compreender como estes eventos são controlados e coordenados, é necessário estudar a expressão genética subjacente em diferentes organismos, estadios de desenvolvimento celular e a sua regulação. Enquanto que muitos dos estudos realizados foram executados para uma análise individual dos genes e seus padrões de expressão, uma análise compreensiva dos perfis de expressão do genoma associada à diferenciação das células estaminais em cardiomiócitos ainda não existe. **No presente trabalho foi realizada uma meta-análise de resultados publicados em 4 trabalhos independentes prévios, num total de 25 *microarrays*, que definiram a expressão diferencial "pangenómica" durante a diferenciação de células estaminais embrionárias em cardiomiócitos ou durante a transdiferenciação de células somáticas em cardiomiócitos.** Este tipo de análise é no entanto essencial, pois a utilização das expressões de um único *microarray* é pouco fiável, podendo este estar seriamente comprometido por artefactos de natureza humana ou das próprias condições experimentais. Foi realizado um breve estudo relativamente ao risco de ocorrer a formação de teratomas a partir da terapia com células estaminais, com o objectivo de verificar se os genes que são comuns ao cancro e às celulas estaminais são semelhantes aos genes responsáveis pela formação dos cardiomiócitos e/ou cardiomiócitos induzidos.

Este tipo de análise é bastante importante, porque não se baseia apenas nos valores de expressão dos genes das experiências, esta análise vai também procurar validar a expressão dos genes por estudos estatísticos, sendo apenas considerados os genes que têm valores *p-value* ajustados significativos (<0.1). Este tipo de tratamento dos *microarrays* torna possível que os dados obtidos sejam mais fiáveis, podendo considerar que os genes adquiridos na análise apresentam consistentemente o mesmo padrão de expressão nos vários estudos em processos similares, procurando assim incluir genes que ainda não tenham sido ligados ao desenvolvimento e regeneração cardíaca. O estudo dos genes importantes para o desenvolvimento cardíaco definiu certos factores de transcrição essenciais para o desencadeamento desse processo. Esses mesmos factores marcam/promovem de tal forma

o desenvolvimento de linhagens de células cardíacas que a expressão exógena destes factores em células somáticas com funções diferentes leva a uma modificação radical da função e das propriedades destas células, tornando-as em células semelhantes a cardiomiócitos. O mecanismo molecular deste processo chamado transdiferenciação ainda é pouco claro, mas é provavel que involva genes que também sejam importantes para o desenvolvimento cardíaco ou diferenciação.

Para elucidar os mecanismos regulatórios subjacentes, foi construída uma rede de interações (*networks*) dos vários genes obtidos, com base em dados publicados de outros artigos. Foram tidos em consideração os factores de transcrição mais relevantes (tais como o Hand2, Mef2c e Gata4) que têm a capacidade de controlar o destino das células cardíacas. A combinação dos dados de expressão e interacção providenciaram um panorama detalhado da dinâmica dos mecanismos de regulação. Foi possível verificar qual a expressão temporal dos genes obtidos através da sua correlação e que tipo de interacção proteína-proteína existia entre os diversos genes.

A meta-análise dos vários estudos de expressão de genoma utilizados neste trabalho, faz com que este trabalho seja único e original, pois tal tipo de análise nunca foi realizada no contexto de tentar encontrar "novos" genes que estejam ligados ao desenvolvimento cardíaco. Este trabalho permitiu elucidar os diferentes aspectos do desenvolvimento e recuperação cardíaca e que genes podem estar envolvidos nesse processo. Através deste trabalho também foi possivel identificar, com algum grau de confiança, alguns genes potencialmente importantes e que ainda não foram completamente associados ao desevolvimento cardíaco, tal com é o caso dos genes Meis1, Smyd1 ou Zfpm2 e providenciou muitos outros indicadores para possíveis futuras experiências laboratoriais.

A combinação dos diversos *microarrays* foi um passo importante para compreender melhor os diferentes aspectos que estão envolvidos intrinsecamente com o desenvolvimento cardíaco e a medicina regenerativa. A sua posterior combinação com as redes de interação entre os genes levou a uma melhor interpretação dos resultados, possibilitando a compreensão do funcionamento temporal e como interagem entre si.

Palavras-chave: Cardiomiócitos, desenvolvimento cardíaco, Gata4, Meis1, microarray, networks, Smyd1

INDEX

Figure Index i
Table Index vii
Abbreviations x
Introduction 1
1. Biological Background 4
1.1. Embryonic Stem Cells 4
1.2. iPSc 5
1.3. Transdifferentiation of Cardiac Fibroblasts 6
1.4. Cancer Cells 7
2. Methodology 9
2.1. Microarray Technologies & Data Sets 9
2.1.1. General Genetic Expression in Mouse Embryonic Stem Cells 11
2.1.2. Rosetta 1 11
2.1.3. Rosetta2 12
2.1.4. Stanford 12
2.1.5. Human Gene Atlas 12
2.1.6. Mouse Gene Atlas 12
2.1.7. Reprogramming Mouse Fibroblast into Functional Cardiomyocites by Defined Factors 13
2.1.8. Reprogramming Non-Myocytes with Cardiac Transcription Factors 14
2.1.9. hiPSc Differentiation Toward Cardiomyocytes 14
2.2. Gene Expression Analysis 16
2.2.1. Preprocessing of Microarray Data 16
2.2.2. Detection of Differential Expression 17
2.2.3. Clustering of Gene Expression Data 18
2.2.4. Functional Enrichment Analysis 18
2.3. Heart Expression 20
2.4. Network Analysis 21
2.5. Gene Comparative Analysis with Stem and Cancer Cells 22
2.5.1. Gene List for Cancer and Pluripotency 22
2.5.2. Assessing Significance of Common Genes 22
3. Results 23

3.1. General Genetic Expression in Mouse Embryonic Stem Cells 24
3.2. Genetic Expression in Different Tissues 40
3.3. Genetic Expression Between Cardiac Fibroblast and Induced Cardiomyocytes and Differentiation into Cardiomyocytes 44
3.3.1. Reprogramming Mouse Fibroblast into Functional Cardiomyocites by Defined Factors 45
3.3.2. Reprogramming Non-Myocytes with Cardiac Transcription Factors 52
3.3.3. hiPSc Differentiation Toward Cardiomyocytes 58
3.4. Comparison of Biological Process, Molecular Functions and KEGG Pathways between Cardiac Data Sets and Mouse ESc 71
3.4.1. Biological Process 71
3.4.2. Molecular Functions 73
3.4.3. KEGG Pathways 74
3.5. Comparison of Individual Gene Profiles Between Cardiac Data Sets and Mouse ESc75
3.5.1. Positive Early Gene Expression 76
3.5.2. Positive Intermediate Gene Expression 79
3.5.3. Positive Late Gene Expression 82
3.6. Network Strutural Analysis 87
3.6.1. Correlation Network 87
3.6.2. PPI Data Network 90
3.7. Gene Comparative Analysis with Stem Cells and Cancer Cells 92
3.8. “HeartEXpress” Interactive Platform 97
4. Conclusion 99
5. References 102
Annex 109
Annex I 110
Annex II 117
Annex III 118
Annex IV 119
Annex V 120

FIGURE INDEX

1. Biological Background

1.4. Cancer Cells

Figure 1.4.1. Possible roles for the four transcription factors in the generation of iPSc or tumor cells (adapted from Yamanaka, 2007). 8

2. Methodology

2.1. Microarray Technologies and Data Sets

Figure 2.1.1. One-colour channel microarray platform, Affymetrix GeneChip (on the left) and Illumina (on the right). 10

2.2. Gene Expression Analysis

2.2.1. Preprocessing of Microarray Data

Figure 2.2.1.1. Simple steps from CEL files to expression set. It starts with an affybatch file, it is submitted to a background correction, normalization, pm correction and summarization and finally is turned in to an expression set (adapted from Gautier *et al*., 2003). 17

2.4. Network Analysis

Figure 2.4.1. A simplified schematic overview of the Cytoscape functionality. (Schematic adapted from Shannon *et al*., 2012). 21

3. Results

3.1. General Genetic Expression in Mouse Embryonic Stem Cells

Figure 3.1.1. Cluster dendrogram for Gaspar *et* al. 2012 data set. 24

Figure 3.1.2. Expression of Nanog homeobox on Gaspar *et* al. 2012 data set. . . 26

Figure 3.1.3. Expression of SRY-box containing gene on Gaspar *et* al. 2012 data set. 26

Figure 3.1.4. Expression of DNA (cytosine-5-)-methyltransferase 3-like on Gaspar *et* al. 2012 data set. 27

Figure 3.1.5. Expression of POU domain, class 5, transcription factor 1 on Gaspar *et* al. 2012 data set. 28

Figure 3.1.6. Expression of developmental pluripotency associated 2 on Gaspar *et* al. 2012 data set. 28

Figure 3.1.7. Expression of neurofilament (light polypeptide) on Gaspar *et* al. 2012 data set. 30

Figure 3.1.8. Expression of T Brachyury on Gaspar *et* al. 2012 data set. 31

Figure 3.1.9. Expression of eomesodermin homolog on Gaspar *et* al. 2012 data set. 32

Figure 3.1.10. Expression of heart and neural crest derivatives expressed transcript on Gaspar *et* al. 2012 data set. 33

Figure 3.1.11 Expression heart and neural crest derivatives-expressed protein 2 on Gaspar *et* al. 2012 data set. 34

Figure 3.1.12. Expression of collagen and calcium binding EGF domains 1 on Gaspar *et* al. 2012 data set. 35

Figure 3.1.13. Expression of Cbp/p300-interacting transactivator 2 on Gaspar *et* al. 2012 data set. 36

Figure 3.1.14. Expression of troponin T type 2 (cardiac) on Gaspar *et* al. 2012 data set. 37

Figure 3.1.15. Expression of actin, alpha, cardiac muscle 1 on Gaspar *et* al. 2012 data set. 37

Figure 3.1.16. Expression of myosin, heavy chain 7, cardiac muscle, beta on Gaspar *et* al. 2012 data set. 38

Figure 3.1.17. Expression of ankyrin repeat domain 2 (stretch responsive muscle) on Gaspar *et* al. 2012 data set. 39

3.2. Genetic Expression in Different Tissues

Figure 3.2.1. Mean expression in density of transcription factor that are up regulated vs. down regulated Rosetta 1. 41

Figure 3.2.2. Mean expression in density of all genes that are up regulated vs. down regulated Rosetta1. 41

Figure 3.2.3. Genes with the highest averaged expression in all tissue in Rosetta 1 data set. 42

Figure 3.2.4. Genes with the highest averaged expression in all tissue in Rosetta 1 data set. 42

Figure 3.2.5. Group of genes that are clustering together in Rosetta 1. 43

3.3. Genetic Expression Between Cardiac Fibroblast And Induced Cardiomyocytes and Differentiation into Cardiomyocytes

3.3.1. Reprogramming Mouse Fibroblast Into Functional Cardiomyocytes by Defined Factors

Figure 3.3.1.1. Cluster dendrogram for the Ieda *et* al. 2010 data set. 45

Figure 3.3.1.2. Expression of myocyte enhance factor 2C in several cells types on Ieda e*t al*., 2010 data set. 46

Figure 3.3.1.3. Expression of GATA binding protein 4 in several cells types on Ieda e*t al*., 2010 data set. 46

Figure 3.3.1.4. Expression of T-box 5 in several cells types on Ieda e*t al*., 2010 data set. 47

Figure 3.3.1.5. Expression of myosin, heavy chain 7, cardiac muscle, beta in several cells types on Ieda e*t al*., 2010 data set. 48

Figure 3.3.1.6. Expression of Fibroblast growth factor 1 in several cells types on Ieda e*t al*., 2010 data set. 48

Figure 3.3.1.7. Expression of troponin T type 2 (cardiac) in several cells types on Ieda e*t al*., 2010 data set. 49

Figure 3.3.1.8. Expression of actin, alpha, cardiac muscle 1 in several cells types on Ieda e*t al*., 2010 data set. 49

Figure 3.3.1.9. Expression of fibroblast growth factor receptor 1 in several cells types on Ieda e*t al*., 2010 data set. 50

Figure 3.3.1.10. Expression of discoidin domain receptor tyrosine kinase 2 in several cells types on Ieda e*t al,* 2010. data set. 50

Figure 3.3.1.11. Expression of fibroblast growth factor 2 in several cells types on Ieda e*t al*., 2010 data set. 51

3.3.2. Reprogramming Non-Myocytes With Cardiac Transcription Factors

Figure 3.3.2.1. Cluster dendrogram for the Song *et* al., 2012 data set. 52

Figure 3.3.2.2. Expression of GATA binding protein 4 on Song e*t al*., 2012 data set. 53

Figure 3.3.2.3. Expression of heart and neural crest derivatives expressed transcript 2 on Song e*t al*., 2012 data set. 53

Figure 3.3.2.4. Expression of myocyte enhance factor 2C on Song e*t al*., 2012 data set. 54

Figure 3.3.2.5. Expression of T-box 5 on Song e*t al*., 2012 data set. 54

Figure 3.3.2.6. Expression of myosin heavy chain beta on Song e*t al*., 2012 data set. 55

Figure 3.3.2.7. Expression of troponin T type 1 on Song e*t al*., 2012 data set. 55

Figure 3.3.2.8. Expression of actin, alpha cardiac muscle 1 on Song e*t al*., 2012 data set. 56

Figure 3.3.2.9. Expression of fibroblast growth factor receptor 1 on Song e*t al*., 2012 data set. 57

Figure 3.3.2.10. Expression of discoidin domain receptor family, member 1, on Song e*t al*., 2012 data set. 57

3.3.3. hiPSc Differentiation Toward Cardiomyocytes

Figure 3.3.3.1. Cluster dendrogram for the Uosaki *et* al., 2011 data set. 58

Figure 3.3.3.2. Expression of Nanog homeobox on Uosaki e*t al*., 2011 data set. 60

Figure 3.3.3.3. Expression of Pou domain, class 5, transcription factor 1 on Uosaki e*t al*., 2011 data set. 60

Figure 3.3.3.4. Expression of SRY-box containing gene 2 on Uosaki e*t al*., 2011 data set. 60

Figure 3.3.3.5. Expression of T Brachyury on Uosaki e*t al*., 2011 data set. 62

Figure 3.3.3.6. Expression of eomesodermin homolog on Uosaki e*t al*., 2011 data set. 62

Figure 3.3.3.7. Expression of ISL LIM homeobox 1 on Uosaki e*t al*., 2011 data set. 64

Figure 3.3.3.8. Expression of heart and neural crest derivatives expressed transcript 1 on Uosaki e*t al*., 2011 data set. 64

Figure 3.3.3.9. Expression of GATA binding protein 4 on Uosaki e*t al*., 2011 data set. 66

Figure 3.3.3.10. Expression of Nkx2 homeobox 5 on Uosaki e*t al*., 2011 data set. 67

Figure 3.3.3.11. Expression of myocyte enhancer factor 2C on Uosaki e*t al*., 2011 data set. 67

Figure 3.3.3.12. Expression of T-box 5 on Uosaki e*t al*., 2011 data set. 68

Figure 3.3.3.13. Expression of myosin, heavy chain 7, cardiac muscle, beta on Uosaki e*t al*., 2011 data set. 69

Figure 3.3.3.14. Expression of actin, alpha, cardiac muscle 1 on Uosaki e*t al*., 2011 data set. 69

Figure 3.3.3.15. Expression of troponin T type 2 (cardiac) on Uosaki e*t al*., 2011 data set. 70

3.4. Comparison of Biological Process, Molecular Functions and KEGG Pathways between Cardiac Data Sets and Mouse ESC

3.4.1. Biological Process

Figure 3.4.1.1. Heat map for biological processes. 72

3.4.2. Molecular Functions

Figure 3.4.2.1. Heat map for molecular function. 73

3.4.3. KEGG Pathways

Figure 3.4.3.1. Heat map for KEGG pathways. 74

3.5. Comparison of Individual Gene Profiles Between Cardiac Data Sets and Mouse ESc

3.5.1. Positive Early Gene Expression

Figure 3.5.1.1. Venn diagram with number of shared genes for the comparative analysis for heart expression. 76

Figure 3.5.1.2. Biological function (on the right) and expression (on the left) for some of the obtained genes in the overlap for positive early gene expression. 78

3.5.2. Positive Intermediate Gene Expression

Figure 3.5.2.1. Biological function (on the right) and expression (on the left) for the obtained genes in the overlap for positive intermediate gene expression. 79

Figure 3.5.2.2. Venn diagram with number of shared genes for the comparative analysis for heart expression. 80

3.5.3. Positive Late Gene Expression

Figure 3.5.3.1. Venn diagram with number of shared genes for the comparative analysis for heart expression. 82

Figure 3.5.3.2. Biological function (on the right) and expression (on the left) for the obtained genes in the overlap for positive late gene expression. 86

3.6. Network Structural Analysis

3.6.1. Correlation Network

Figure 3.6.1.1. Protein-Protein Interaction Network filtered from publically available interaction studies. 88

Figure 3.6.1.2. Gene expression network with correlation >0.75 in cytoscape for Day 0 expression values in Gaspar (2012) data set. 89

3.6.2. PPI Data Network

Figure 3.6.2.1. Figure 3.7.2.1. Protein-Protein Interaction Network filtered from publically available interaction studies. 90

3.7. Gene Comparative Analysis With Stem Cells and Cancer Cells

Figure 3.7.1. Venn diagram with number of shared genes for the comparative analysis for genes down-regulated during ESc differentiation. 93

Figure 3.7.2. Venn diagram with number of shared genes for the comparative analysis for genes up-regulated during ESc differentiation. 95

3.8. "HeartEXpress" Interactive Platform

Figure 3.8.1. HeartEXpress user interface. 97

Figure 3.8.2. A mouse click on the displayed expression of a gene will open another page displaying all genes that are co-expressed with the chosen gene. 98

2. Methodology

2.3. Heart Expression

Table 2.3.1. Heart expression analysis parameters. 20

3. Results

3.1. General Genetic Expression in Mouse Embryonic Stem Cells

Table 3.1.1. Biological functional analysis in Day 0 on Gaspar *et* al. 2012 data set. 27

Table 3.1.2. Biological functional analysis in Day 1 on Gaspar *et* al. 2012 data set. 29

Table 3.1.3. Biological functional analysis in Day 2 on Gaspar *et* al. 2012 data set. 30

Table 3.1.4. Biological functional analysis in Day 3 on Gaspar *et* al. 2012 data set. 32

Table 3.1.5. Biological functional analysis in Day 4 on Gaspar *et* al. 2012 data set. 33

Table 3.1.6. Biological functional analysis in Day 5 on Gaspar *et* al. 2012 data set. 34

Table 3.1.7. Biological functional analysis in Day 6 on Gaspar *et* al. 2012 data set. 36

Table 3.1.8. Biological functional analysis in Day 7 on Gaspar *et* al. 2012 data set. 38

Table 3.1.9. Biological functional analysis in Day 10 on Gaspar *et* al. 2012 data set. 39

3.3. Genetic Expression Between Cardiac Fibroblast And Induced Cardiomyocytes and Differentiation into Cardiomyocytes

3.3.3. hiPSc Differentiation Toward Cardiomyocytes

Table 3.3.3.1. Temporal alignment between some key marker genes present in Mouse and Human. (General Genetic Expression In Mouse Embryonic Stem Cells and hiPSc Differentiation Toward Cardiomyocytes) 59

Table 3.3.3.2. Biological functional analysis in Day 0 on Uosaki e*t al*., 2011 data set. 61

Table 3.3.3.3. Biological functional analysis in Day 2 on Uosaki e*t al*., 2011 data set. 63

Table 3.3.3.4. Biological functional analysis in Day 5 on Uosaki e*t al*., 2011 data set. 65

Table 3.3.3.5. Biological functional analysis in Day 7 on Uosaki e*t al*., 2011 data set. 66

Table 3.3.3.6. Biological functional analysis in Day 9 on Uosaki e*t al*., 2011 data set. 68

Table 3.3.3.7. Biological functional analysis in Day 11 on Uosaki e*t al*., 2011 data set. 70

3.4. Comparison of Biological Process, Molecular Functions and KEGG Pathways between Cardiac Data Sets and Mouse ESC

Table 3.4.1. Data sets used for the comparison of the multiple processes. 71

3.5. Comparison of Individual Gene Profiles Between Cardiac Data Sets and Mouse ESc

Table 3.5.1. Gene lists used for comparative analysis regarding heart gene expression in early, intermediate and late stage. 75

3.5.1. Positive Early Gene Expression

Table 3.5.1.1. Number of genes present in the overlap with determined biological functions in an early stage. 77

3.5.2. Positive Early Gene Expression

Table 3.5.2.1. Number of genes present in the overlap with determined biological functions in an intermediate stage. 80

3.5.1. Positive Early Gene Expression

Table 3.5.3.1. Number of genes present in the overlap with determined biological functions in a late stage. 83

3.7. Gene Comparative Analysis With Stem Cells and Cancer Cells

Table 3.7.1. Gene lists used for comparative analysis. 93

Table 3.7.2. Significance of common genes in different combinations of gene lists. 93

Table 3.7.3. Gene lists used for comparative analysis. 95

Table 3.7.4. Significance of common genes in different combinations of gene lists. 95

ABBREVIATIONS

ACFGHMT – Adult Cardiac Fibroblast with Gata4, Hand2, Mef2c and Tbx5

CF – Cardiac Fibroblasts

c-Myc – v-myc myelocytomatosis viral oncogene homolog

ESc – Embryonic Stem cells

eSet – Expression Set

Gata4 – Transcription factor, gata binding protein 4

GEO – Gene Expression Omnibus

GHMT – Gata4, Hand2, Mef2c and Tbx5

GMT – Gata4, Mef2c and Tbx5

GO – Gene Ontology

Hand2 – Transcription factor, heart and neural crest derivatives expressed transcript 2

hESc – human Embryonic Stem cells

(h)iPSc – (human) induced Pluripotent Stem cells

iCM – induced Cardiomyocytes

iPSc – induced Pluripotent Stem cells

Klf6 – Kruppel-like factor 6

Mef2c – Transcription factor, myocyte enhancer factor 2c

mESc – mouse Embryonic Stem cells

mESCDiff – mouse Embryonic Stem Cell Differentiation

MyoCD – Myocardin

Nanog – Nanog homeobox

Pou5f1 – Transcription factor also known as Oct4, POU class 5 homeobox 1

PPI – Protein-Protein Interaction

Tbx5 – Transcription factor T-box 5

TP53 – Tumor suppressor protein p53

WiCM – Week induced Cardiomyocytes

INTRODUCTION

The process of heart development is highly complex, involving a series of precisely spatially and temporally ordered events on molecular level. To better understand how these events are controlled and coordinated, it is necessary to study the gene expression and its regulation, ideally on a genome-wide level. Using microarray technology, several studies genome-wide expression profiles associated with heart development have been carried out. Although they have given comprehensive views of expression changes during cardiogenesis, they may have their pitfalls. It is well known that single microarray experiment might be seriously compromised by artefacts. Thus, it is important to compare and consolidate the genome-wide microarrays studies. However, to date no meta-analysis of microarray data for cardiogenesis have been attempted. Such lack of assessment can be considered as a serious obstacle on the way to understand how to improve our knowledge in cardiac regenerative medicine.

This thesis aims to provide such critical meta-analysis, which integrates publically available data sets for cardiogenesis to obtain a detailed and reliable comprehensive view of expression changes in cardiogenesis. Using the consolidated data, I elucidate the networks underlying heart development which are still only rudimentarily understood. Such endeavour will help to get a better view of this intrinsically complex process and give valuable cues for regenerative medicine applied to heart.

The thesis can be divided in three different parts: In the first part, I will describe the collection of publicly available data sets from various studies. These data sets comprise expression data for various types of cells, such as, embryonic stem cells, induced cardiomyocytes derived from cardiac fibroblast, cancer cells and several types of tissue samples. Such a broad basis can help to distinguish transcriptional patterns which are specific to heart development from more generic gene expression patterns.

The second part focuses on the analysis of those data sets and provides a comprehensive view of the dynamics of gene expression during cardiomyogenesis. The analyses included some standard pre-processing, background correction and summarization of expression values. I also applied clustering methods to identify potential co-regulated genes.

Subsequently, I have identified transient expression patterns through clustering methods and evaluated their functional relevance to obtain indications which processes are active during the different stages of differentiation. Besides stem cell differentiation, a focus of the thesis is set on the reprogramming of somatic cells into cardiomyocytes. Originally, it has been demonstrated in the landmark studies led by Shinya Yamanaka that fibroblasts can be turned into to so called "induced Pluripotent Stem cells" (iPSc) through expression of mere four transcription factors which have a critical role in Embryonic Stem cells (ESc) properties [1]. These iPSc appear to be indistinguishable from ESc in their morphology, proliferation and gene expression, but care might be needed, as iPSc have been linked to teratoma formation [1] and immune responses [2]. Notably, it is also possible to induce cardiomyocytes from iPSc efficiently [3]. Interestingly, the ectopic expression of pro-cardiogenic transcription factors in fibroblast will originate cells with properties similar to those of cardiomyocytes. For reprogramming fibroblasts into cardiomyocytes, a single master regulator has not been found for cardiac differentiation, but several core transcription factors are being extensively studied [4].

As last part of this work, a network analysis was carried out. Here, I constructed a molecular interactions networks to examine regulatory actions occur based on published interaction data.

Combining expression and interaction data, gives a detailed picture of the dynamic regulatory mechanisms underlying heart development. In the end, by putting all things together, I will seek to identify and characterize any additional key regulators, which will possibly provide some major candidates for a future experimental validation in cardiomyogenesis.

Reprogramming of cells toward cardiac fate

There are core set of transcription factors that are highly conserved through evolution, controlling cell fate, cardiac gene expression and heart development. Remarkably, it was reported a few years ago that the exogenous expression of Gata4, Mef2c and Tbx5 were capable of converting neonatal cardiac fibroblast into cardiomyocytes-like cells *in vitro* [4]. Ieda and co-workers also revealed that it was possible to reprogram fibroblasts directly into "induced cardiomyocytes" (iCM) without first becoming a stem or progenitor cells.

The cooperative interaction between these transcription factors is consistent with their ability to activate cardiac gene expression and activate each other's expression in adult cardiac cells [5]. A subsequent study has indicated that the most effective combination of transcription factors for this process is Gata4, Hand2, Tbx5 and Mef2c. [5]. The efficiency of cellular reprogramming into induced cardiomyocytes by these 4 transcription factors is comparable to the reprogramming iPSc by pluripotency factors. Finding an ideal combination is not easy, as including additional factors might promote but might also hinder reprogramming. For instance, Song (2012) determined that Nkx2-5, a primary marker of cardiomyocytes, decreased the efficiency of the five transcription factors used in his experiment. Another gene that is important for cardiomyogenesis is Isl1 that is transiently expressed in early cardiac progenitor cells before further cardiac differentiation. However, it may not be activated during reprogramming, as Inagawa and colleagues [6] observed that fibroblast are converted directly into differentiated cardiomyocytes without passing through a progenitor cell state, showing that expression of Isl1 is not essential for creating iCMs. To confirm the correct transdifferentiation of cells into induced cardiomyocytes, the expression of standard cardiac markers necessary for cardiomyocyte functions, such as cardiac troponin T, also known as Tnnt2 and alpha-myosin heavy chain, also known as Myh7, in late cell differentiation or in differentiated cell stage [5] was analysed.

While the reprogramming event appears to be stable at the epigenetic level, the global gene expression of iCMs and neonatal cardiomyocytes are similar, but not identical. Despite remarkable progress, it still remains important to optimize the combination of transcription factors for reprogramming of cells such as fibroblasts into fully functional induced cardiomyocytes-like cells, as so far the efficiency of reprogramming is fairly limited. This achievement would be even greater if the cultured cells do not need to go through an induced pluripotent cell (iPSc) stage.

Moreover, it is important in this context, to carefully assess the similarity of gene expression patterns of induced pluripotent cells with cancer cells, because these can share some characteristic features such as the capacity of infinite division.

Finally, combination of various expression studies and the subsequent analysis of the interaction networks can help elucidate the different aspects of heart morphology, development and repair.

1. BIOLOGICAL BACKGROUND

For a better understanding of the differentiation and/or transdifferentiation process of an ESc, iPSc and somatic cell towards a cardiac cell fate, I analysed several publically available microarrays data sets obtained from embryonic stem cells and induced pluripotent stem cells at various stages of differentiation, as well as transdifferentiation of cardiac fibroblasts into cardiomyocytes. I used this comparative analysis to distinguish commonly expressed genes from genes activated only under certain conditions that are relevant for cardiac differentiation. In the following, I will introduce and describe the main types of cells whose expression was analysed in my study.

1.1. EMBRYONIC STEM CELLS

ESc are derived from the inner cell mass of blastocyst embryos and they have a unique capacity to proliferate extensively while maintaining pluripotency [2].

ESc can be easily identified, isolated and maintained in a pluripotency state. Notably, it is possible to keep them in culture for long periods of time, serving as a back-up cells, ready to use [7]. As they are pluripotent, they can differentiate into germ cells or any derivative of the three primary embryonic germ layers: Endoderm (lungs); Mesoderm (muscle, blood, heart); Ectoderm (epidermal tissue and nervous system). They can also be genetically manipulated and numerous protocols have been established that allow the differentiation of embryo-derived stem cells into almost any type of cell [7]. For maintenance and regulation of pluripotency, specific transcription factors in ESc such as the Pou5f1, Sox2 and Nanog, play a very important role.

Oct4 expression is restricted to the blastomeres of the developing mouse embryo, in the inner cell mass of the blastocyst, epiblast and germ layers. It is also expressed in pluripotent stem cells, including embryonic stem cells, embryonic germ cells and embryonic carcinoma cells [2]. Oct4 plays an important role in the maintenance of pluripotency and promoting differentiation [2].

Sox2 also marks the pluripotent lineage of the early mouse embryo and, unlike Oct4, Sox2 is also expressed by multipotent cells of extra embryonic ectoderm [2]. Its expression is also associated with uncommitted dividing stem cells and precursor cells of the central nervous system [2].

Nanog is expressed in ESc and not expressed in differentiated cells, indicating that this gene expression is responsible for pluripotency cell maintenance [8]. Endogenous Nanog expression in parallel with Stat3 drives ESc self-renewal [8]. Nanog overexpression is enough for clonal expansion of ESc, maintaining Oct4 levels elevated. This shows that this transcription factor is essential for defining ESc identity [8]. Down regulation of Nanog cause ESc to lose pluripotency and to differentiate into extraembryonic endoderm lineage [9].

Nevertheless, there remain some challenges in determine optimal conditions to maintain cells pluripotent or promote their efficient and reliable differentiation using ESc: (i) Experimental conditions or small variations of culture techniques can lead to different outcomes; (ii) Onset and shutdown of important biological factors occur in a narrow time window, so we could fail to detect or interpret correctly important biological processes; (iii) comprehensive functional genomics technologies, stringent statistical criteria and bioinformatics analysis are necessary to get a deep view into the complex biological processes occurring at the different time points of the differentiation.

Despite these potential obstacles, ESc lines are considered as encouraging donor sources to repair or replace damaged tissue, reverse diseases and injuries and cell transplantation therapies for diseases such as diabetes, cardiovascular and blood diseases [2, 10]. For applications related to tissue regeneration ("regenerative medicine"), it will be crucial to understand in detail how stem cells organise *in vivo* the generation, maintenance and regeneration of tissue, while preventing or suppressing abnormal growth and avoiding depletion [7].

1.2. IPSC

It was showed that iPSc can be generated from adult human and mouse fibroblast and other somatic cells by ectopic expression of four transcription factors (Klf4, Sox2, c-Myc and Pou5f1 - also known by Oct4) [1, 2] that play important roles in the maintenance of pluripotency. Although, other combinations of factors have been shown to also achieve reprogramming of iPSc, the quartet of Oct4, Sox2, Klf4 and c-Myc (OSKM), referred commonly as Yamanka's factors, are the most commonly used for reprogramming [1, 2, 11, 12]. The established human iPSc are similar to hESc in many aspects such as morphology, proliferation, gene expression and promoter activities, surface markers, *in vitro*

differentiation and teratomas formation [1]. Besides Oct4 and Sox2, two transcription factors are used:

c-Myc is an oncogene found in human cancers, involved in the transactivation of CBP and p300, which have histone acetylase activities, immortalizing, regulating the expression of non-coding RNAs and opening the chromatin of the cells [2].

Klf4 is a member of the family of Krüppel-like transcription factors and has an ambivalent role. It is highly expressed gene in differentiated post-mitotic epithelial cells of the skin and fibroblast. High levels of Klf4 RNA can be found in cells during growth arrest and it is almost undetectable in cells that are in exponential phase of proliferation, indicating to function as a tumor suppressor. Depending on the status of its target genes (especially of p21 and p53), it can also act as an oncogene promoting proliferation. This role is also supported by a high expression observed in some cancers as well as in mouse ESc [2].

During the reprogramming process, c-Myc promotes the immortalization and opens the chromatin, whereas Klf4 is essential to suppressed p53 (a tumor suppressor gene) and c-Myc-induced apoptosis [2, 13], so the balance of expression between the two factors might be critical for the reprogramming process from somatic cells/fibroblast into iPSc.

Forced expression of c-Myc and Klf4 alone would result in generation of tumor cells (figure 1.4.1), so it would be necessary to combine their induction with other factors, and here Oct4 plays an important role that prevents cells from turning into tumor cells. However, Oct4 alone is not sufficient to induce pluripotency, being Sox2 also required to active multiple target genes for generating iPSc [1, 2].

1.3. TRANSDIFFERENTIATION OF CARDIAC FIBROBLASTS

The heart is composed approximately by 30% of cardiomyocytes and 60-70% of cardiac fibroblast, being cardiac fibroblasts the prevailing cell type in the adult mammalian heart [4, 5]. The large population of cardiac fibroblast existing in the heart could be a potential source for induced cardiomyocytes for the purpose of regenerative medicine [4].

Indeed, cardiac fibroblast were already successfully reprogrammed into pluripotent cells and subsequently directly reprogrammed or transdifferentiated into induced cardiomyocytes (iCM), by combination of cardiac lineage-associated transcription factors [1, 2, 5, 14, 15]. A core set of evolutionary transcription factors controls cardiac gene expression and heart development, amongst which,Gata4 is considered a master regulator in heart

development, being capable of binding Mef2c and Tbx5 to their specific target sites, leading to complete activation of the cardiac programming process [5].

Recently, Gata4, Mef2c and Tbx5 (GMT), master regulators of cardiac development, were reported to be capable of converting neonatal fibroblast into cardiomyocytes in vitro when exogenously overexpressed during 2 weeks [4, 15]. In another study, these transcription factors in in combination with Hand2 (GHMT), another transcription factor involved in heart development, were used to convert cardiac fibroblasts into functional cardiomyocytes in vitro and directly in vivo in mouse infracted hearts [5].

The direct conversion of cardiac fibroblasts into cardiomyocytes constitutes an attractive paradigm. In this case, reprogramming into iPSc before cardiac differentiation is not necessary and this direct conversion would probably significantly lower the risk of originating teratomas or non-specific tissue. This approach is also attractive for future heart therapies, since a great amount of cardiac fibroblast can be cultivated from a simple cardiac biopsy, and performing an *in vitro* transduction with the defined factors to obtain a large amount of iCM is relatively easy, before their delivery into the damaged heart. More interestingly, is the potential delivery of cardiomyocytes reprogramming factors to convert *in vivo* cardiac fibroblasts of the scar tissue generated in infracted hearts into functional cardiomyocytes. Furthermore, it is proven that the expression of the core GHMT factors in mouse infracted hearts reduces fibrosis and improves cardiac function *in vivo* [5]. It is also possible that other mechanisms like enhancement survival of cardiomyocytes, differentiation of activated cardiac progenitors into cardiomyocytes, blockade of the activation of cardiac fibroblasts are affected in a positive way by this core transcription factors in the heart myocardial infarction [5].

1.4. CANCER CELLS

Cancer cells are cells that grow and divide at a high rate, without any regulation from the host organism. Many factors can play an important role in the development of cancer cells, such as genetics, age, environment, immune system. Due to these factors, cells can lose their ability to mark and destroy damaged cells and lead to cancer cells. One of the most studied processes of the apoptosis of damaged cells is the functionality of the tumor suppressor tp53[16], in which this gene loses the functionality and does not mark for apoptosis.

Stem cells and cancer cells share some characteristic features such as the capacity to infinite division and self-renewal [17, 18]. Thus, common processes and genes may be activated in both types of cells. Indeed, studies have showed that both stem and cancer cells can share a significant number of common activated genes, which points to the execution of a common molecular program in both types of cells[17].

For most cancers, the transforming genetic mutations are still not exactly known. Nevertheless, some types of cancers appear to arise from mutations that accumulate in stem cells [16, 17]. In general, there are many apparent connections between stem cells and cancer that are important to understand [17]. In particular, understanding the process for the control of self-renewal of normal stem cells could give new insights into the origins of cancer.

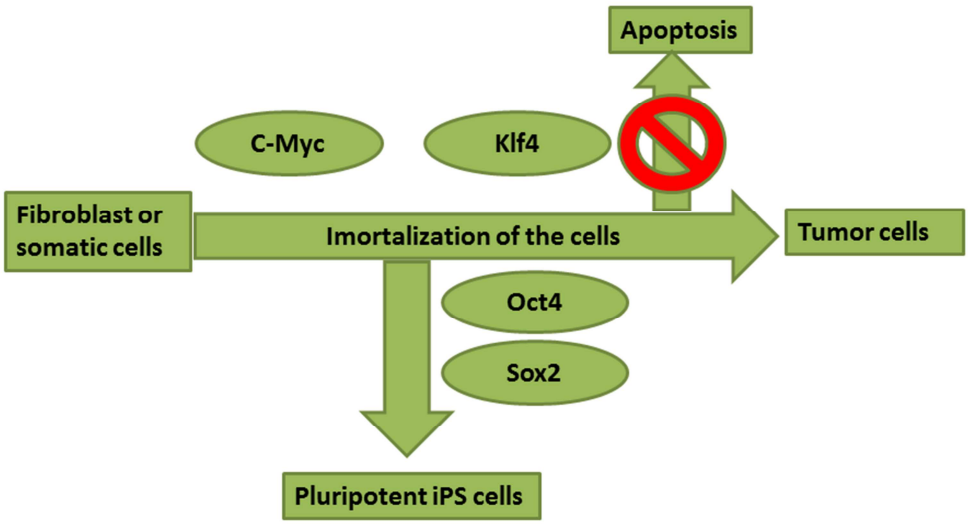


Figure 1.4.1. Possible roles for the four transcription factors in the generation of iPSc or tumor cells (adapted from Yamanaka, 2007).

2. METHODOLOGY

This chapter details the microarrays chosen for the gene expression data sets analysis and the methods used to filter and treat all the data sets gathered along this study. It contains a brief description of the data sets, bioinformatic tools used in this study and the network analysis procedures. Finally, it describes methods used for the comparative gene analysis between stem cells and cancer cells.

2.1. MICROARRAY TECHNOLOGIES & DATA SETS

Although, the DNA content is the same in every cell of an organism, tissues show an amazing diversity of functions due to the tissue-specific gene expression. To understand better this diversity, genome-wide measurements of expression have been made in many types of tissues from various organisms. Nowadays, the study of expression of a remarkably large number of genes is routinely performed using the microarrays technology.

Microarrays have therefore become a great tool to perform genome-wide measurements, being a crucial tool to create a comprehensive gene expression atlas for different organisms, to facilitate rapid identification of new marker genes for improved diagnosis and target genes. This kind of atlas would increase general understanding in gene expression and give new information about functions of genes.

Generally, microarrays are based on the hybridization of labelled transcripts to a complementary nucleotide sequences attached to a solid surface. There are two types of microarrays, two-channel and one-channel microarray. The two-channel microarrays, in two samples are co-hybridized on the same array and the one-channel microarray, where only a single sample is hybridized on a microarray.

In this work the main types of microarrays are one-channel microarrays from the companies Affymetrix and Illumina. These microarrays platforms can be reliable due to their accuracy and precision.

Figure 2.1.1. One-colour channel microarray platform, Affymetrix GeneChip [19] (on the left) and Illumina [20] (on the right).

Affymetrix produces an oligonucleotide microarray (also called GeneChips) that are manufactured using photolithography [19]. Photolithography is a process of using light to control the manufacture of multiple layers of material. For GeneChip production, Affymetrix uses photolithography masks that contains tiny holes designed to let light through for the sequence that is receiving the next nucleotide[19]. The sequences that are protected from light will not receive another nucleotide to the DNA strand and the ones that are not protected from the light will receive a nucleotide to add in the growing DNA chain[19]. Each mask is designed to add new nucleotides in different sequences and this process is repeated over and over again with a new mask until the desired sequences are completed [19].

Illumina microarray or BeadArrays contain of oligonucleotides immobilized in beads that are held in microwells on the array substrate. It is highly reproducible since it was a high level of bead type redundancy (in average 30 beads per probe) [20]. These beads are randomly distributed across the array and the unique sequences present in each bead are used for identifying the location of each bead [20]. Each probe location and sequence combination on Illumina bead chip is carefully selected bioinformatically [20]. Hybridization of the whole-gene expression assay offers the highest capability for simultaneously profile more than 47000 transcripts. To identify the unique sequences on the beads, Illumina uses sequencing by synthesis technology, which consists in sequencing tens of millions of clusters on the array [20]. During each sequencing cycle, a nucleotide is added to the nucleic acid chain. Nucleotide label serves as a terminator for polymerization, so after each inserted

nucleotide, the fluorescent dye is imaged to identify the base used. Inserted nucleotides are identified directly from signal intensity measurements during each cycle [20].

2.1.1. GENERAL GENETIC EXPRESSION IN MOUSE EMBRYONIC STEM CELLS

A dataset produced by Gaspar et al. 2012 [21], which profiled the RNA expression in the first 10 days of *in vitro* differentiation of mouse CGR8 ESc, was used in my study. CGR8 is a cell line that was established from the inner cell mass of a 3.5 day male pre-implantation embryo (*Mus musculus*, strain 129) [21]. These pluripotent cells retain their ability to participate in normal embryonic development. Supplementation of LIF (Leukaemia Inhibiting Factor) in the culture media allows culture of undifferentiated CGR8 ESc without the use of any feeder layers.

For this study, cRNA was prepared according to the standard Affymetrix protocol. From the 45,101 probe sets represented on the Mouse 430 version 2 array, expression data of 30,526 gene associated transcripts were analysed after eliminating transcripts without annotation and of unknown origin, as well as hypothetical transcripts or proteins. These expression measurements were performed in triplicates from biologic independent samples [21]. In total, the dataset comprises of 27 samples taken at different time points days of age (0, 1, 2, 3, 4, 5, 6, 7, 10 days).

The results of the microarray were deposited in the EBI ArrayExpress, data set E-TABM-672. The *CEL* files were downloaded from there and were submitted to background correction, normalization and summarization of gene expression (*rma*), through a package called *limma* implemented in programming language “R” (see below for a description of these methods).

2.1.2. ROSETTA 1

The data set Rosetta 1 was generated using Agilent spotted oligonucleotide microarrays and contains expression patterns for 10000 genes in 52 different types of tissue and cell lines [22].

For further analyses and to facilitate the comparison, the samples were assigned to 19 main tissue classes based on their physiology and histology [23]. The data was log transformed and subsequently normalized using the quantile normalization which is based on the transformation of the original value to the corresponding quantile value of a

reference chip [23].This data set was retrieved from BMC Genomics Site in the additional files, additional file 1 (http://www.biomedcentral.com/1471-2164/11/305/additional) [23].

2.1.3. ROSETTA2

Data set Rosetta2 was generated using Agilent spotted oligonucleotide microarrays and contained expression patterns for 50000 genes in 54 different types of tissue [24]. As in the Rosetta 1 data set, the data used was merged into 19 main tissue to facilitate comparison [23]. This data set was retrieved from BMC Genomics Site in the additional files tab, additional file 2 (http://www.biomedcentral.com/1471-2164/11/305/additional) [23].

2.1.4. STANFORD

In this study, Shyamsundar and co-workers (2005) measured gene expression in 115 tissue samples on a dual channel cDNA microarray containing 39,711 human cDNA, representing 26,260 different genes [25]. Tables with the pre-processed expression values for the analysed data were taken from (http://www.biomedcentral.com/1471-2164/11/305/additional) [23].

2.1.5. HUMAN GENE ATLAS

Su (2004) created a genome-wide expression profile of 79 human tissues using Affymetrix HG-U133A and customized GNF1H comprising 42.865 probe sets and expression summaries were obtained using the Affymetrix Microarray Suite 5 (MAS5) [26].

That data processed by MAS5 was downloaded from the BioGPS website (http://biogps.org/downloads/), the annotations for the GNF1H were also taken from the same website. Annotations for the Affymetrix HG-U133A were obtained through R/Bioconductor package (hgu133a.db). Probes that could not be mapped were excluded from the analysis.

2.1.6. MOUSE GENE ATLAS

Su (2004) also created a genome-wide expression for mouse, containing 61 types of tissue and a total of 153 samples. They used a custom design array with the whole genome expression, containing 36.182 transcripts, with the identification by Affymetrix as GNF1M array [26].

The used expression data set was downloaded from the BioGPS website (http://biogps.org/downloads/), the annotations for the GNF1M were also taken from the same website. Annotations for the Affymetrix GNF1M were obtain trough a package present in R, called org.Mm.eg.db (Genome wide annotation for Mouse), extracting from the package the annotation about gene id, symbol and description of the gene. Probes that could not be mapped were excluded from the analysis. These 61 physiologically normal tissues were obtained from adult (10-12 weeks) C57Bl/6 mice (4 male, 3 female) by dissection [26].

2.1.7. REPROGRAMMING MOUSE FIBROBLAST INTO FUNCTIONAL CARDIOMYOCITES BY DEFINED FACTORS

In a landmark study, Ieda and co-workers showed that it is possible to reprogram murine cardiac fibroblasts into cells with great similarities to cardiomyocytes [4].

In their study, they combined the exogenous expression of three fundamental heart development transcription factors (Gata4, Tbx5 and Mef2c) to transdifferentiate *in vitro* rapidly and efficiently post-natal cardiac fibroblast directly into differentiated cardiomyocytes-like cells. These so called iCM expressed cardiac-specific markers and have a global gene expression profile analogous to native cardiomyocytes. [4]. Notably, the authors also showed that the reprogramming occurs in a direct manner i.e. the fibroblasts are reprogrammed to iCMs without first becoming a stem or progenitor cells. Importantly, this induced transdifferentiation can also occur *in vivo*, as fibroblast cells showed that were transplanted into mouse hearts one day after transduction of the three factors and differentiated into iCMs [4].

These findings suggest that functional cardiomyocytes can be directly obtained from differentiated somatic cells by induction of defined factors [4]. Ieda and co-workers also performed genome wide expression analyses using Affymetrix Mouse Gene 1.0 ST Array. Besides profiling neo-natal murine cardiomyocytes and cardiac fibroblasts, the expression of successfully and non-successfully reprogrammed fibroblasts was measured. The success of reprogramming was measured using transgenic GFP under control of an aMHC promoter. GFP^+ and GFP^- were collect by fluorescent-activated cell sorting (FACS) after 2 and 4 weeks of culture [4]. Microarray analyses were performed in triplicate from independent biologic samples, according to standard Affymetrix Genechip protocol [4].

The results of the microarray were deposited in GEO, with the accession number GSE22292. For my study, the *CEL* files were downloaded from there and were submitted to background correction, normalization and summarization of gene expression (*rma*), through a package present in "R" called *affy*.

2.1.8. REPROGRAMMING NON-MYOCYTES WITH CARDIAC TRANSCRIPTION FACTORS

Song (2012) demonstrated that the combination of GHMT (Gata4, Hand2, Mef2c, Tbx-5) factors could reprogram cardiac fibroblast into functional cardiomyocytes-like cells *in vitro* and *in vivo*. First they could demonstrate that the chosen combination of transcription factors correlated with their ability to convert fibroblasts into so called cardiac-like induced myocytes which express markers for cardiomyocytes. They showed that exogenous GHMT expression in non-cardiomyocyte in the infarcted heart reduces fibrosis and improves cardiac function *in vivo* [5].

Notably, it appears that the efficiency of transdifferentiation is greater in vivo than in vitro, indicating that the native environment of the heart is a more permissive environment than the plastic tissue culture dishes for functional reprogramming [5]. In this study, the obtained RNA was isolated from the uninfected cardiac fibroblasts as well as, cardiac fibroblasts transduced for 2 or 4 weeks with either an empty vector or a GHMT retrovirus and cardiomyocytes from adult mice. Microarray analysis was performed on the platform of Illumina Mouse-6 Beadchip and analysed using GeneSpring GX software (Agilent) [5].

Data from this study was deposited in GEO with the accession number GSE37057. The *CEL* files were downloaded from there and were submitted to background correction, normalization and summarization of gene expression (*rma*), through a package present in "R" called *affy*.

2.1.9. HIPSC DIFFERENTIATION TOWARD CARDIOMYOCYTES

Besides direct conversion from other types of differentiated cells, differentiation of hESc/iPSc could be a promising cell source of cardiomyocytes in cardiac regenerative medicine [3]. Uosaki and colleagues differentiated hiPSc towards cardiomyocytes applying sequential administration of activin, bone morphogenetic protein 4 (BMP4), fibroblast growth factor 4 (FGF4) and Dickkopf 1 homolog (DKK1). For expression profiling, they used Affymetrix Human Gene 1.0 ST arrays. at day 0, 2, 5, 7, 9 and 11 during differentiation [3].

Data from this study was deposited in GEO with the accession number GSE28191**.** The *CEL* files were downloaded from there and were submitted to background correction, normalization and summarization of gene expression (*rma*), through a package present in "R" called *affy*. For my study, 33010 genes were included that were annotated.

2.2. GENE EXPRESSION ANALYSIS

In the next sub-chapters it will be addressed the main computational algorithms and approaches that I applied for the data analysis including, the comparison and clustering of temporal gene profiles. Most of the computational analysis was performed in R which is a programming environment with favourable numerical capability, flexible visualization and a wide range of statistical and mathematical algorithms. Notably, R is gaining a widespread usage within the computational biology and bioinformatics, and is the basis for the Bioconductor platforms which provides many add-on packages for microarray data analysis and toolset to work with [27].

Examples for the infrastructure concepts could be the *ExprsSet* class from the *Biobase* package. The fundamental object for gene expression analysis in R/Bioconductor is the *ExprsSet*, which is a data structure binding together array-based expression measurements with information about the samples [27]. Important Bioconductor package used for my work are *Affy* and *Limma,* which are outlined below [27].

2.2.1. PREPROCESSING OF MICROARRAY DATA

Before analysing expression changes of genes, the microarray data need to be pre-processed. This includes performing background correction, normalization, pm correction and summarization. Figure 2.2.1.1 presents an overview of these steps which were carried out using the Affy package in this study. For Affymetrix experiments, *CEL* files that contain probe intensities derived from the scan of the GeneChip. *CEL* data is a very simple structure, storing all probe intensities from the chip. Information about the probe identity, location of each probe and very limited sequence data is stored in the .cdf object generated by the program.

As there is always some amount of background noise in every scanned image, so this background correction function will check the distribution of probe intensities and this will be used to estimate overall background noise level and adjust it.

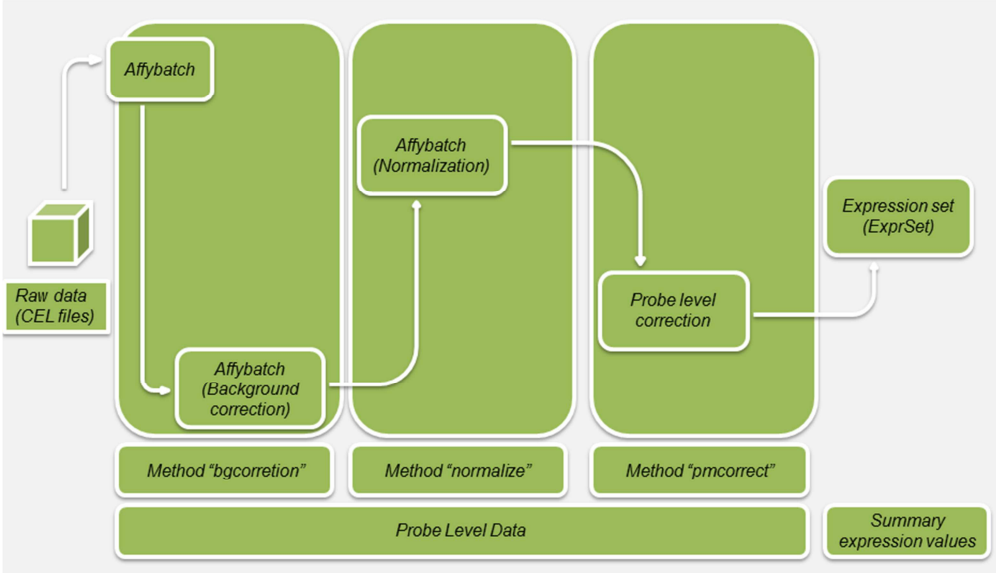


Figure 2.2.1.1. Simple steps from CEL files to expression set. It starts with an affybatch file, it is submitted to a background correction, normalization, pm correction and summarization and finally is turned in to an expression set (adapted from Gautier *et al*., 2003 [28]).

Normalization is required because no step in the hybridization process is perfectly controlled, so the quantity of RNA in a sample varies slightly from chip to chip. Normalization procedures attempt to detect and correct systematic differences between chips, so data from different chips can be directly compared [28].

Pm correction (perfect match correction) is for mismatch probes that exist in *Affymetrix GeneChips* to quantify non-specific and cross-hybridization. This mismatch probes are used to correct perfect match probes. As each transcript is targeted on the *GeneChip* by multiple probes, a final step is the summarization over different probes associated with a gene one or more probe sets.

2.2.2. DETECTION OF DIFFERENTIAL EXPRESSION

To detect differential expression, a linear model approach was applied using *the Limma* package. This method requires construction of two types of matrices: (i) the first one is the *design matrix* which associates the RNA samples with the array, and (ii) the *contrast matrix* that specifies which are the comparisons that the user would like to make between the RNA samples [29]. For statistical analysis and assessing differential expression, *limma* uses an empirical *Bayes* method to moderate the standard errors of the estimated fold changes [29].

2.2.3. CLUSTERING OF GENE EXPRESSION DATA

Cluster 3.0 is a program that provides a computational environment for analysing data from DNA microarray experiments and other genomic data sets. This program can only read tab-limited (.txt files) in a particular format. After data input, it is submitted to a set of filters that allows the treatment of the data, such as, removing all genes that have missing values in at least one condition of the experiment.

Posteriorly it was performed a hierarchical clustering, which organizes genes in a hierarchical structure based on their expression pattern. It was used *Single Linkage Clustering,* in which the distances between two items x and y is minimum of all pair wise distances between items contained in x and y. This kind of hierarchical clustering can be used to cluster large sets of gene expression [30].

Output files are opened with *Java treeview*, a program that can provide a graphical environment for analysing data from microarrays experiments and other types of data sets [31]. This program allows an interactive graphical analysis of the results from the *Cluster 3.0,* reading the output files.

2.2.4. FUNCTIONAL ENRICHMENT ANALYSIS

To detect whether differential genes tend to be associated with a particular processes, functional enrichment analysis was performed. Here, the statistical significance is calculated based on the observed number of genes associated with the same process among the total set of differentially expressed genes. Moreover, biological processes, molecular function and cellular components can be tested based on the annotation in Gene Ontology (GO), which organizes genes into categories for a number of different organisms [32]. The statistical evaluation was performed using R as well as using the *FatiGo* tool from the *Babelomics* (http://babelomics.bioinfo.cipf.es/) website [33].

FatiGO is a simple and powerful tool to extract relevant GO terms for a list of genes with respect to a set of genes of reference (usually the rest of the genome). This program uses the most widely accepted ontologies in GO, which organizes information for molecular function, biological processes and cellular components for a number of different organisms [32]. *FatiGO* program returns 4 main columns of interest: GO, that represents what are the biological processes, list of gene input, so it compares to the all genes that are present in the

biological functions; list of positive hits, which means the number of genes that participate in the biological functions; adjusted *p-value*, is ordered by increasing value, facilitating the selection of GO terms with the most significance for the experience [32].The *FatiGO* program present in *Babelomics* website was used for the studies presented in sub-chapters 3.1 and 3.3.3.For the comparison of biological processes, molecular functions and KEGG pathways between cardiac data sets and mESc (sub-chapter 3.4), all 3 gene ontology processes were analysed through R, which facilitates comparison and systematic inspection of the results.

2.3. HEART EXPRESSION

This study was made to understand which genes could be linked between the different data set analysed, so it was compared several genes associated to heart development and morphogenesis, after the choice of certain parameters for each stage of development (early, intermediate and late) present in table 2.3.1. It was not possible to choose from Song data set [5] and Ieda data set [4] the early and intermediate stage, since this two data sets do not have time course gene expression.

Table 2.3.1. Heart expression analysis parameters. (Description of parameters: Song *et al*., 2012 – ACFGHMT – Adult Cardiac Fibroblast transdifferentiated with transcription factors Gata4, Hand2, Mef2c and Tbx5; Ieda *et al*., 2010 – 2WiCM – 2Weeks induced Cardiomyocytes, 4WiCM – 4Weeks induced Cardiomyocytes.)

Authors	Abbreviation	Parameter	Stage
Song *et al*.,2012	CF into iCM	ACFGHMT	Late
Ieda *et al.,* 2010	Rep into iCM	2WiCM, 4WiCM	Late
Hideki *et al*., 2011	iPSc into iCM	Day 0	Early
		Day 2, Day5	Intermediate
		Day 7, Day 9	Late
Gaspar *et al*., 2012	mESCDiff	Day 0, Day 1	Early
		Day 3, Day 4	Intermediate
		Day 6, Day 7	Late

Genes of these data sets have been compared between themselves for the early, intermediate and late stage. The log fold change chosen for this analysis has a cut-off of 0.5 log fold change expression (these values were obtained through “limma”).

For comparison, all genes were mapped with mouse EntrezGene Ids. It was recorded which genes are present in each data set and which genes are overlapping in the different combinations of these data sets.

It was created a web based program (HeartExpress: http://heartexpress.sysbiolab.eu/) [34] that provides researchers a direct access to the collected data and enables independent investigations. Heart Express allows to explore a genome-wide expression data for heart development and morphogenesis.

2.4. NETWORK ANALYSIS

For the analysis of networks, the Cytoscape software was utilized [35]. This software can generally be used for analysing protein-protein, protein-DNA and genetic interactions that are increasingly available for different model organisms [36]. Cytoscape has the purpose of modelling and integrate biomolecular interaction networks and states [36]. Cytoscape core handles basic features such as network layout and mapping of data attributes to visual displays properties. Dynamic states on molecules and molecular interactions are handled as attributes on node and edges, whereas static hierarchical data, such as, protein functional ontologies are supported by use of annotations [36]. Besides the graphical image presented by Cytoscape, there are other functionalities that allow the user to modulate the network (figure 2.5.1.).

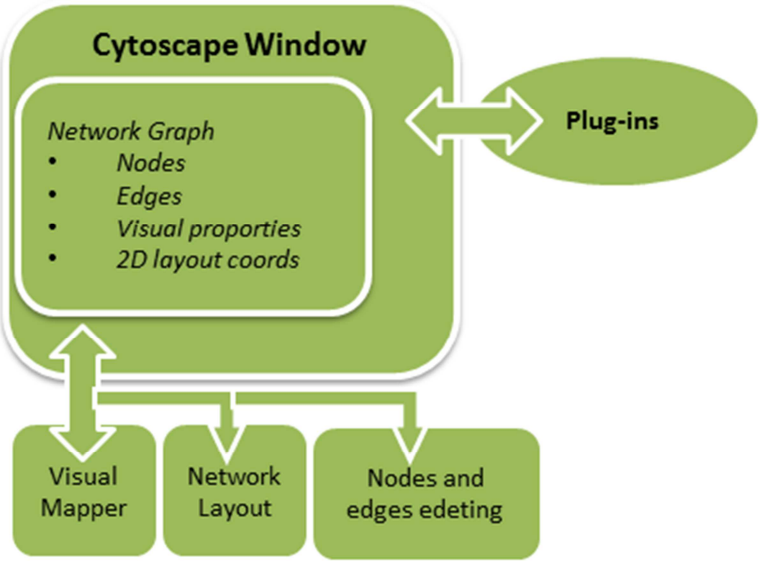


Figure 2.4.1. A simplified schematic overview of the Cytoscape functionality. Besides the Cytoscape main window, which gives graphical presentation of interactions of the genes, the other functionalities allow the user to modulate the results, according to what is needed. (Schematic adapted from Shannon *et al*., 2012 [36]).

This tool was used to visualize the sets of the identified genes with early, intermediate and late expression during cardiomyogenesis within a network context and understand their behaviour. First, a large molecular interaction network was generated. This large network was examined for smaller sub-networks that could be of interest for this purpose of the study. This approach helped to identify other key players that have not been considered to be involved in heart development to date.

2.5. GENE COMPARATIVE ANALYSIS WITH STEM AND CANCER CELLS

To examine the link between cancer and stem cells, several relevant gene sets were gathered: a list of genes connected to pluripotency (*Oct4 screen hits [37]*) and two lists connected to cancer cells (Cancer Gene Census [38] & Genetic Association Database [39]).

2.5.1. GENE LIST FOR CANCER AND PLURIPOTENCY

Oct4 screen hits [37] data set has originated from a genome-wide RNA interference screen to identify genes which regulate self-renewal and pluripotency properties in hESc [37]. A genome wide screen was performed and genes targets identified which are associated with a reduced Oct4 expression. In total, 566 genes were detected.

Cancer Gene Census [38] data set results from an effort by Sanger Institute to catalogue all genes for which mutations have been causally implicated in cancer. Number of genes present in the used data set: 487 genes. (http://www.sanger.ac.uk/genetics /CGP/Census/).

Genetic Association Database [39] is an archive of human genetic association studies of complex diseases and disorders. The database provides summary data extracted from published papers in peer reviewed journals on candidate gene and GWAS studies. The goal of this database is to allow the user to rapidly identify polymorphism from the large volume of polymorphism and mutational data. Number of genes present in the used data set: 3214 genes. (http://geneticassociationdb.nih.gov/).

2.5.2. ASSESSING SIGNIFICANCE OF COMMON GENES

The statistical significance of the detected overlap between the different lists, the hypergeometric test was applied. The calculated significance indicates whether more common genes were found in two lists than expected by chance. The hypergeometric test is equivalent to Fisher's test. Significance (p-value) of enrichment in common genes was calculated for each potential combination. For visualization of the number of genes present in the overlap, Venn diagrams were produced.

3. RESULTS

Several microarrays data sets were analysed and compared in various ways to obtain a comprehensive view of transcriptional changes during cardiomyogenesis. Various types of approaches were performed, since there were many types of microarrays, which range from temporal gene expression to gene expression in specific types of tissues, such as, in cardiac fibroblasts and induced cardiomyocytes. First, I examined temporal gene expression data for *in vitro* differentiation of mouse ESc, although this differentiation method is not specific for cardiomyogenesis and generates cells derived from the three germ layers, the analysis gave important cues about the temporal order of expression during stem cell differentiation. Furthermore, in the molecular mechanisms involved in ESc differentiation towards cardiac cell lineages are comparable to the developmental mechanisms.

Subsequently, I analysed microarray data sets which are closely associated with cardiomyogenesis. The results obtained from the four microarray data sets (Table 2.3.1.), were used to get biological processes, molecular functions and KEGG pathways and to compare the overlapping gene expression in order to derive a more accurate and reliable sets of genes that were expressed in an early, intermediate and late stage of cardiac cell lineage differentiation. This last method helps to filter from the significant genes linked to heart development that were previously described in the literature. Also, network-based analyses were carried out to interpret the observed gene expression changes. Finally, to assess if any genes associated with heart development could be linked to cancer stem cells or stem cells expression, a brief comparison of the genes included in different gene sets conclude the analyses carried out within this study. To compile all this gathered and finely tuned information, it was developed a website that contains data, where an independent user can check his genes of interest.

3.1. GENERAL GENETIC EXPRESSION IN MOUSE EMBRYONIC STEM CELLS

Differentiation of stem cells is a complex process with a defined temporal order. To elucidate the underlying dynamics, I analysed a microarray data set, which captured expression changes in murine stem cells during differentiation induced by the hanging drop method [21]. Although this approach generates not only cardiomyocytes, but rather a mixture of differentiated cell types representative of the three germ cell layers, it provides valuable insights in the temporal order of the expression of genes. To gain an overview of the expression patterns as well as to assess the quality of microarray measurements, it was checked the gene expression day by day, searching for the genes that are more expressed at each time point. This type of analysis was made to determine (specific) gene expression variation through the differentiation time points.

A dendrogram cluster of gene expression was made and the genes associated in three different groups. Notably, no outlier was detected indicating a high reproducibility of the measurements. Moreover, day 0, 1, 2; day 3, 4, 5; day 6, 7, 10 are clustering together, showing that gene expression between them are similar and clearly dividing themselves in early, intermediate and late stage. The most expressed genes in day 0 to day 2 are genes related to the undifferentiated cell state; from day 3 to day 5 expressed genes are linked to cell fate commitment and anatomical structure formation involved in heart morphogenesis and from day 6 to day 10 the most expressed genes are linked to organ and tissue development.

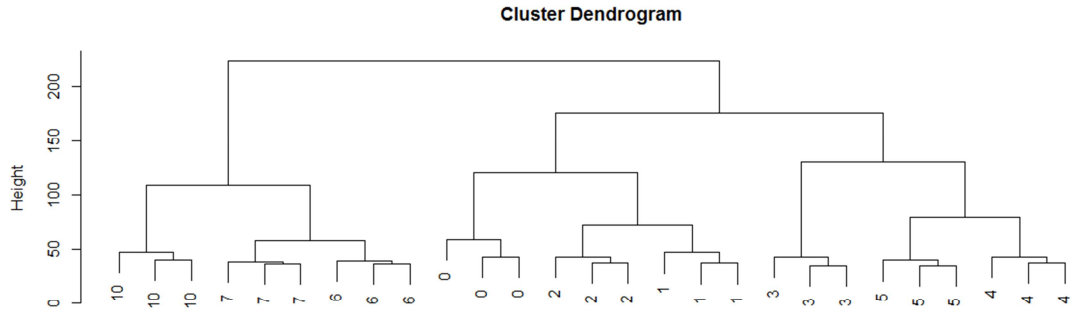


Figure 3.1.1. Cluster dendrogram for Gaspar *et* al. 2012 data set [21]. It is possible to observe that the respective days are clustering together and dividing themselves according to an early/intermediate/late stage.

To get the most up-regulated genes in each day, it was obtained “topTables” in R, trough the “limma” package. These “toptables” contains all genes that have a defined significant adjusted p-value, that in this case is <0.1, and ease the visualization of the genes that are most expressed in each day of differentiation. Amongst these genes, I selected those with a strong positive fold change and a significant adjusted p-value to ensure the specificity of expression of these genes to the specific time points. Subsequently, a functional analysis through Babelomics 4.3 (http://babelomics.bioinfo.cipf.es) was performed to determine the biological processes in which the selected genes participate at the relevant time points [33].

In the following section, I will present the expression profiles along the differentiation process of genes important ESc undifferentiated state, including Nanog, Sox2 and Pou5f1 and genes involved in differentiation into cardiac cells, for instance, Gata4, Tbx5, and Mef2c. The expression profiles of genes involved in the transition from pluripotent stem cells to mesoderm and endoderm, such as, T-brachyury and Eomes are also presented.

DAY 0

Genes up-regulated at day 0, in comparison to the rest of the time points, are associated with maintenance and proliferation of undifferentiated cells, such as Nanog and Sox2 and show a strong expression (figure 3.1.2 and 3.1.3). Dnmt3l is a *DNA Methyltransferase* that methylates DNA. It can cause epigenetic modifications and mediate transcriptional repression through interaction with histone deacetylase1 [40].

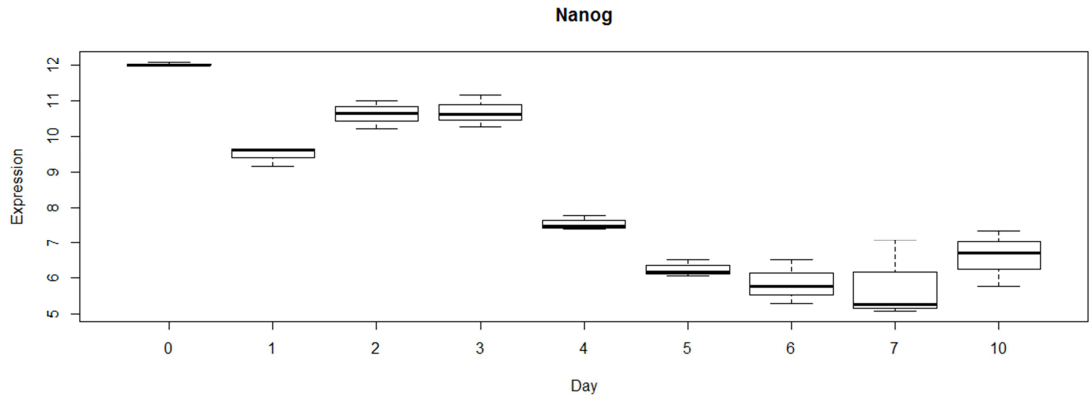


Figure 3.1.2. Expression of Nanog homeobox on Gaspar *et* al. 2012 data set.

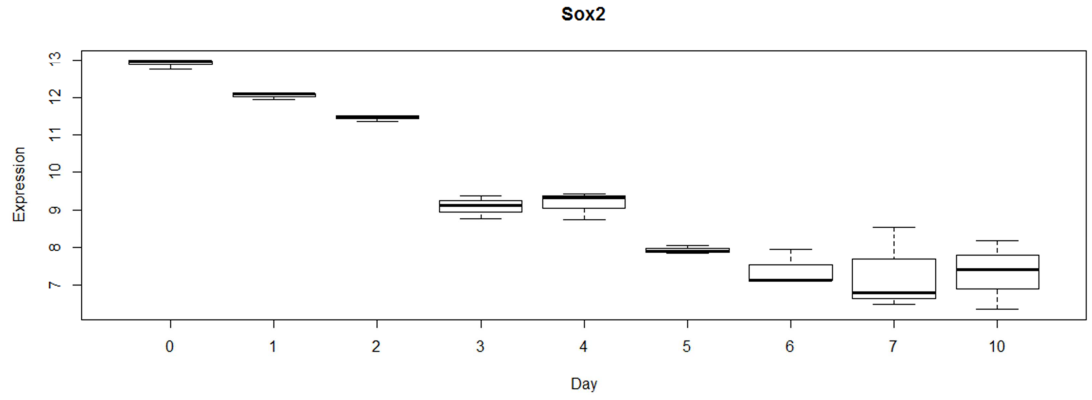


Figure 3.1.3. Expression of SRY-box containing gene on Gaspar *et* al. 2012 data set.

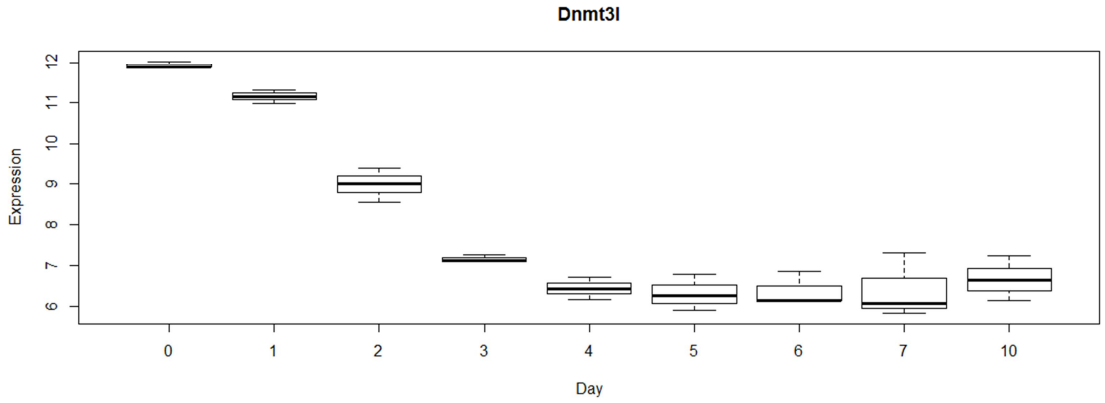


Figure 3.1.4. Expression of DNA (cytosine-5)-methyltransferase 3-like on Gaspar *et* al. 2012 data set.

In table 3.1.1. the main genes, that are overexpressed, are linked to stem cell maintenance and development, negative regulation of differentiation, regulation of cell differentiation and meiosis. All these processes together are promoting stem cell maintenance and proliferation, while inhibiting cell differentiation.

Table 3.1.1. Biological functional analysis in Day 0 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0019827	stem cell maintenance	253	9	3.56	3.85E-09
GO:0048864	stem cell development	253	9	3.56	3.85E-09
GO:0048863	stem cell differentiation	253	9	3.56	3.29E-08
GO:0045596	negative regulation of cell differentiation	253	15	5.93	4.74E-06
GO:0007283	spermatogenesis	253	15	5.93	1.38E-04
GO:0032526	response to retinoic acid	253	6	2.37	1.42E-03
GO:0048468	cell development	253	26	10.28	2.00E-03
GO:0045595	regulation of cell differentiation	253	17	6.72	3.27E-03
GO:0033189	response to vitamin A	253	6	2.37	3.27E-03
GO:0007548	sex differentiation	253	9	3.56	3.53E-03
GO:0006520	cellular amino acid metabolic process	253	12	4.74	3.53E-03
GO:0033273	response to vitamin	253	7	2.77	3.64E-03
GO:0007127	meiosis I	253	5	1.98	4.30E-03
GO:0009880	embryonic pattern specification	253	5	1.98	4.30E-03
GO:0050793	regulation of developmental process	253	24	9.49	4.43E-03
GO:0045165	cell fate commitment	253	9	3.56	4.43E-03

DAY 1

Pou5f1 has a higher peak of expression at day 1, and that could mean that its overexpression could be responsible for the initiation of cell differentiation and regulation of development process [41], while is also involved in stem cell maintenance [13].

Dppa2 is also significantly over-expressed in pluripotent cells. Knock down of this gene induces differentiation of stem cells causing a slight down regulation of Sox2 and Pou5f1 expression [42]. Dppa2 expression decreased prior to Pou5f1 at day 1 and day 2, indicating that Dppa2 might regulate the expression of master transcription factor Pou5f1 [42]. Thus, Dppa2 could also be a good marker for pluripotency and the regulation of ESc pluripotent state [42].

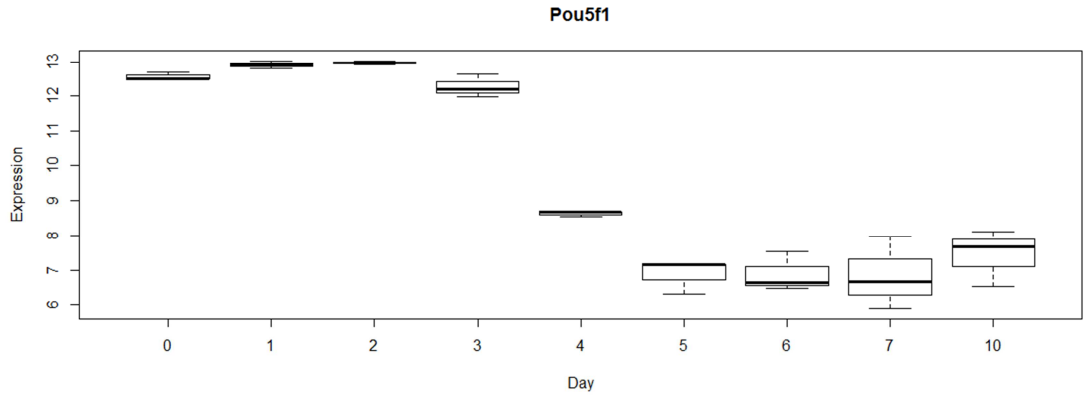


Figure 3.1.5. Expression of POU domain, class 5, transcription factor 1 on Gaspar *et* al. 2012 data set.

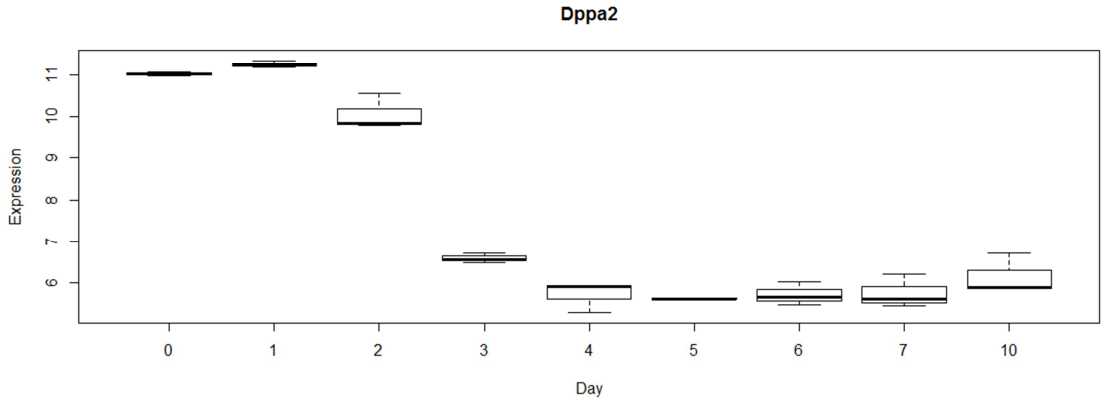


Figure 3.1.6. Expression of developmental pluripotency associated 2 on Gaspar *et* al. 2012 data set.

Although most of the positively differentially expressed genes are linked to pluripotency and negative regulation of cell differentiation, a number of genes related to development have their expression increased at this stage, suggesting that ESc are moving towards cell fate commitment (last biological process present in the table 3.1.1. and 3.1.2.).

Table 3.1.2. Biological functional analysis in Day 1 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positive	%Perc	p-value
GO:0007399	nervous system development	138	21	15.22	4.37E-04
GO:0048468	cell development	138	20	14.49	4.47E-04
GO:0007283	spermatogenesis	138	11	7.97	5.16E-04
GO:0050793	regulation of developmental process	138	19	13.77	6.20E-04
GO:0022008	neurogenesis	138	15	10.87	1.27E-03
GO:0009887	organ morphogenesis	138	17	12.32	1.27E-03
GO:0048598	embryonic morphogenesis	138	11	7.97	1.53E-03
GO:0045814	negative regulation of gene expression, epigenetic	138	4	2.9	1.76E-03
GO:0016458	gene silencing	138	5	3.62	2.29E-03
GO:0006366	transcription from RNA polymerase II promoter	138	15	10.87	2.73E-03
GO:0007420	brain development	138	10	7.25	2.73E-03
GO:0048839	inner ear development	138	6	4.35	2.73E-03
GO:0060052	neurofilament cytoskeleton organization	138	3	2.17	2.73E-03
GO:0045595	regulation of cell differentiation	138	12	8.7	3.30E-03
GO:0010628	positive regulation of gene expression	138	13	9.42	3.50E-03
GO:0009952	anterior/posterior pattern specification	138	7	5.07	3.50E-03
GO:0045165	cell fate commitment	138	7	5.07	3.50E-03

DAY 2

At day 2, the commitment to cells of the nervous system is initiated, as verified by the increased expression of Nefl, a gene responsible for nervous system development which is the most abundant cytoskeletal component of neurons [43]. All main processes in the neural commitment occur during this day, and the expression of genes involved in central nervous system to forebrain development is detected (table 3.1.3).

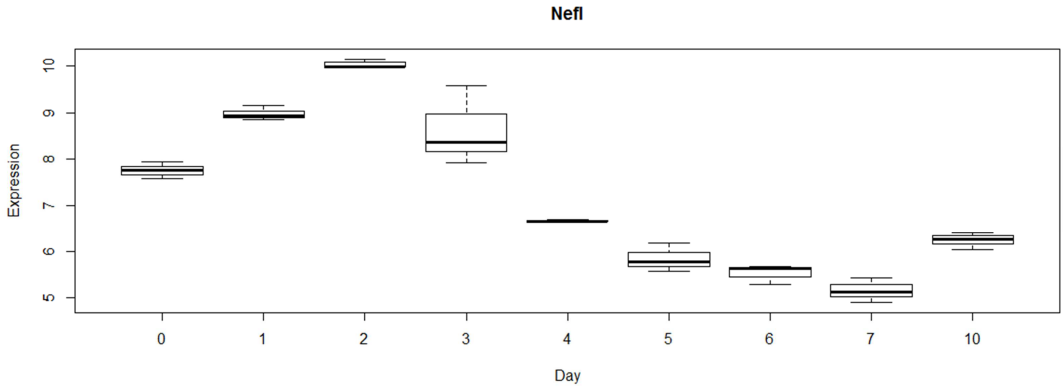


Figure 3.1.7. Expression of neurofilament (light polypeptide) on Gaspar *et* al. 2012 data set.

Genes involved in the general nervous system development are more expressed at day 3 in comparison to day 2 (table 3.1.3 and table 3.1.4.). Comparing day 2 with day 3, it is clear that the processes in the top 5 in day 2 are linked to nervous system development, but they are no longer present under the top ranked of day 3, although they still have a higher significance. Processes linked to pluripotent stem cell maintenance are also starting to fade and are not present anymore in day 3.

Table 3.1.3. Biological functional analysis in Day 2 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positive	%Perc	Adj p-value
GO:0007399	nervous system development	89	20	22.47	8.13E-07
GO:0048468	cell development	89	17	19.1	5.46E-05
GO:0030900	forebrain development	89	8	8.99	1.73E-04
GO:0007417	central nervous system development	89	11	12.36	2.06E-04
GO:0048598	embryonic morphogenesis	89	10	11.24	2.09E-04
GO:0048729	tissue morphogenesis	89	8	8.99	2.96E-04
GO:0009887	organ morphogenesis	89	14	15.73	3.25E-04
GO:0019827	stem cell maintenance	89	4	4.49	4.29E-04
GO:0048864	stem cell development	89	4	4.49	4.63E-04
GO:0007420	brain development	89	9	10.11	4.78E-04
GO:0009888	tissue development	89	13	14.61	4.83E-04
GO:0007492	endoderm development	89	4	4.49	5.02E-04
GO:0060052	neurofilament cytoskeleton organization	89	3	3.37	6.32E-04
GO:0050793	regulation of developmental process	89	14	15.73	8.02E-04
GO:0048863	stem cell differentiation	89	4	4.49	8.34E-04

DAY 3

T Brachyury (also known as T) is considered to be one of the most important early markers for early mesendoderm and is widely used in the field of developmental biology to track meseodermal germ layer development [21, 44]. Brachyury is initially expressed in all mesodermal cells and its expression declines when cells suffer patterning and specification processes to give rise to more specified mesodermal-derived tissues such as skeletal muscle, cardiac muscle and connective tissue [44]. The expression profile and values we observed (figure 3.1.9), show a significantly high expression around day 3. T Brachyury also participates, in interaction with other genes, on the neural tube development and segmentation [44].

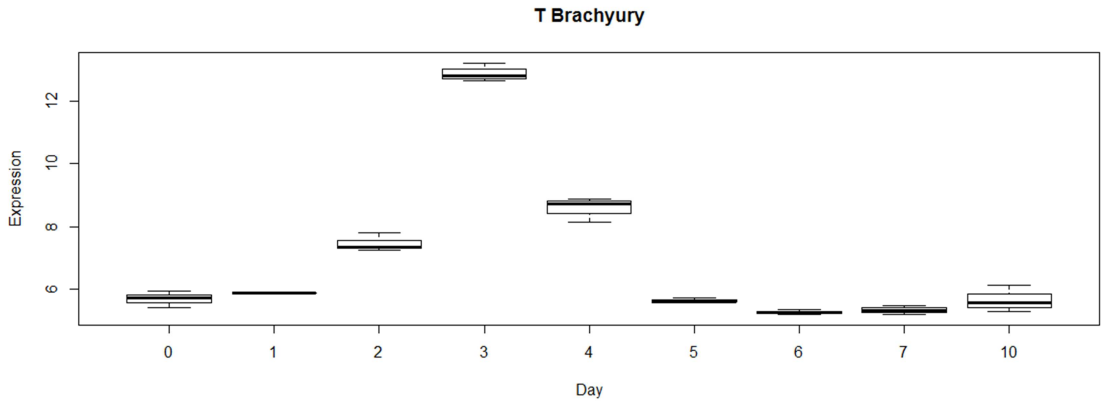


Figure 3.1.8. Expression of T Brachyury on Gaspar *et* al. 2012 data set.

Similar to T Brachyury, the expression of Eomes is transient and peaks at day3 (figure 3.1.10). Eomes is essential for mesoderm formation and defines a conserved molecular pathway controlling the gem layer formation and also controlling gastrulation and trophoblast differentiation in mammals [45]. This gene is precociously detected in trophoblast cell lineage, starting in the trophoectoderm. Its expression starts in the posterior part of the epiblast and later extends distally into the primitive streak and nascent mesoderm [45].

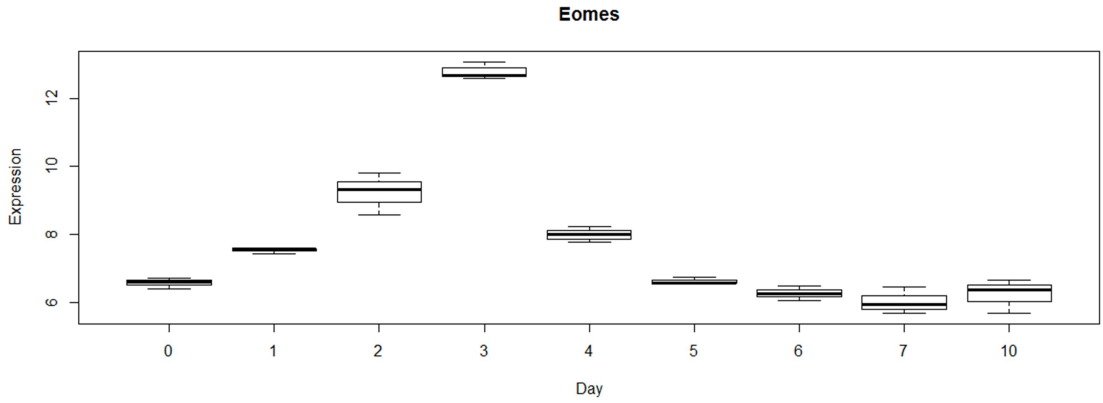


Figure 3.1.9. Expression of eomesodermin homolog on Gaspar *et* al. 2012 data set.

Thus, the expression of T Brachyury and Eomes peak at day 3 and these genes are the key regulators for embryonic morphogenesis and development, gastrulation, tissue development and anatomical structure formation involved in morphogenesis, but essentially for mesoderm development. The expression of these two genes presents a peak, because they are most active during tissue morphogenesis and are not involved in other types of processes.

Table 3.1.4. Biological functional analysis in Day 3 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0007389	pattern specification process	100	29	29	5.78E-29
GO:0048598	embryonic morphogenesis	100	28	28	1.88E-25
GO:0009790	embryo development	100	33	33	2.02E-24
GO:0007369	gastrulation	100	16	16	1.21E-20
GO:0009888	tissue development	100	31	31	1.21E-20
GO:0009887	organ morphogenesis	100	31	31	9.00E-20
GO:0009952	anterior/posterior pattern specification	100	18	18	5.16E-19
GO:0007399	nervous system development	100	32	32	3.66E-18
GO:0007498	mesoderm development	100	13	13	1.60E-17
GO:0048646	anatomical structure formation involved in morphogenesis	100	22	22	3.01E-17
GO:0048729	tissue morphogenesis	100	18	18	4.27E-17
GO:0001707	mesoderm formation	100	11	11	7.60E-17
GO:0048332	mesoderm morphogenesis	100	11	11	1.13E-16
GO:0007417	central nervous system development	100	21	21	3.38E-15
GO:0048468	cell development	100	28	28	5.79E-15
GO:0045165	cell fate commitment	100	15	15	8.20E-15

DAY 4

Hand1 (figure 3.1.11) is involved in early heart differentiation and plays an essential, but poorly understood role in cardiac morphogenesis [46], and is required for a normal progression of cardiac and extraembryonic development [47]. Risebro (2006) described Hand1 as an important cardiac transcription factor in regulating cardiomyocytes exit from cell cycle, to control the balance between cell proliferation and differentiation. Its high expression originates a pool of undifferentiated cardiomyocyte precursors [46].

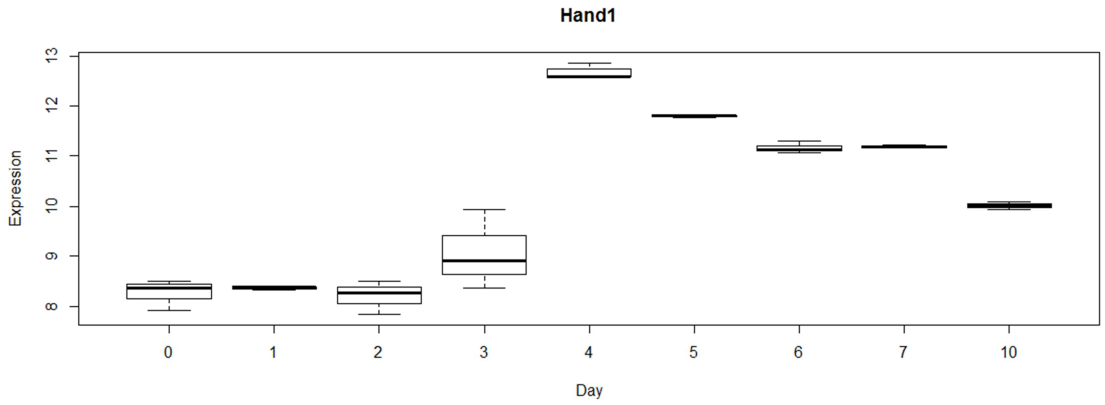


Figure 3.1.10. Expression of heart and neural crest derivatives expressed transcript on Gaspar *et* al. 2012 data set.

In table 3.1.5. it is possible to observe that most of the biological processes related to heart development and angiogenesis are starting to appear significantly enriched in over-expressed genes, indicating that genes related to heart development start to be expressed from day 4. By day 4 biological processes linked to pluripotency do not appear any more among the most significant biological functions.

Table 3.1.5. Biological functional analysis in Day 4 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0009887	organ morphogenesis	84	30	35.71	6.38E-20
GO:0007389	pattern specification process	84	18	21.43	6.84E-15
GO:0007507	heart development	84	17	20.24	6.84E-15
GO:0009790	embryo development	84	23	27.38	2.22E-14
GO:0003007	heart morphogenesis	84	11	13.1	1.72E-12
GO:0009888	tissue development	84	21	25	2.19E-11
GO:0035295	tube development	84	14	16.67	2.94E-11
GO:0001568	blood vessel development	84	15	17.86	2.94E-11
GO:0001944	vasculature development	84	15	17.86	3.37E-11
GO:0048646	anatomical structure formation involved in morphogenesis	84	16	19.05	8.02E-11
GO:0048598	embryonic morphogenesis	84	15	17.86	2.13E-10
GO:0008283	cell proliferation	84	21	25	2.95E-10

DAY 5

Hand2 is an early cardiac marker and is required in the formation of the heart, and is specifically involved in the development of the secondary heart field [14, 46, 47]. Smyd1 activity is necessary for Hand2 expression in cardiac precursors and Smyd1 is a direct target of Mef2c, which plays a role in myogenesis, suggesting that these transcription factors regulate development of ventricular cardiomyocytes [14] and have been shown to regulated transcription factors by mediating distinct chromatin modifications [48].

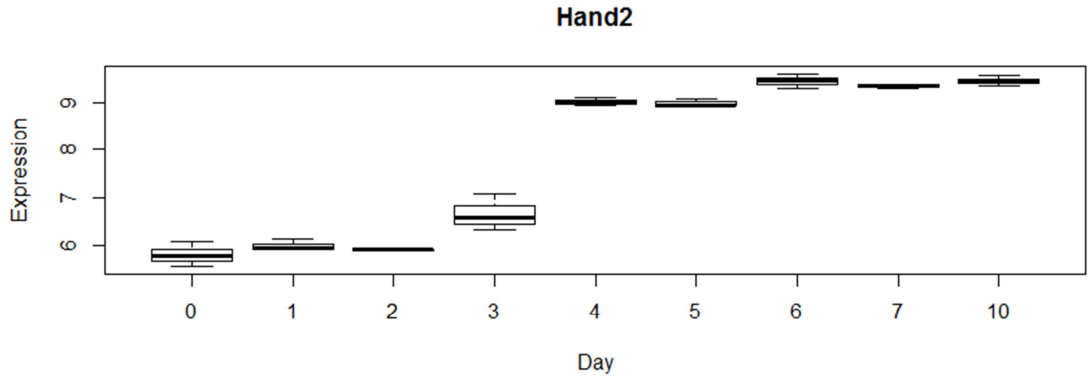


Figure 3.1.11. Expression Heart- and neural crest derivatives-expressed protein 2 on Gaspar *et* al. 2012 data set.

Once again, genes involved in heart development process still appear at day 5 of ESCs differentiation (table 3.1.6.), as expected, but not in the top 10 biological functions. This could mean that the genes linked to heart development are not displaying a strong expression, but the process is still active.

Table 3.1.6. Biological functional analysis in Day 5 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0048856	anatomical structure development	118	58	49.15	2.08E-31
GO:0048513	organ development	118	51	43.22	1.31E-29
GO:0009653	anatomical structure morphogenesis	118	41	34.75	4.44E-27
GO:0009887	organ morphogenesis	118	35	29.66	5.78E-27
GO:0009888	tissue development	118	29	24.58	1.49E-20
GO:0001501	skeletal system development	118	23	19.49	8.86E-20
GO:0030154	cell differentiation	118	40	33.9	1.84E-19
GO:0009790	embryo development	118	26	22.03	4.83E-18
GO:0048523	negative regulation of cellular process	118	34	28.81	3.71E-16
GO:0001944	vasculature development	118	19	16.1	5.18E-16
GO:0007507	heart development	118	17	14.41	8.32E-16
GO:0048646	anatomical structure formation involved in morphogenesis	118	18	15.25	4.85E-14
GO:0030334	regulation of cell migration	118	13	11.02	1.01E-12
GO:0003007	heart morphogenesis	118	11	9.32	1.59E-12

DAY 6

Ccbe1 and Cited2 have a peak of expression at day 6 of ESCs differentiation and their expression has been highly correlated to general organ development and anatomical structure morphogenesis. Also, some studies have indicated that they are linked to heart development and morphogenesis, and quite intimately linked to cardiac progenitors [49, 50].

Facucho-Oliveira (2011) confirmed that this gene is expressed in early cardiac progenitors that emerge from the primitive streak. Ccbe1 is expressed in the cardiogenic mesoderm and in the cardiogenic plate [49]. The expression of Ccbe1 in the cardiac mesoderm and absence in later stages of heart development suggests that its expression in only present in multipotent and proliferative cardiac progenitors and down regulated upon commitment into a more specific cardiac cell type [49]. However, during ESCs differentiation Ccbe1 expression is maintained at high levels until day 10, suggesting that cardiac progenitors persist in time or that Ccbe1 is expressed in non-cardiac cell types.

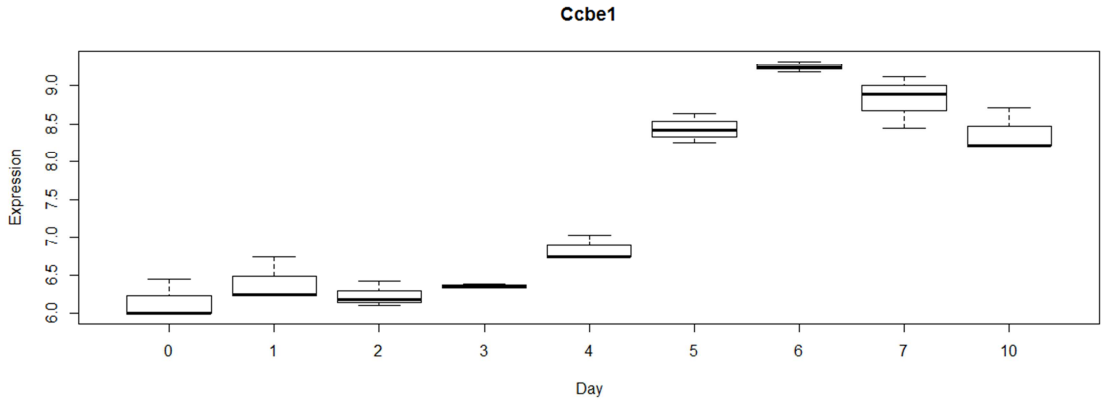


Figure 3.1.12. Expression of collagen and calcium binding EGF domains 1 on Gaspar *et* al. 2012 data set.

Cited2 is essential for the embryonic development of many different structures, such as, adrenal, neural crest, liver, lung, lens and placenta [50]. Macdonald (2012) showed that Cited2 also played a cell autonomous role in mouse cardiac myocytes, being required for myocardial thickening and coronary vessels development [50]. Although, Cited2 expression peaks at day 6, its expression is already high in undifferentiated cells (day 0) and decreases at days 1-2 before increasing again at day 3. Interestingly, Cited2 is required to ESCs self-renewal by controlling the expression key genes involved in the maintenance of pluripotency and for the optimal differentiation of ESCs towards cardiac cell lineages (José Bragança, personal communication). The results presented in Figure 3.1.14 suggest that the expression

of Cited2 is down-regulated at the beginning of ESCs differentiation to allow the differentiation process, but its expression rises before the initiation of the cardiac differentiation.

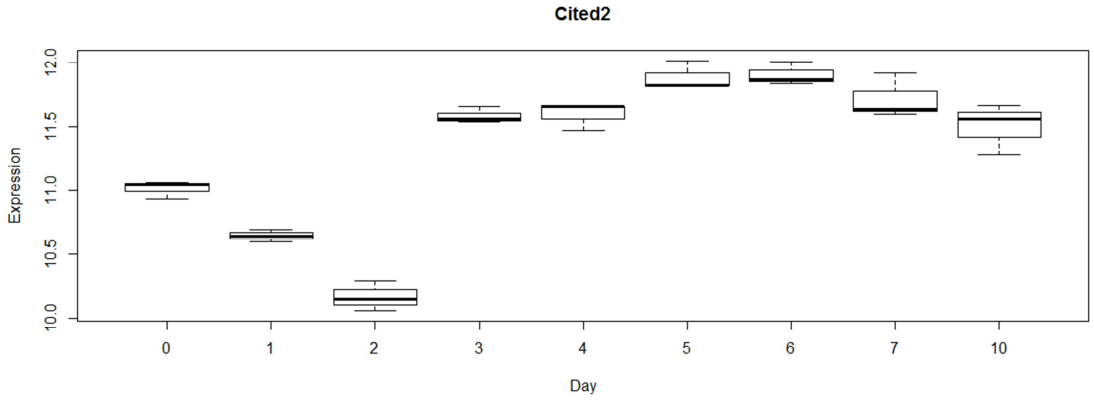


Figure 3.1.13. Expression of Cbp/p300-interacting transactivator 2 on Gaspar *et* al. 2012 data set.

It is still possible to see processes such as heart development and morphogenesis in day 6 (table 3.1.7). This day shows an increasing number of genes that participate in heart development and morphogenesis.

Table 3.1.7. Biological functional analysis in Day 6 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0048513	organ development	228	89	39.04	1.28E-47
GO:0009653	anatomical structure morphogenesis	228	67	29.39	5.88E-39
GO:0009887	organ morphogenesis	228	51	22.37	1.14E-32
GO:0009888	tissue development	228	48	21.05	4.02E-31
GO:0007155	cell adhesion	228	47	20.61	5.91E-27
GO:0001944	vasculature development	228	32	14.04	1.86E-25
GO:0030154	cell differentiation	228	63	27.63	4.34E-25
GO:0001568	blood vessel development	228	31	13.6	1.73E-24
GO:0007507	skeletal system development	228	28	12.28	1.87E-24
GO:0048646	heart development	228	29	12.72	1.08E-20
GO:0048468	extracellular matrix organization	228	41	17.98	2.72E-20
GO:0007517	anatomical structure formation involved in morphogenesis	228	24	10.53	1.74E-18
GO:0048514	muscle organ development	228	24	10.53	6.02E-18
GO:0008015	blood vessel morphogenesis	228	19	8.33	1.13E-14
GO:0001525	cell development	228	19	8.33	1.13E-14
GO:0048738	heart morphogenesis	228	14	6.14	1.13E-14

DAY 7

Most of the highly expressed genes at day 7 are markers of mature cardiac cells, such as, Tnnt2 (figure 3.1.14), Actc1 (figure 3.1.15) and Myh7 (figure 3.1.16). These genes indicate that the heart-cells differentiation process is coming to an end, since in day 10 (table 3.1.9) the relevant biologic functions are no longer among most significant enriched in over-expressed genes. .

These biomarkers are very important, because they help us to track cells of interest , i.e. in our case cardiac cells, differentiated [3, 51] or transdifferentiated [4, 52, 53], and to examine whether these cells present a similar expression profile compared to normal cardiac cells [4, 5].

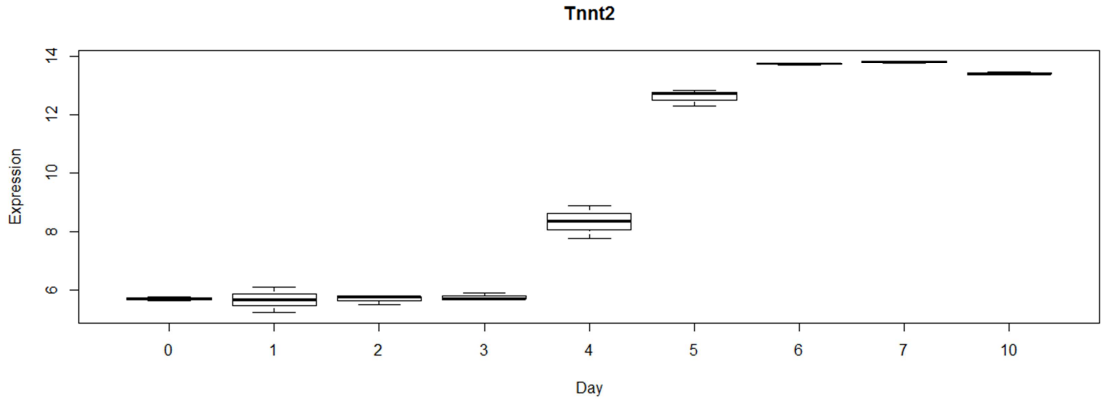


Figure 3.1.14. Expression of troponin T type 2 (cardiac) on Gaspar *et* al. 2012 data set.

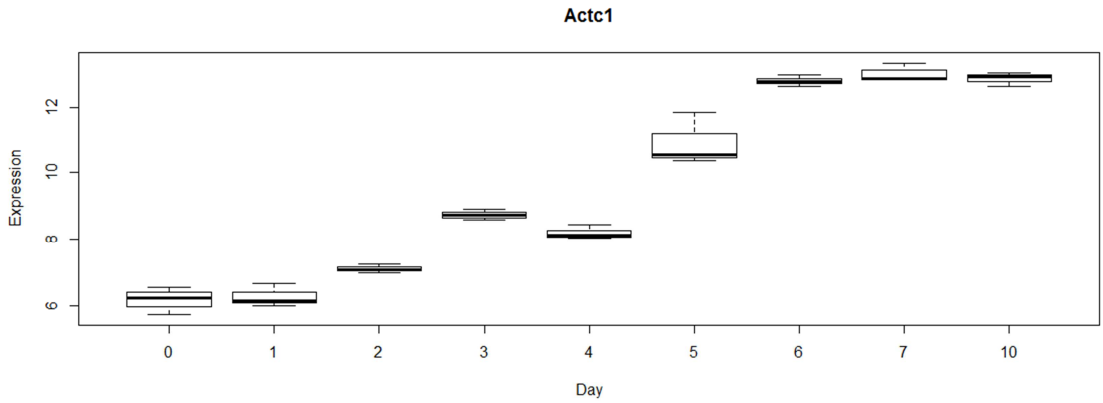


Figure 3.1.15. Expression of actin, alpha, cardiac muscle 1 on Gaspar *et* al. 2012 data set.

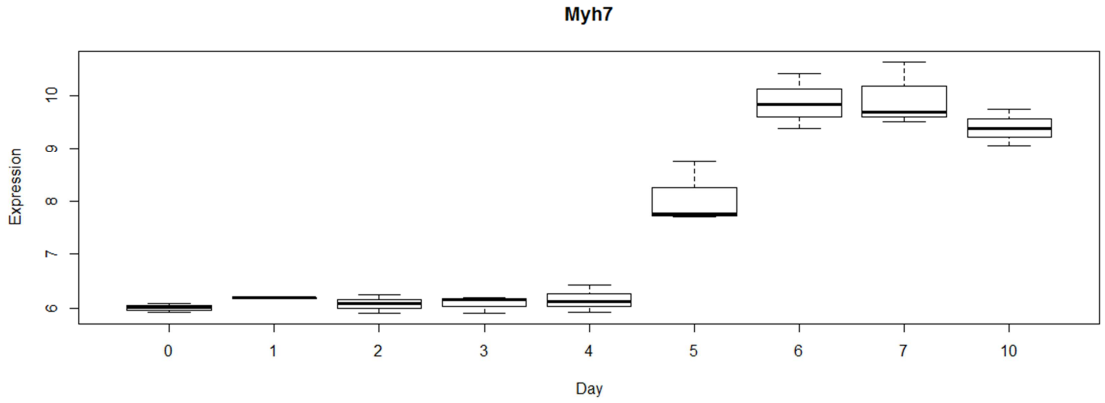


Figure 3.1.16. Expression of myosin, heavy chain 7, cardiac muscle, beta on Gaspar *et* al. 2012 data set.

In comparison of day 7 (table 3.1.8.) against day 10 (table 3.1.9), it is possible to assert that between these time points, biological functions linked to heart development are not listed among the processes with the highest enrichment in differentially expressed genes.

Table 3.1.8. Biological functional analysis in Day 7 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perk	Ad p-value
GO:0048513	organ development	283	100	35.34	2.37E-52
GO:0009653	anatomical structure morphogenesis	283	71	25.09	1.61E-39
GO:0009887	organ morphogenesis	283	57	20.14	6.85E-37
GO:0009888	tissue development	283	50	17.67	3.56E-31
GO:0007155	cell adhesion	283	50	17.67	8.35E-28
GO:0001944	vasculature development	283	33	11.66	4.05E-26
GO:0030154	cell differentiation	283	68	24.03	1.26E-25
GO:0001568	blood vessel development	283	32	11.31	4.05E-25
GO:0001501	heart development	283	34	12.01	1.58E-24
GO:0007507	anatomical structure formation involved in morphogenesis	283	28	9.89	4.38E-24
GO:0030198	cell development	283	20	7.07	2.69E-21
GO:0048646	muscle organ development	283	29	10.25	2.97E-20
GO:0048523	skeletal system development	283	55	19.43	1.14E-19
GO:0007517	blood vessel morphogenesis	283	25	8.83	2.14E-19
GO:0009605	heart morphogenesis	283	45	15.9	4.37E-19
GO:0048514	blood circulation	283	24	8.48	7.93E-18
GO:0048468	angiogenesis	283	39	13.78	2.81E-17
GO:0003007	cardiac muscle tissue development	283	16	5.65	4.21E-17
GO:0050793	extracellular matrix organization	283	39	13.78	6.59E-17
GO:0042221	muscle contraction	283	50	17.67	9.39E-17

DAY 10

The analysis of genes expressed at day 10 (table 3.1.9), show that they are mainly linked to fibroblast cells, so it seems that most of the genes linked to heart development and morphology are showing lower levels of expression or at least the pathways for heart development are slowly shutting down.

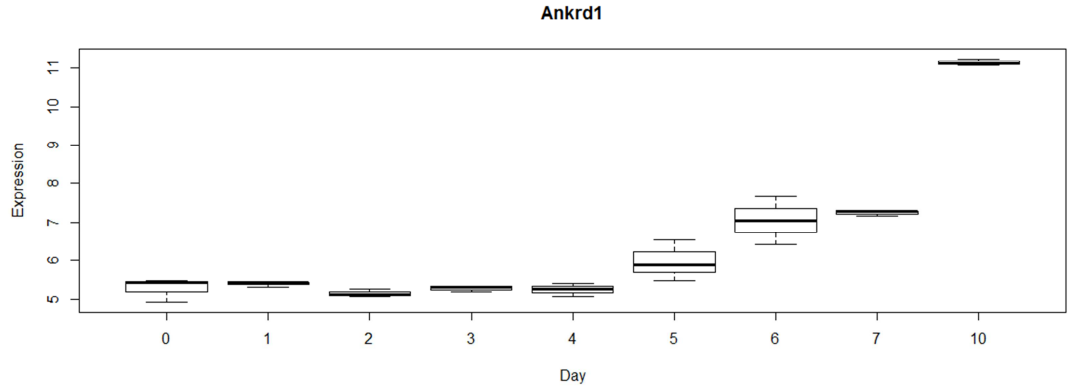


Figure 3.1.17. Expression of ankyrin repeat domain 2 (stretch responsive muscle) on Gaspar *et* al. 2012 data set.

Nevertheless, it is still possible to see some genes linked to heart development and morphology expressed, like Ankrd1 (figure 3.1.17). Interestingly, expression of Ankrd1 is peaking only at day 10 (or later). Ankrd1 is known to be a transcription co-factor and early differentiation marker of cardiac myogenesis expressed in the heart during embryonic development [54].The observed expression pattern may supports a potential role of Ankrd1 as a negative regulator of cardiac expression, although this is not yet demonstrated *in vivo* [55].

Table 3.1.9. Biological functional analysis in Day 10 on Gaspar *et* al. 2012 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0048513	organ development	509	143	28.09	1.76E-55
GO:0009653	anatomical structure morphogenesis	509	106	20.83	7.69E-46
GO:0009605	response to external stimulus	509	91	17.88	2.20E-37
GO:0009888	tissue development	509	74	14.54	1.12E-36
GO:0009887	organ morphogenesis	509	77	15.13	1.12E-36
GO:0042221	response to chemical stimulus	509	102	20.04	3.06E-34
GO:0009611	response to wounding	509	67	13.16	4.02E-34
GO:0030154	cell differentiation	509	109	21.41	8.54E-33
GO:0010033	response to organic substance	509	71	13.95	2.52E-29
GO:0050793	regulation of developmental process	509	74	14.54	2.52E-29
GO:0008283	cell proliferation	509	68	13.36	9.74E-27
GO:0001944	vasculature development	509	45	8.84	1.07E-26
GO:0006950	response to stress	509	109	21.41	1.44E-26
GO:0007155	cell adhesion	509	67	13.16	1.52E-26
GO:0001568	blood vessel development	509	44	8.64	5.10E-26
GO:0042127	regulation of cell proliferation	509	57	11.2	1.45E-24

3.2. GENETIC EXPRESSION IN DIFFERENT TISSUES

The analyses described in the previous section showed a characteristic sequence of activated genes during stem cell differentiation. Moreover, genes could be classified with respect to the time point when their expression reaches a maximum. Genes associated with stem cell maintenance and proliferation showed an early peak in their expression, whereas genes associated with organ development are up-regulated at later time points. These characteristic profiles point to distinct cellular 'programmes' for tissue generation. A fundamental question is whether these programmes also underlie tissue regeneration, or whether they are restricted to developmental phase of an organism. To obtain some insights into this matter, I analysed the expression in differentiated tissue using five microarray data sets (Rosetta1, Rosetta2, Stanford, Mouse Gene Atlas and Human Gene Atlas). First, I assessed whether genes down-regulated during stem cell differentiation remain generally down-regulated in a wide range of tissue. Similarly, I examined whether genes up-regulated during differentiation remain at a high expression level. These sets of up-and down-regulated genes during differentiation were obtained directly from the study by Gaspar et al (2012), which used a filter approach for their identification (supplementary table 1, Annex V). Due to the importance of transcription factors in this context, special lists were generated including only these proteins.

In general, we will see that the down-regulated genes show a lower expression and the up-regulated genes show a higher expression. As all 5 data sets showed similar patterns in their gene expression, only the Rosetta1 experiment is presented here, whereas the results from the other 4 data sets (Rosetta2, Stanford, Mouse Gene Atlas and Human Gene Atlas) will be displayed in the supplementary figure 1-8 (Annex I).

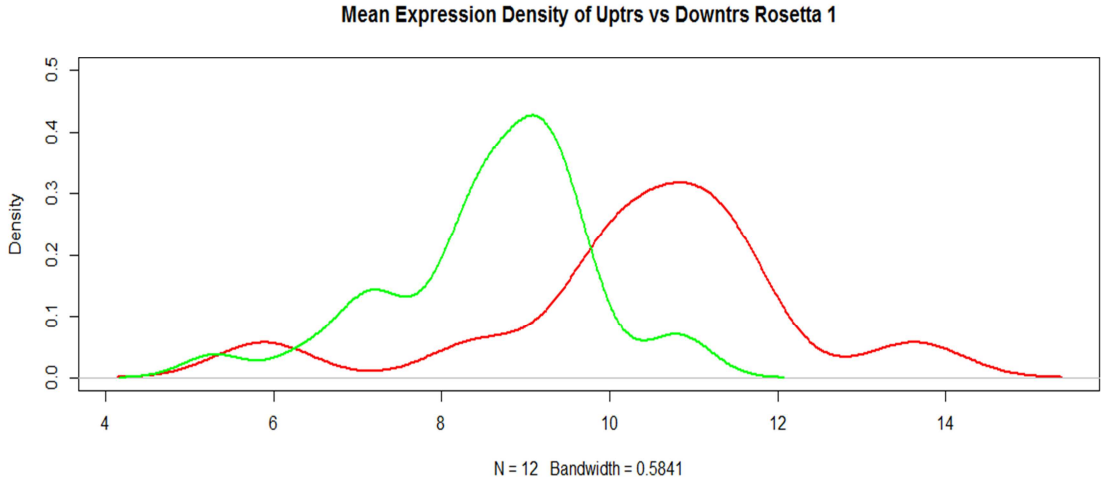


Figure 3.2.1. Mean expression in density of transcription factor that are up regulated vs. down regulated Rosetta 1.

As expected, the distribution of expression values for down-regulated genes show a peak at a lower expression compared to the up-regulated genes. This means that the down-regulated genes display a lower expression values in differentiated cells with a peak at log2 of 9 (meaning that expression intensity is around 512) and the up regulated genes present a higher expression with a peak at log2 of 11 (expression around 2048). We can observe this type of pattern in both line plots, showing that the selection applied was behaving as expected, as in the other data sets it was the same behaviour (supplementary figure 1-8, Annex I).

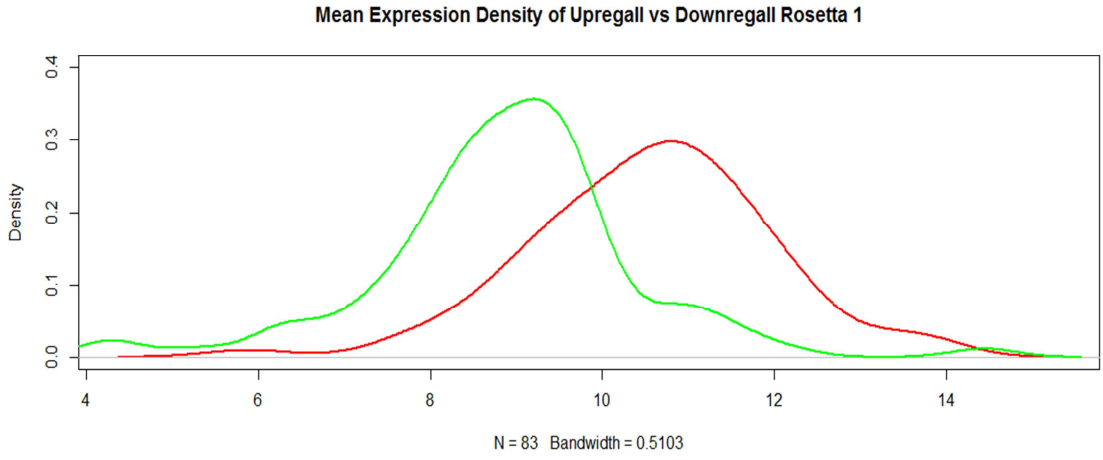


Figure 3.2.2. Mean expression in density of all genes that are up regulated vs. down regulated Rosetta1.

Additionally I generated some barplots and line plots to show average expression of the up- and down-regulated genes and transcription factors in all tissues. Essentially, the bar plots show individual gene expression, and the objective is to look for the most and less expressed genes present in each data set, for example, Lgals1, KLF6, TPM1 are the most expressed genes in every data set for tissue expression , while Nanog, Pou5f1, Sox2, Esrrb, Fgf4 were the lowest expressed genes in almost every data set (Supplementary figure 9-16, Annex I).

I also calculated the average of each gene in all tissues, to check which genes were most commonly up- and down-regulated in differentiated cells (Figure 3.2.3).

As expected, most of the genes that are up-regulated in differentiated cells, such as CD44, Klf6, Mef2a, Tnnc1, Lgals1 and Rpl13 which are majorly linked to regulation of cell proliferation.

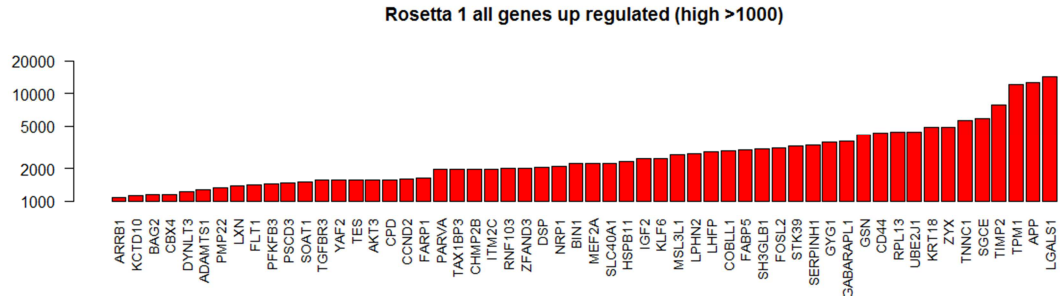


Figure 3.2.3. Genes with the highest averaged expression in all tissue in Rosetta 1 data set.

On the other hand, the down-regulated genes were essentially genes that are usually highly expressed in undifferentiated cells such as ESRRB, DNMT3L and FGF4.

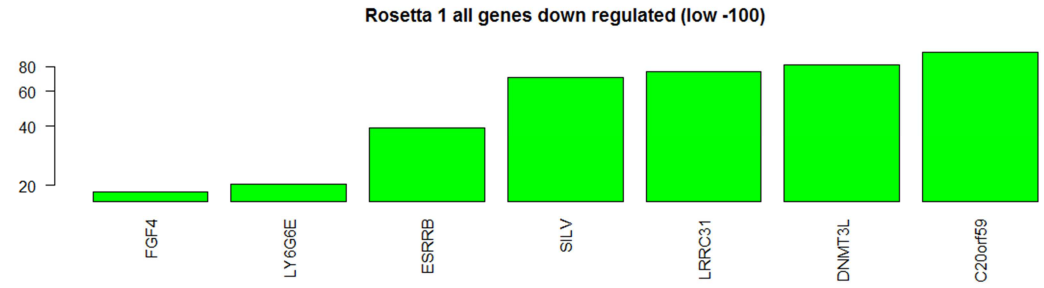


Figure 3.2.4. Genes with the highest averaged expression in all tissue in Rosetta 1 data set.

After identification of genes up- and down-regulated, I clustered and analysed them. Remarkably, this cluster showed that the transcription factors responsible for the undifferentiated cell state are clustering all together (Pou5f1, also known as OCT4), Nanog, Jarid2), indicating that these genes remain co-regulated in differentiated tissue. A striking observation here is that the gene expression of this set of genes appears to correlate with known regenerative capacity of tissues. For instance, liver which is known to have high regenerative capacity shows a prominent up-regulation of these genes, whereas heart, which is known to have limited regenerative capacities, displays low expression of this gene set. This observation strongly suggests that it is possible to define a gene signature for the regenerative capacity of tissues. Further analyses need to be performed to clarify this remarkable finding.

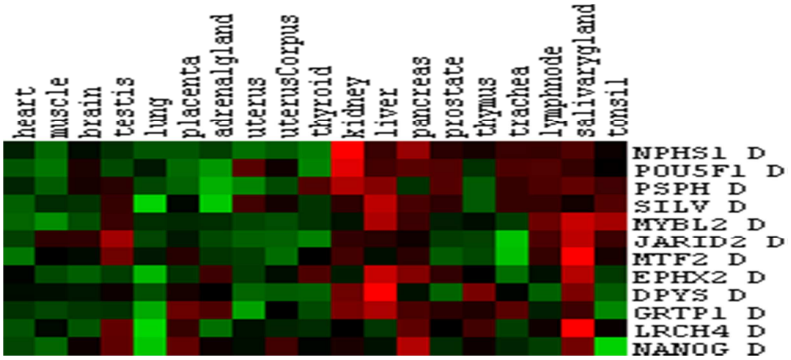


Figure 3.2.5. Group of genes that are clustering together in Rosetta 1.

3.3. GENETIC EXPRESSION BETWEEN CARDIAC FIBROBLAST AND INDUCED CARDIOMYOCYTES AND DIFFERENTIATION INTO CARDIOMYOCYTES

In this sub chapter, we compare the expression pattern of different experiments in which cardiomyocytes were generated *in vitro*:

- Gene expression in mouse neonatal cardiomyocytes, cardiac fibroblasts, cardiac fibroblast that failed reprogramming into iCMs (GFP- cells), and reprogrammed iCMs (GFP+) [4];
- Heart repair by reprogramming non-myocytes with cardiac transcription factors [5];
- Time-course expression of human iPSc differentiation toward cardiomyocytes [3].

The analysis focuses on genes coding for proteins of known importance for heart morphogenesis and development as well as their interacting partners, such as transcription factors that are believed to be related to heart cell fate, namely Gata4, Mef2c, Tbx5, Hand2, Isl1 and Nkx2-5 [3-5]. These three studies give diversified options for the reprogramming of iCM, such as utilization of three or four reprogramming transcription factors or the direct differentiation of hiPSc towards cardiac cell fate. Moreover, with these three studies it is possible to cover more genes for study of the differentiation and/or transdifferentiation into cardiomyocytes.

3.3.1. REPROGRAMMING MOUSE FIBROBLAST INTO FUNCTIONAL CARDIOMYOCITES BY DEFINED FACTORS

Reprogramming of fibroblast cells to iPSc shows the possibility that a somatic cell can be reprogrammed to an alternative differentiated fate without first becoming a pluripotent stem cells or a progenitor cells [4]. So Ieda's group (2010) reported that the combination of three transcription factors essential for heart developmental (i.e. Gata4, Mef2c and Tbx5) could rapidly and efficiently reprogram cardiac fibroblast directly into functional differentiated cardiomyocyte-like cells, so called induced Cardiac Myocytes (iCM) [4]. iCM expressed cardiac-specific markers and had a global gene expression profile similar to cardiomyocytes and contracted spontaneously.

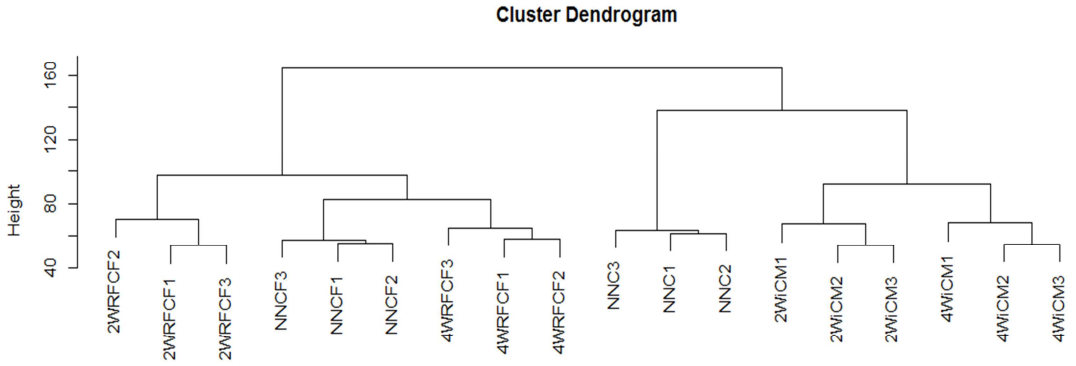


Figure 3.3.1.1. Cluster dendrogram for the data set published by Ieda *et* al. 2010. 2WRFCF (2 Weeks Reprogramming Fibroblast Cells Failed); 2WRFCF (4 Weeks Reprogramming Fibroblast Cells Failed); NNCF (Neonatal Cardiac Fibroblasts); 2WiCM (2 Weeks Induced Cardiomyocytes); 4WiCM (4 Weeks Induced Cardiomyocytes); NNC (Neonatal Cardiomyocytes).

For the analysis of this data set, I first generated a cluster dendrogram to examine if the different cell types and replicates were clustering together. Otherwise if they do not cluster together correctly, these measurements could be compromised by artefacts. Assessment of the dendrogram, however, shows that is not the case. Moreover, we can observe that replicates are clustering together and the 2 main branches are divided in cardiomyocytes and cardiac fibroblast. I will present some of the analysis of the most relevant genes in each category, these genes being cardiac differentiation, cardiomyocytes, and cardiac fibroblast markers.

CARDIAC TRANSDIFFERENTIATON MARKERS

Gata4, Mef2c, and Tbx5 were the factors used to induce cardiac fibroblast to cardiomyocyte transdifferentiation. These gene are highly expressed during the transdifferentiation process, and show a higher expression in induced cardiomyocytes (2WiCM and 4WiCM) than in neonatal cardiomyocytes (NNC) (figure 3.3.1.2; figure 3.3.1.3; figure 3.3.1.4). As explain previously in this work, these transcription factors are known to be important genes in the transdifferentiation into cardiac cells [4, 5, 15, 52, 53].

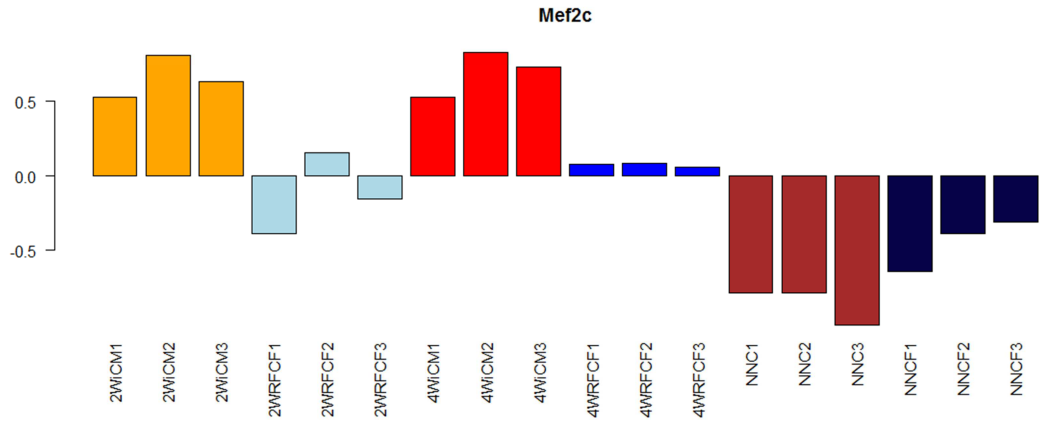


Figure 3.3.1.2. Expression of myocyte enhance factor 2C in several cells types on Ieda e*t al*., 2010 data set.

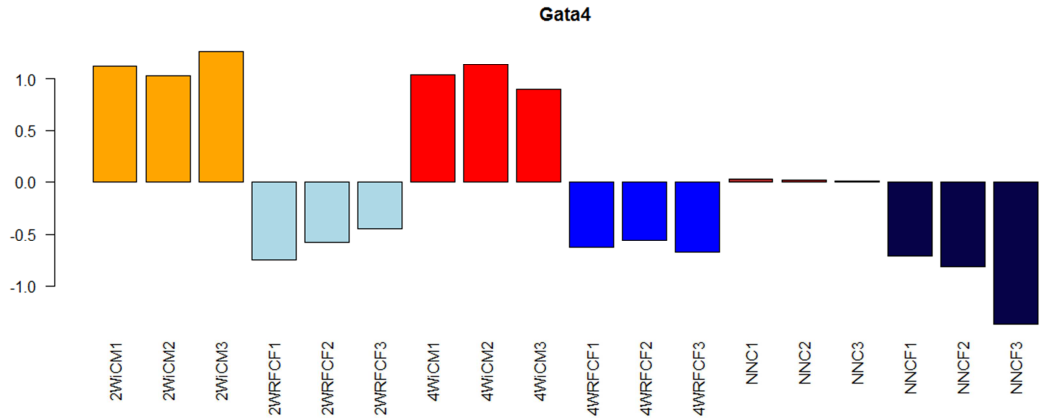


Figure 3.3.1.3. Expression of GATA binding protein 4 in several cells types on Ieda e*t al*., 2010 data set.

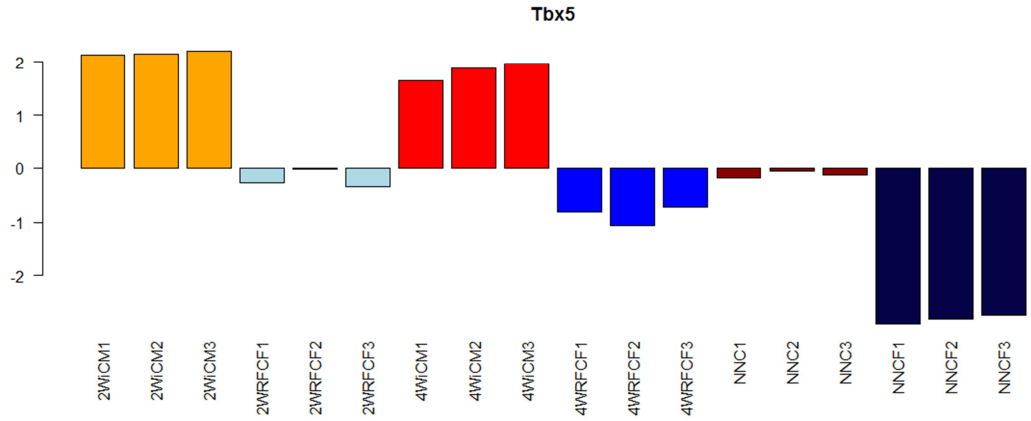


Figure 3.3.1.4. Expression of T-box 5 in several cells types on Ieda e*t al*., 2010 data set.

Isl1 and Nkx2-5, which are key factors involved in heart development and are marker genes for cardiac progenitor cells showed an unexpected expression. The expression of Nkx2-5 in iCMs is lower than in in NNC, and Isl1 shows a very odd expression (Supplementary figure 1-2, Annex II). Nkx2.5 is express in different cardiac subtypes (such as ventricular cardiomyocytes, atrial cardiomyocytes and other cells...), whereas the iCMs were shown to behave like ventricular cardiomyocytes [4]. The number of Isl1-expressing cells in neonatal hearts decrease rapidly and do not mark mature cardiomyocytes, which confirms that iCM are fully mature cardiomyocytes and the transdifferentiation process of cardiac fibroblast do not go via an intermediate progenitor cell type.

MATURE CARDIAC MARKERS

In the following barplots, I present some of the most important cardiac markers for mature cardiomyocytes. As expected, they are expressed in neonatal cardiac cells but were not detected in cardiac fibroblasts (figures 3.3.1.5, 3.3.1.6, 3.3.1.7 and 3.3.1.8). However, these markers are also expressed in iCM at week 2 and 4, showing that these cells are getting transdifferentiated correctly into functional cardiomyocytes. To confirm if the cardiac fibroblasts transdifferentiated correctly into induced cardiomyocytes I also assessed the expression profiles of standard cardiac markers, such as, Myh7 (figure 3.3.1.5.), Fgf1 (figure 3.3.1.6.) and Tnnt2 (figure 3.3.1.7). We can observe a comparable expression in the neonatal cardiac cells and the induced cardiomyocytes which was similar expression to the neonatal cardiomyocytes.

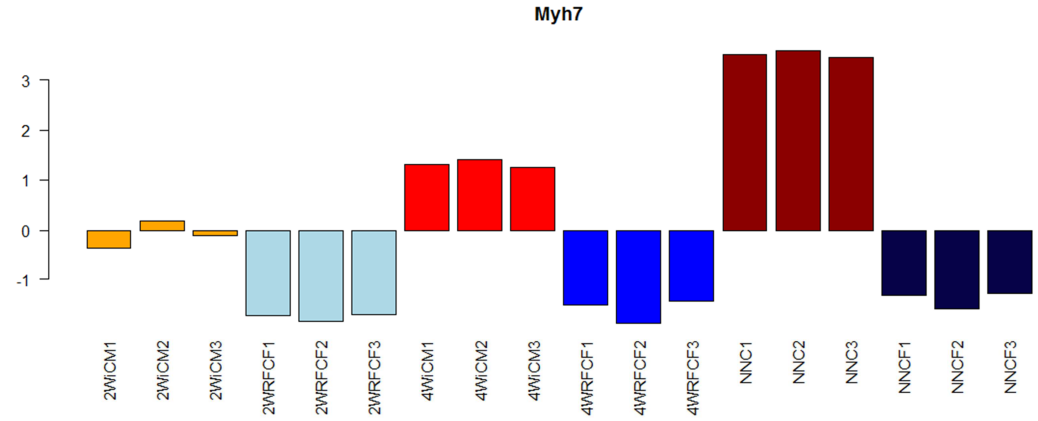


Figure 3.3.1.5. Expression of myosin, heavy chain 7, cardiac muscle, beta in several cells types on Ieda e*t al*., 2010 data set.

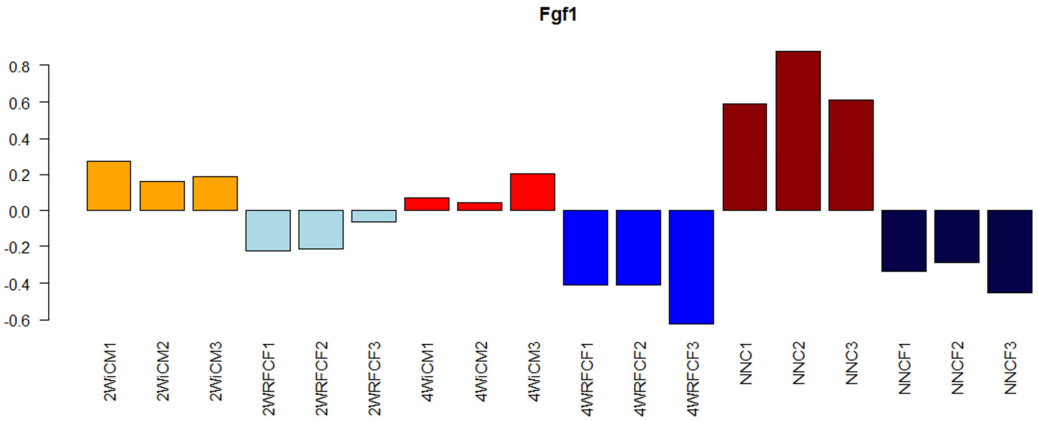


Figure 3.3.1.6. Expression of Fibroblast growth factor 1 in several cells types on Ieda e*t al*., 2010 data set.

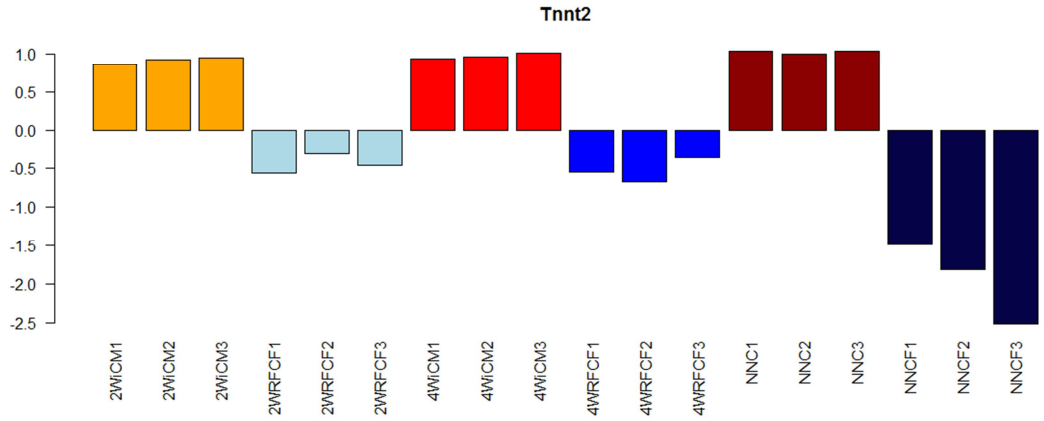


Figure 3.3.1.7. Expression of troponin T type 2 (cardiac) in several cells types on Ieda e*t al.*, 2010 data set.

The expression of Actc1, which is involved in cell motility, heart development and morphogenesis and cell differentiation, was consistent in reprogrammed iCMs and neonatal cardiomyocytes (figure 3.3.1.8).

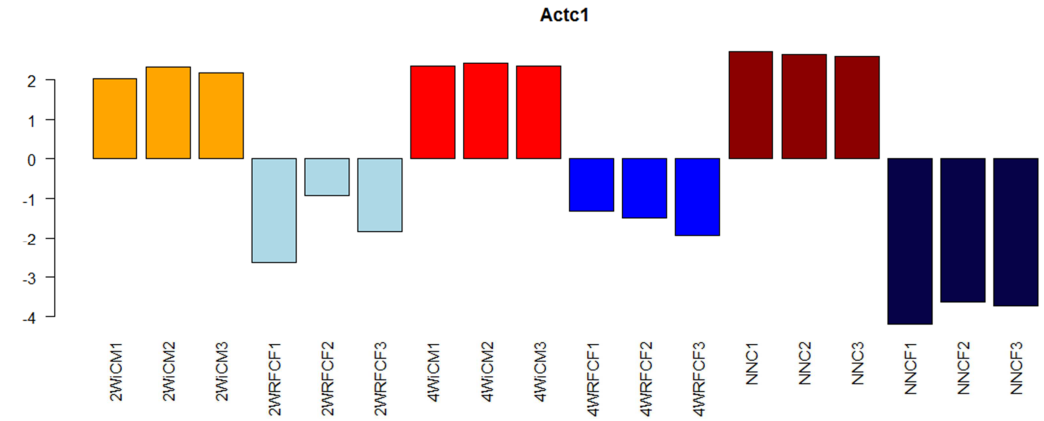


Figure 3.3.1.8. Expression of actin, alpha, cardiac muscle 1 in several cells types on Ieda e*t al.*, 2010 data set.

CARDIAC FIBROBLASTS MARKERS

Interestingly, we observed that genes related to cardiac fibroblast, such as, Fgfr1 (figure 3.3.1.9.), Ddr2 (figure 3.3.1.10.) and Fgf2 (figure 3.3.1.11) are expressed in early transdifferentiated induced cardiomyocytes and in neonatal cardiac fibroblast. This could mean that they can be somewhat important in the cell development or establishment of these cells into cardiac muscle, or it could mean that the characteristics of the cardiac fibroblast - or at least the normal expression of the genes linked to cardiac fibroblast - start to fade away.

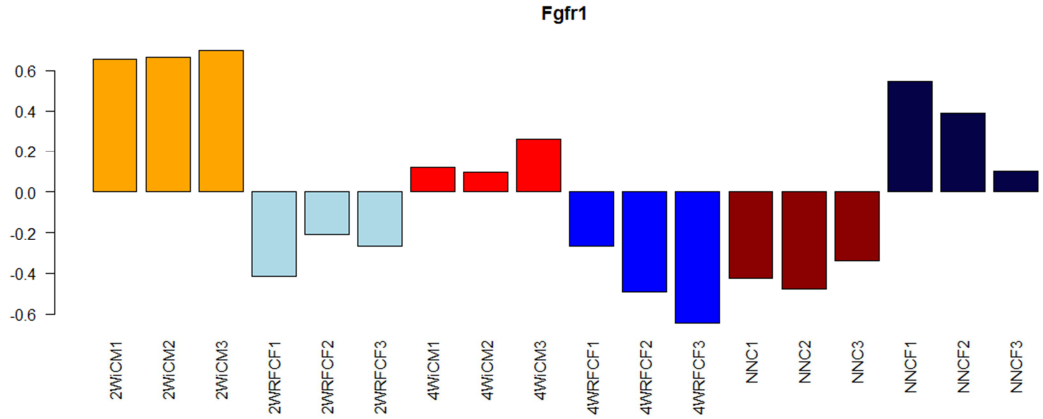


Figure 3.3.1.9. Expression of fibroblast growth factor receptor 1 in several cells types on Ieda e*t al*., 2010 data set.

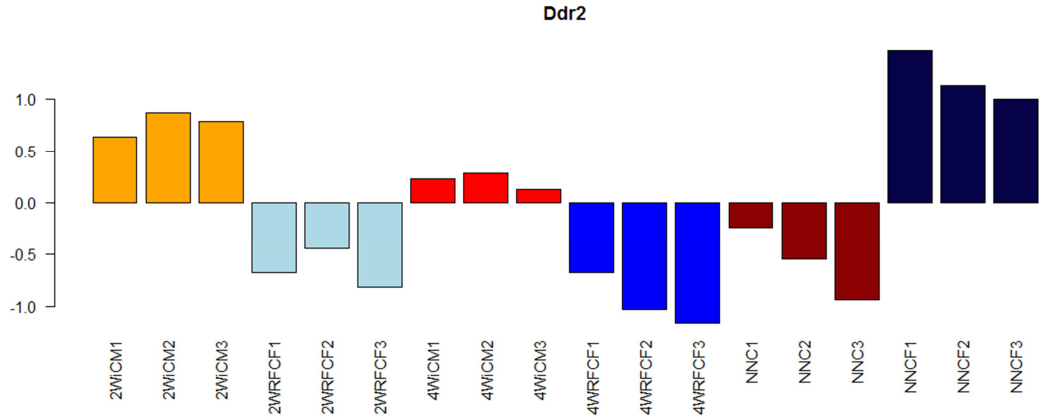


Figure 3.3.1.10. Expression of discoidin domain receptor tyrosine kinase 2 in several cells types on Ieda e*t al,* 2010 data set.

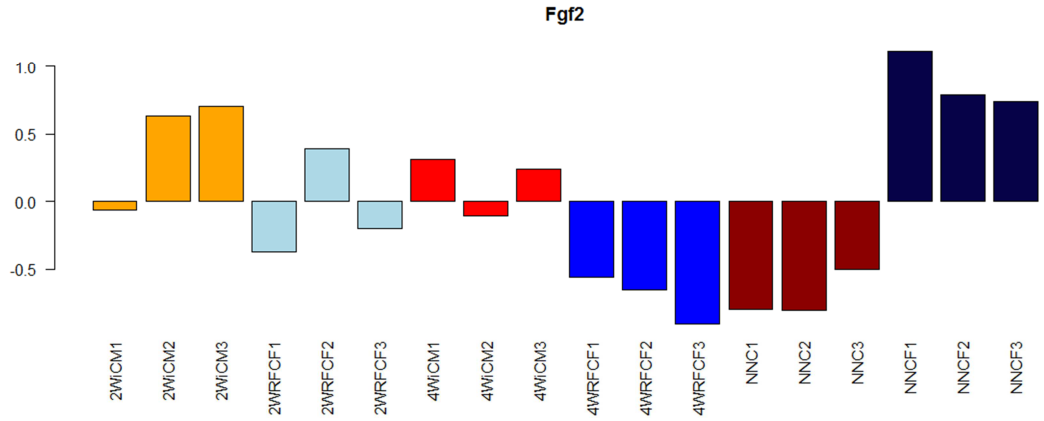


Figure 3.3.1.11. Expression of fibroblast growth factor 2 in several cells types on Ieda e*t al.*, 2010 data set.

3.3.2. REPROGRAMMING NON-MYOCYTES WITH CARDIAC TRANSCRIPTION FACTORS

As cardiac fibroblasts account for the majority of cells present in the heart and potential cellular source for restoration of cardiac function, Song's research team (2012) used four transcription factors, Gata4, Hand2, Mef2c and Tbx5, showing that these genes can cooperatively reprogram cardiac fibroblast into beating cardiac-like myocytes *in vitro* [5]. Song (2012) also showed that forced expression of these factors in dividing non-cardiomyocytes *in vivo* reprograms these cells into functional cardiac-like myocytes and reduces adverse heart remodelling after cardiac infarction [5].

I performed a cluster dendrogram analysis of the genes expressed in adult cardiomyocytes, in adult cardiac fibroblasts (ACF) and reprogrammed cardiomyocytes from the data previously obtained [5]. After the analysis of the data, we observed that the expression pattern of ACF transduced with the reprogramming factors differ from those of control ACF, presenting a differential expression of several and important heart development linked genes, that will be described below.

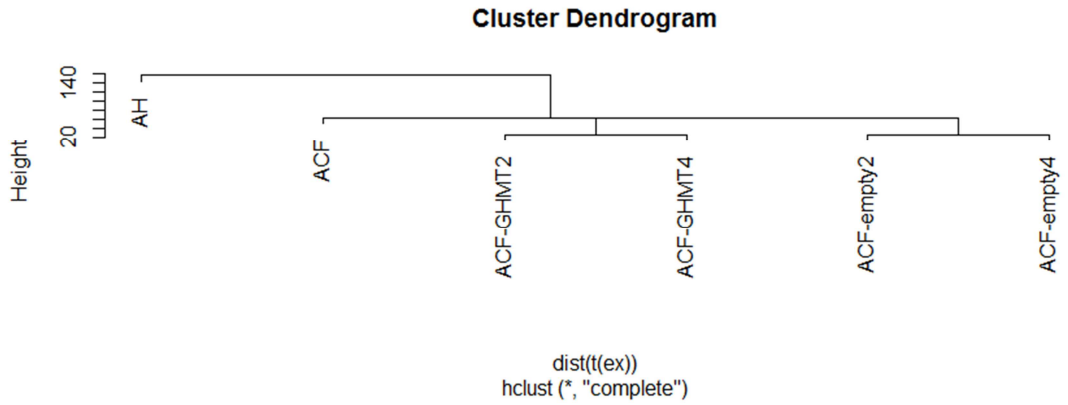


Figure 3.3.2.1. Cluster dendrogram for the Song *et* al., 2012 data set. AH (Adult Heart); ACF (Adult Cardiac Fibroblast); ACF-GHMT2 (2 weeks Adult Cardiac Fibroblast reprogrammed with GHMT transcription factor); ACF-GHMT4 (4 weeks Adult Cardiac Fibroblast reprogrammed with GHMT transcription factors); ACF-Empty2 (2 weeks Adult Cardiac Fibroblast reprogrammed with empty vector); ACF-Empty4 (4 weeks Adult Cardiac Fibroblast reprogrammed with empty vector).

Exogenous Gata4, Hand2, Mef2c and Tbx5 transcription factors (GHMT) were used *in vivo* in an injured myocardium to reprogram heart resident cardiac fibroblasts and led to an improvement in the function and structure of an infracted myocardium. [5]. The apparent efficiency of reprogramming was greater *in* vivo than *in vitro* perhaps due to environment

surrounding the heart, like extracellular matrix, growth factors, surrounding contractile cells and other cell types, being more permissive than artificial tissue culture [5].

CARDIAC TRANSDIFFERENTIATON MARKERS

In this section, I present the expression of the *in vivo* reprogramming transcription factors used for the transdifferentiation of cardiac fibroblast into induced cardiomyocytes-like cells. With the exception of Mef2c, the expression of GHMT factors is not considerably higher than their expression in control ACF and ACF transduced with empty vectors after 2 or 4 weeks (ACF-Empty2 and 4). As expected, the transdifferentiated cardiomyocytes derived from cardiac fibroblast displayed a similar, although smaller, gene expression to normal adult heart cardiomyocytes (figure 3.3.2.2. – 3.3.2.5.).

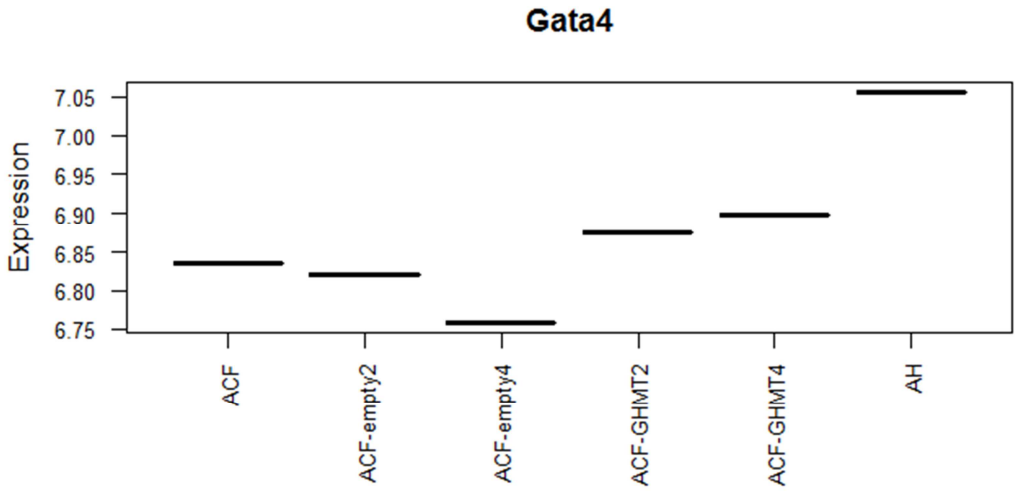


Figure 3.3.2.2. Expression of GATA binding protein 4 on Song e*t al*., 2012 data set.

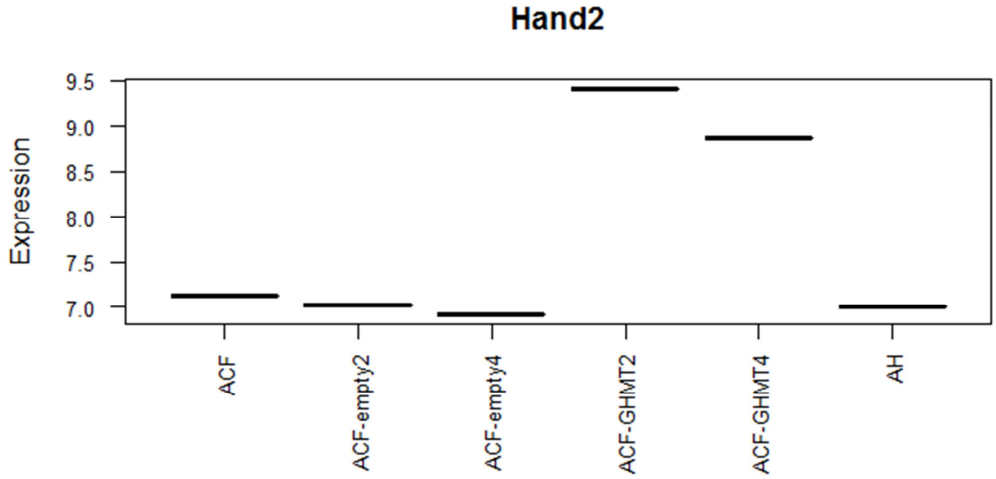


Figure 3.3.2.3. Expression of heart and neural crest derivatives expressed transcript 2 on Song e*t al*., 2012 data set.

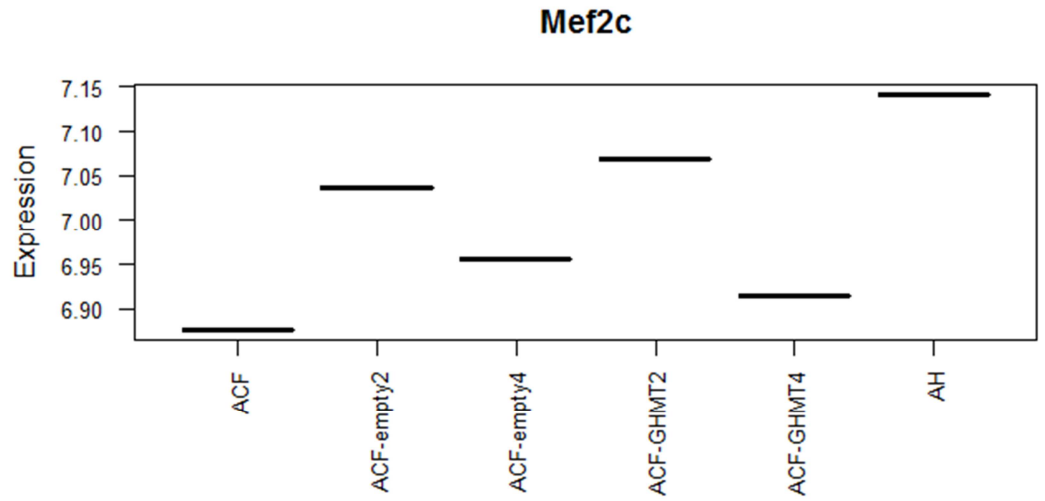


Figure 3.3.2.4. Expression of myocyte enhance factor 2C on Song e*t al.*, 2012 data set.

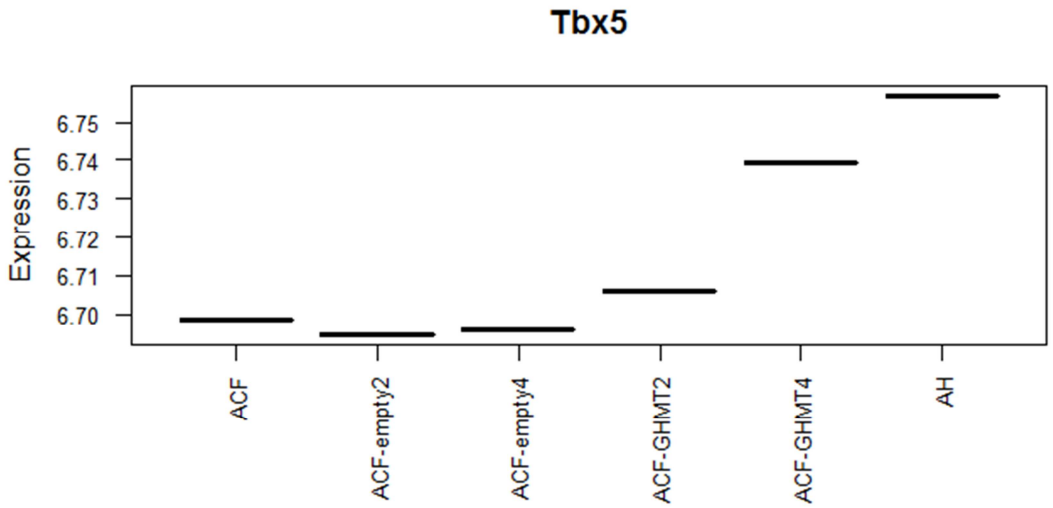


Figure 3.3.2.5. Expression of T-box 5 on Song e*t al.*, 2012 data set.

Observing these results it is possible to infer that this transcription factors can be important to heart development and morphogenesis, but even so it will be required in-depth analysis of this data set to get a better conclusion. In line with the results obtained with the previous study (Ieda et al. 2010), Isl1 expression is not consistent indicating that this factor might not be crucial for the transdifferentiation process (Supplementary figure 1, Annex III).

MATURE CARDIAC MARKERS

The results obtained indicate that mature cardiac markers are differentially expressed in reprogrammed cells when compared to the adult cardiac cells (figure 3.3.2.6 and 3.3.2.7.). Thus, the forced expression of transcription factors Gata4, Hand2, Mef2c and Tbx5 in non-myocytes present in the heart are insufficient to induce the expression of these cardiac markers to endogenous levels.

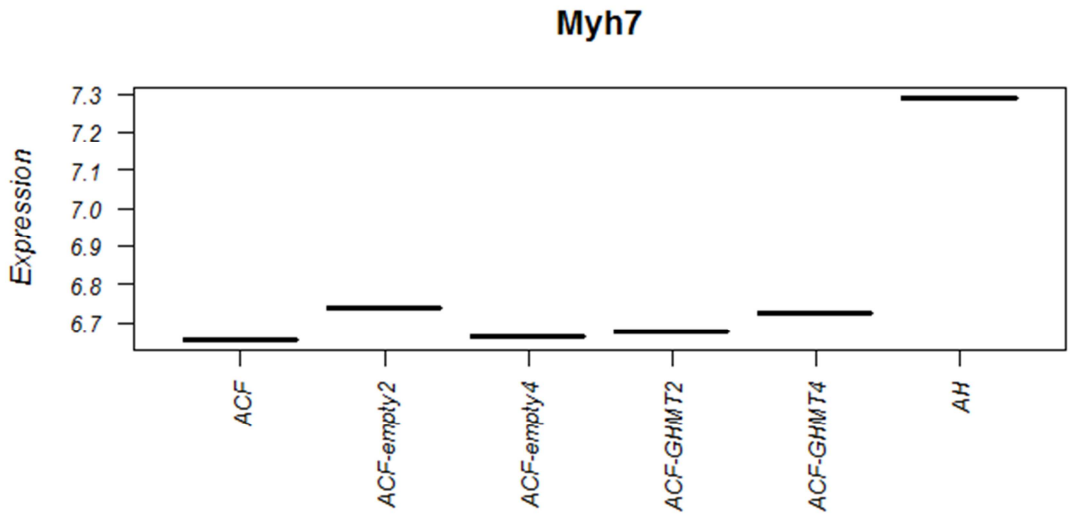


Figure 3.3.2.6. Expression of myosin heavy chain beta on Song e*t al*., 2012 data set.

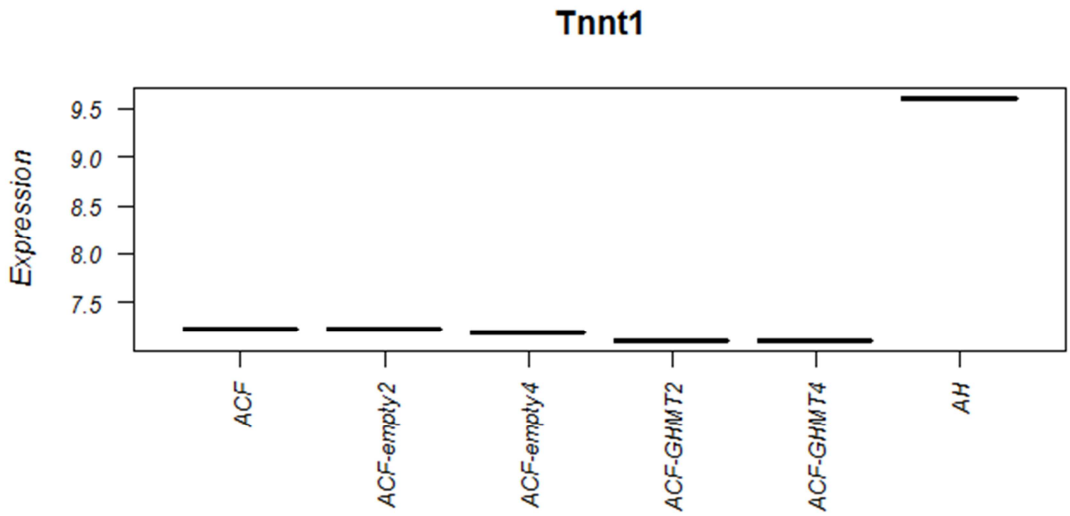


Figure 3.3.2.7. Expression of troponin T type 1 on Song e*t al*., 2012 data set.

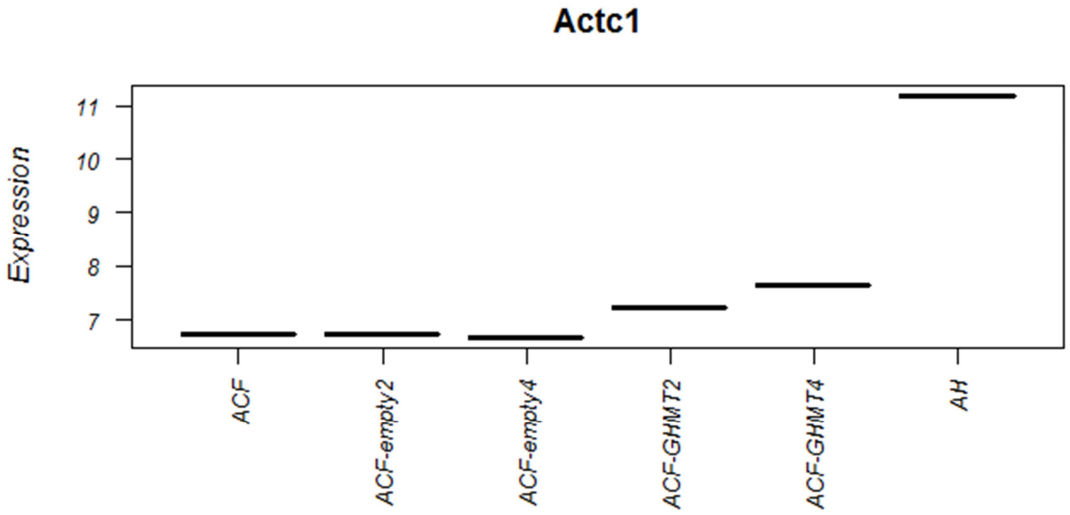


Figure 3.3.2.8. Expression of actin, alpha cardiac muscle 1 on Song e*t al*., 2012 data set.

CARDIAC FIBROBLAST MARKERS

The obtained results are not that obvious as demonstrated in the previous study [5], since the expression patterns of the fibroblast markers are still very high (figure 3.3.2.9 and 3.3.2.10.). This pattern could demonstrate that the cells did not have time to completely transdifferentiate, since they still show fibroblast marker genes. But their studies revealed that the injection of the transcription factors on the heart improve reprogramming of cardiac fibroblast to iCM *in vivo*, that could be due to the native milieu existing in the heart, leading to a better reprogramming [5]. Since the results analysed come from an *in vitro* experiment, it is not possible to see clearly the down regulation of the cardiac fibroblast markers.

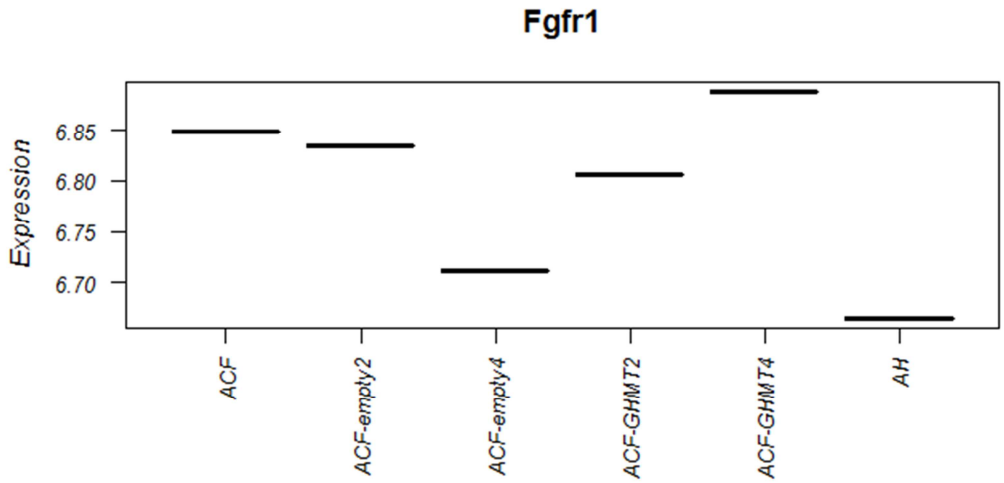


Figure 3.3.2.9. Expression of fibroblast growth factor receptor 1 on Song e*t al*., 2012 data set.

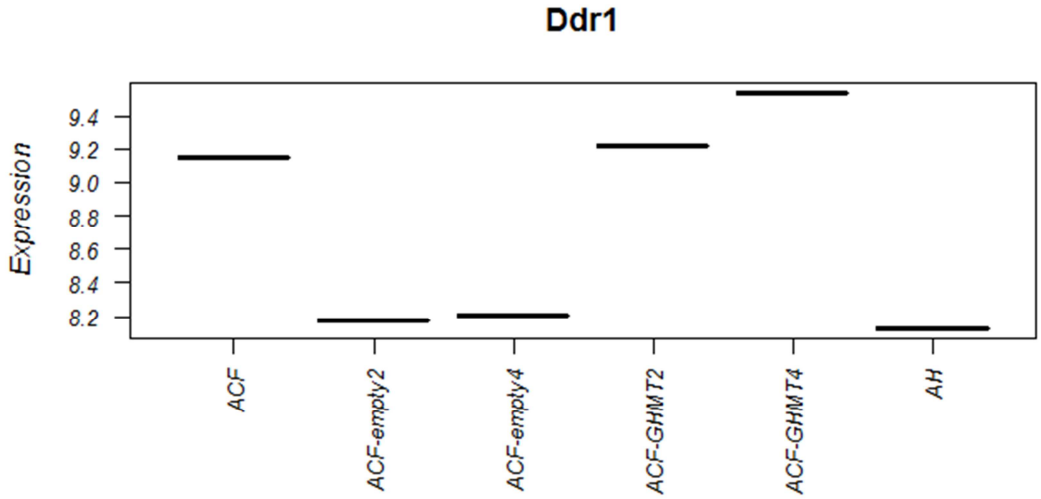


Figure 3.3.2.10. Expression of discoidin domain receptor family, member 1, on Song e*t al*., 2012 data set.

3.3.3. HIPSC DIFFERENTIATION TOWARD CARDIOMYOCYTES

The difference between this Uosaki (2011) data set [3] and the general genetic expression in mouse embryonic stem cells of Gaspar (2012) publication [21] is that the day 0 starts with the beginning of cardiac differentiation, while in Gaspar [21] day 0 corresponds to undifferentiated mouse stem cells that started to differentiate. Uosaki (2011) and colleagues obtained cardiac myocytes through differentiated human iPSc towards cardiomyocytes applying sequential administration of activin, bone morphogenetic protein 4 (BMP4), fibroblast growth factor 4 (FGF4) and Dickkopf 1 homolog (DKK1). Through the cluster dendrogram it is possible to see that the days of cardiac differentiation are clustering as expected, meaning that genetic expression in those days are very similar. So with these results, we can at least see that the expressions for the replicates are matching together, indicating that the expression values obtained are reproducible. For a further analysis it was necessary to check gene expression for each day, to see if specific genes were expression at the expected times according to other studies [13, 21, 44].

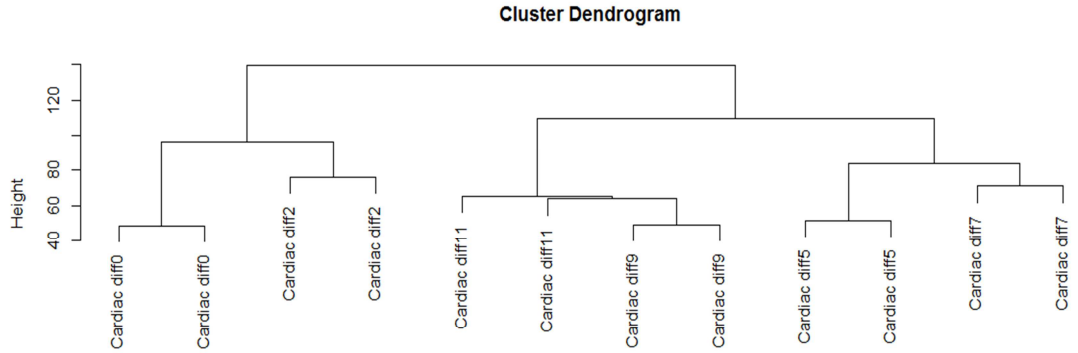


Figure 3.3.3.1. Cluster dendrogram for the Uosaki *et* al., 2012 data set. Cardiac diff0 (Differentiation into cardiomyocytes – Day 0); Cardiac diff2 (Differentiation into cardiomyocytes – Day 2); Cardiac diff5 (Differentiation into cardiomyocytes – Day 5); Cardiac diff7 (Differentiation into cardiomyocytes – Day 7); Cardiac diff9 (Differentiation into cardiomyocytes – Day 9); Cardiac diff11 (Differentiation into cardiomyocytes – Day11).

With the “topTables” obtained through “limma”, that is an R package, it was possible to see which genes were more expressed in each day. To verify which genes were up-regulated each day, they were sorted by “logFC” and it was chosen the ones with a positive fold change.

After filtering the obtained genes, It was performed a biological functional analysis through *Babelomics* (http://babelomics.bioinfo.cipf.es), using *FatiGO*. It was obtained 1 table for each day and sorted then by "p-value", choosing the top scores and most relevant functions.

Table 3.3.3.1. Temporal alignment (in days) between some key marker genes present in Mouse and Human. (General Genetic Expression In Mouse Embryonic Stem Cells [21] and hiPSc Differentiation Toward Cardiomyocytes [3]).

Data sets	Org.	Nanog, Lefty1, Pou5f1	T, Eomes, Sox17	Snai2, Meis1	Lum	Hand2, Zfpm2, Myh6	Myl4
General Genetic Expression In Mouse Embryonic Stem Cells	Mouse	0, 1 & 2	3, 4	6,7	10	10	10
hiPSc Differentiation Toward Cardiomyocytes	Human	0	2	5	7	9	11

I also compare the peak times for gene expression observed for the differentiation of mouse ESc with those observed for hiPSc. According to these findings (table 3.3.3.1.), comparing these two data sets, the peak expression for these marker genes occur in different days. Moreover, it seems that the expression peaks in human come earlier, which could be result of the use of additional factors for differentiation. For an early stage it was used the pluripotency markers Nanog, Lefty1 and Pou5f1; for an intermediate stage it was used cell fate/commitment markers like T-brachyury, Eomes and Sox17; and for a late stage it was used general cardiac markers like Meis1, Hand2 and Myh6. Unfortunately, Gaspar (2012) dataset do not have a measurement for all days, and only has for some days. This was the best comparison I was able to perform. Even so, it is possible to observe that the expression of these specific gene markers start first in human stem cells and in mouse start more or less a day later.

DAY 0

At day 0 of differentiation towards cardiac cells, it is possible to see the main transcription factors for maintenance of the pluripotent stem cells (figure 3.3.3.2.; figure 3.3.3.3; figure 3.3.3.4) are strongly up-regulated. This roughly corresponds to the period of day 0 to day 2 of Gaspar dataset [21], as rapidly in the next time points, their expression decreased quickly.

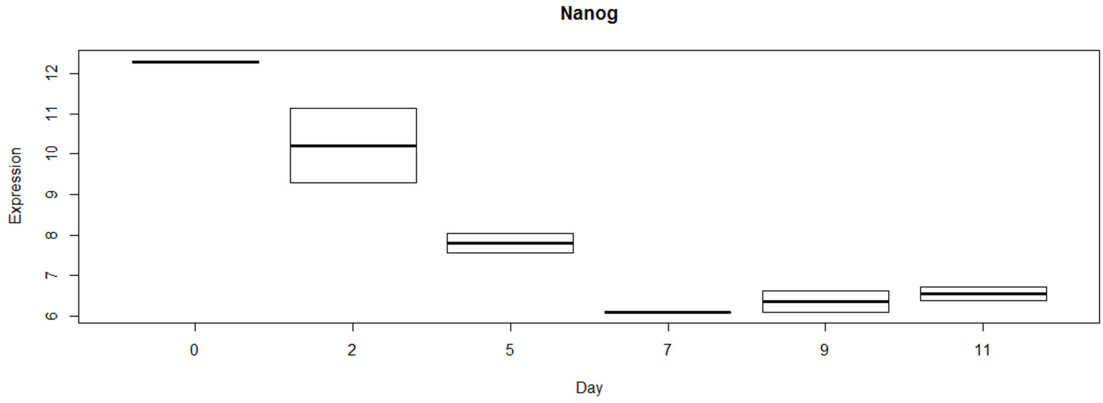


Figure 3.3.3.2. Expression of Nanog homeobox on Uosaki e*t al*., 2011 data set.

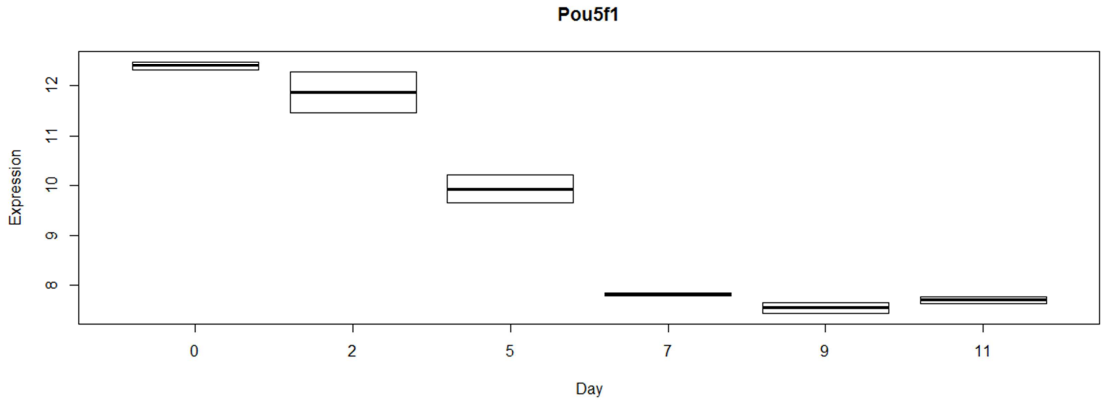


Figure 3.3.3.3. Expression of Pou domain, class 5, transcription factor 1 on Uosaki e*t al*., 2011 data set.

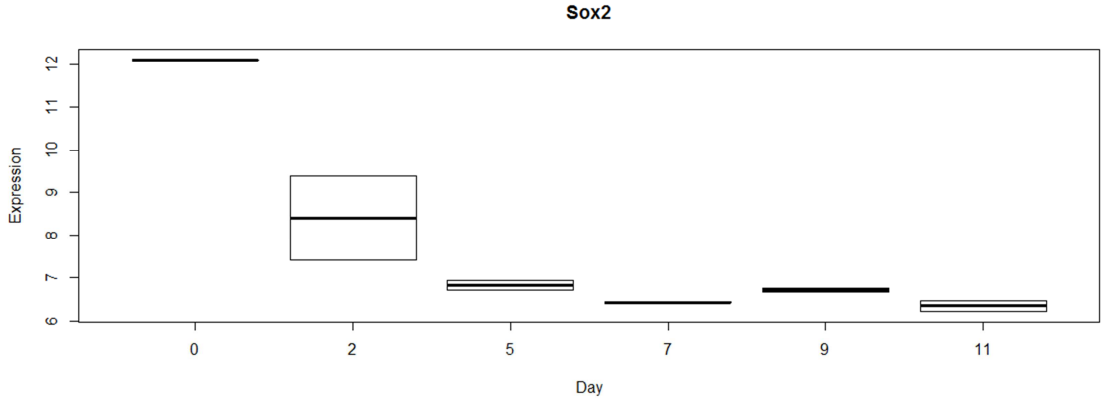


Figure 3.3.3.4. Expression of SRY-box containing gene 2 on Uosaki e*t al*., 2011 data set.

We also observed at day 0 that high expression of transcription factors are not only involved in the pluripotency state of ESCs but are also involved in heart development in particular (Table 3.3.3.2). This observation is in agreement with observations by other groups and strengthen the idea that human ESC although they are pluripotent and undifferentiated, they are primed, in other words are ready, to undergo differentiation processes by expressing messengers of master regulators of cell fate decisions.

Table 3.3.3.2. Biological functional analysis in Day 0 on Uosaki e*t al*., 2011 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0007507	heart development	342	30	8.77	6.73E-18
GO:0009790	embryo development	342	46	13.45	9.13E-18
GO:0048738	cardiac muscle tissue development	342	18	5.26	3.54E-16
GO:0007517	muscle organ development	342	30	8.77	3.90E-15
GO:0048646	anatomical structure formation involved in morphogenesis	342	33	9.65	7.47E-15
GO:0009888	tissue development	342	46	13.45	2.06E-14
GO:0009887	organ morphogenesis	342	46	13.45	2.06E-14
GO:0006936	muscle contraction	342	25	7.31	2.08E-14
GO:0007399	nervous system development	342	55	16.08	1.30E-13
GO:0003007	heart morphogenesis	342	16	4.68	5.30E-13
GO:0048468	cell development	342	47	13.74	1.93E-12

DAY 2

The transient expression T Brachyury, Eomes, is maximal at this time point and quickly fades away in the next time points. These results are comparable to the observations made with mouse ESC in Gaspar's study (sub-chapter 3.1, Day3).

T Brachyury (figure 3.3.3.5) is known as a canonical marker for early mesoderm [21, 44]. Lhx1 is expressed in early mesodermal tissue and later in organ development. Comparing to the mouse data set [21] it is possible to see that the peak expression of T brachyury occurs in day 2, contrarily to what is observed in Gaspar (2012) dataset [44]. Although the figure 3.3.3.5. presents a high expression oscillation for T brachyury.

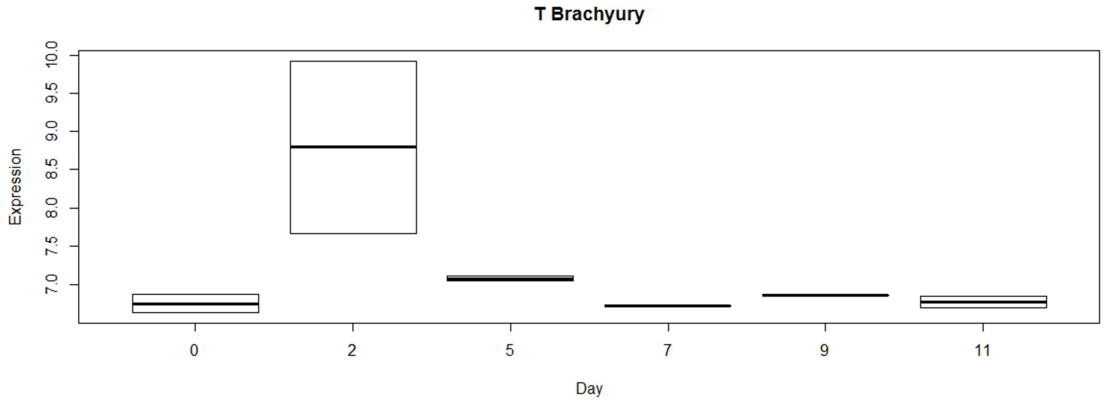


Figure 3.3.3.5. Expression of T Brachyury on Uosaki e*t al*., 2011 data set.

Eomes (figure 3.3.3.6), participating also mesoderm formation, and also controlling gastrulation and trophoblast differentiation in mammals [45], appears to be highly expressed at day2.

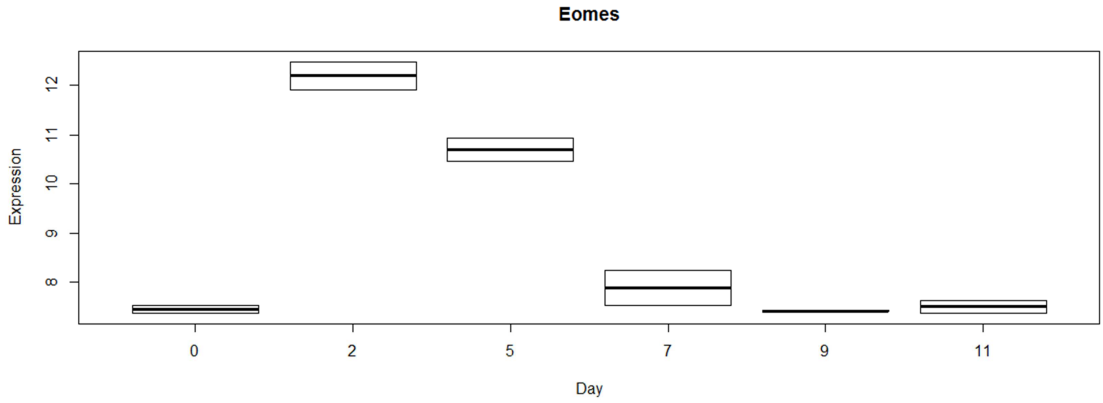


Figure 3.3.3.6. Expression of eomesodermin homolog on Uosaki e*t al*., 2011 data set.

Once again, since the differentiation process of these cells were directed to towards a specific cell fate, in this case cardiac cell fate, they demonstrate a strong biological function linked to heart development and morphogenesis. Additionally, the cardiac differentiation program is also one of the first programs to be initiated in ESc differentiation.

Table 3.3.3.3. Biological functional analysis in Day 2 on Uosaki e*t al*., 2011 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0006936	muscle contraction	154	20	12.99	1.13E-14
GO:0048738	cardiac muscle tissue development	154	14	9.09	1.13E-14
GO:0007507	heart development	154	20	12.99	3.92E-14
GO:0009887	organ morphogenesis	154	32	20.78	6.30E-14
GO:0009888	tissue development	154	32	20.78	7.18E-14
GO:0008015	blood circulation	154	19	12.34	1.19E-12
GO:0007517	muscle organ development	154	19	12.34	2.92E-11
GO:0035051	cardiac cell differentiation	154	9	5.84	2.92E-11
GO:0048729	tissue morphogenesis	154	16	10.39	4.23E-11
GO:0009790	embryo development	154	24	15.58	5.19E-10
GO:0051146	striated muscle cell differentiation	154	9	5.84	5.63E-09
GO:0003007	heart morphogenesis	154	10	6.49	1.46E-08
GO:0060047	heart contraction	154	10	6.49	1.69E-07

DAY 5

Figures 3.3.3.7 and 3.3.3.8 present the expression of Isl1 and Hand1, we can observe a large expression of these genes at day 5 into cardiac differentiation, indicating that they are important in the early stage of heart development and morphogenesis [46, 47]. Indeed, mutations of these genes can lead to several heart disease conditions [47, 56]. These to transcription factors are essential to begin the differentiation of hESc towards a cardiac fate.

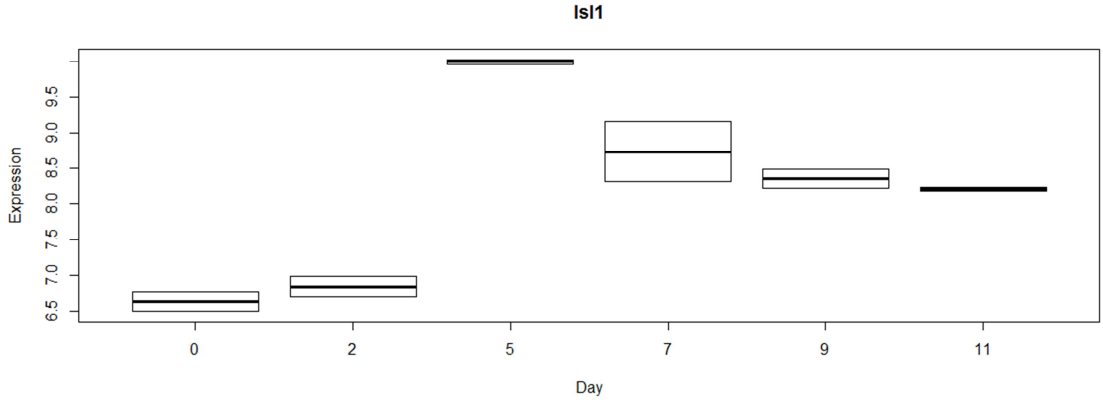


Figure 3.3.3.7. Expression of ISL LIM homeobox 1 on Uosaki e*t al*., 2011 data set.

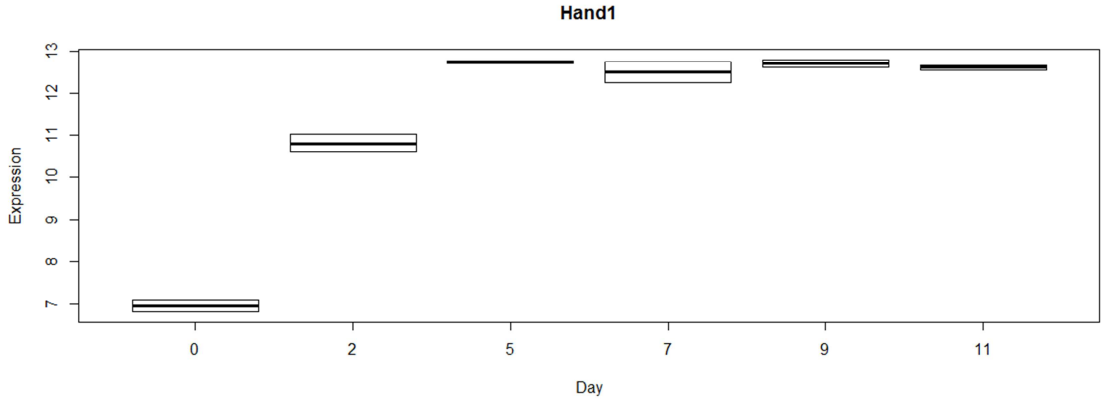


Figure 3.3.3.8. Expression of heart and neural crest derivatives expressed transcript 1 on Uosaki e*t al*., 2011 data set.

Processes related to heart development are quite highly active (table 3.3.3.4) indicating that the cells are specifically going along the desired cardiac differentiation path.

Table 3.3.3.4. Biological functional analysis in Day 5 on Uosaki e*t al*., 2011 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0006936	muscle contraction	231	22	9.52	1.16E-13
GO:0048738	cardiac muscle tissue development	231	14	6.06	3.94E-12
GO:0007517	muscle organ development	231	23	9.96	5.53E-12
GO:0007507	heart development	231	20	8.66	4.45E-11
GO:0008015	blood circulation	231	19	8.23	1.84E-09
GO:0060047	heart contraction	231	13	5.63	5.93E-09
GO:0008016	regulation of heart contraction	231	12	5.19	1.11E-08
GO:0003007	heart morphogenesis	231	11	4.76	3.67E-08
GO:0009887	organ morphogenesis	231	30	12.99	3.67E-08

DAY 7

Gata4 expression increases from day 2 and peaks at day 7 of cardiac differentiation, and could mean this gene could be an important transcription factor that could influence the expression of other genes like, Nkx2-5, Mef2c and Tbx5 [53, 57].

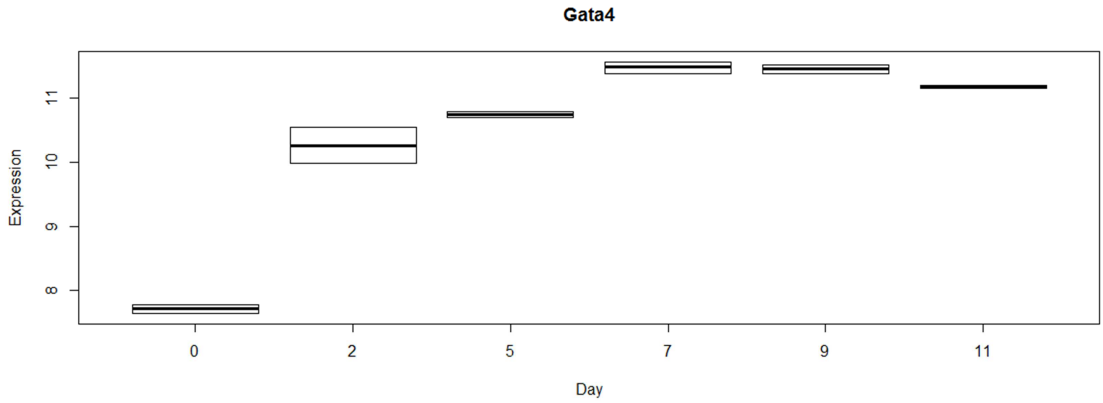


Figure 3.3.3.9. Expression of GATA binding protein 4 on Uosaki e*t al*., 2011 data set.

Gata4 actively participates in heart morphogenesis and development, tissue development, embryo development and vasculature development. In day 7 it is possible to observe that some processes related to anatomical structure morphogenesis start to come up, but still it is strongly expressed genes related to heart development.

Table 3.3.3.5. Biological functional analysis in Day 7 on Hideki *et al*., data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0009887	organ morphogenesis	373	46	12.33	2.74E-11
GO:0007507	heart development	373	25	6.7	2.74E-11
GO:0048646	anatomical structure formation involved in morphogenesis	373	31	8.31	4.96E-11
GO:0048468	cell development	373	48	12.87	8.19E-11
GO:0006928	cellular component movement	373	40	10.72	1.32E-10
GO:0009888	tissue development	373	44	11.8	1.32E-10
GO:0009790	embryo development	373	38	10.19	1.40E-10
GO:0007155	cell adhesion	373	48	12.87	2.28E-10
GO:0007517	muscle organ development	373	25	6.7	9.56E-10

DAY 9

Nkx2-5 (figure 3.3.3.10.), Mef2c (figure 3.3.3.11.) and Tbx5 (3.3.3.12.) exhibit similar expression patterns across all time points, suggesting that they could be interacting with each other, generating a positive feedback for their expression such as they do during embryonic development [53, 57]. They have an ambiguous expression in day 7, but in day 9 they are highly expressed. They essentially participate in muscle organ development, cardiac muscle tissue development, heart development, heart morphogenesis and organ morphogenesis. Nkx2-5 particularly also participates in heart contraction, regulation of heart contraction and striated heart contraction.

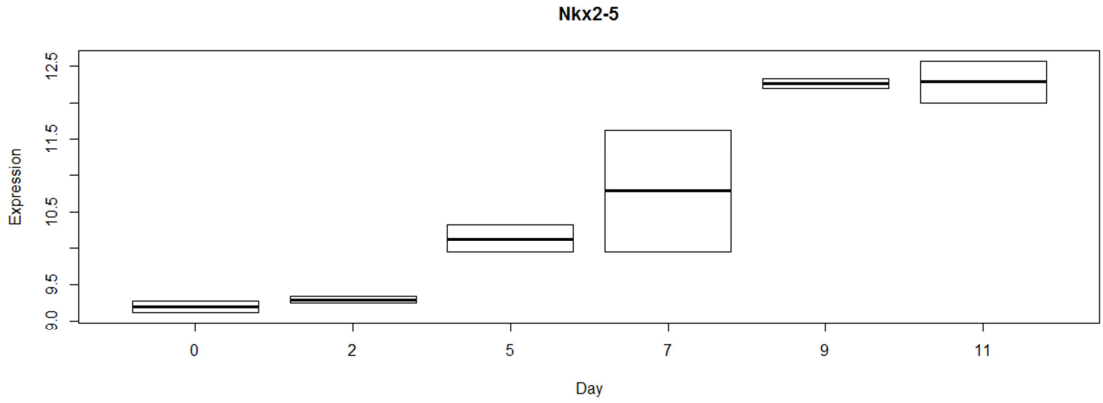


Figure 3.3.3.10. Expression of Nkx2 homeobox 5 on Uosaki e*t al*., 2011 data set.

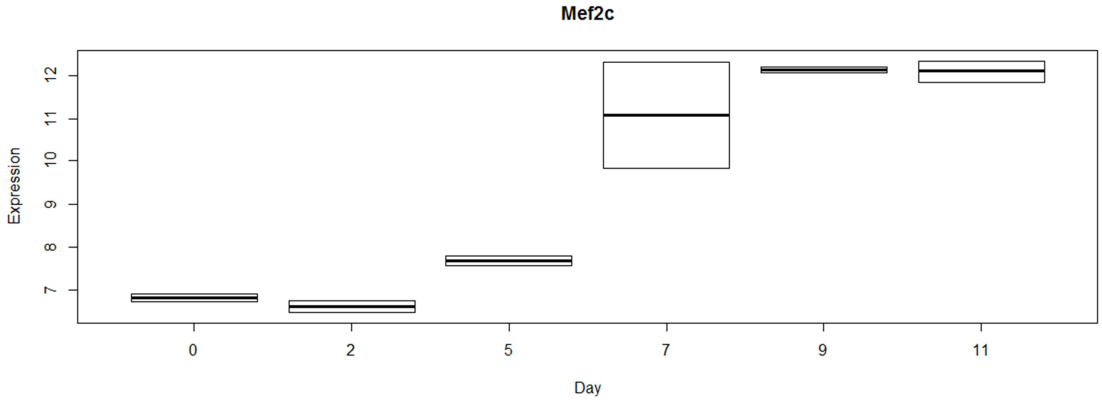


Figure 3.3.3.11. Expression of myocyte enhancer factor 2C on Uosaki e*t al*., 2011 data set.

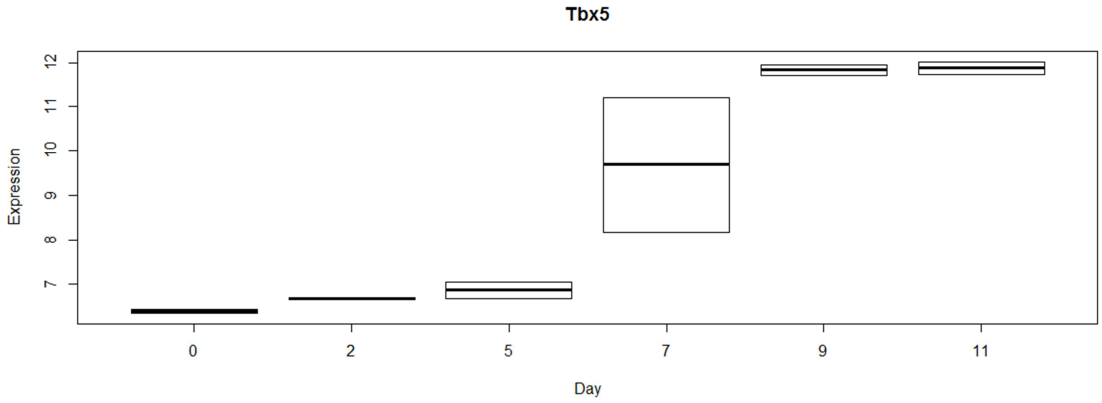


Figure 3.3.3.12. Expression of T-box 5 on Uosaki e*t al*., 2011 data set.

In comparison to day 7 (table 3.3.3.5), in day 9 all processes linked to embryo development and anatomical structure cease to function. This means that day 7 measurements are slightly altered due to some measurement error, since we can see an ambiguous expression in day 7 in these 3 genes, possibly associated with the embryo development and anatomical structure ranking high in table 3.3.3.5.

Table 3.3.3.6. Biological functional analysis in Day 9 on Hideki *et al*., data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0006936	muscle contraction	273	34	12.45	6.74E-26
GO:0007517	muscle organ development	273	36	13.19	5.64E-23
GO:0048738	cardiac muscle tissue development	273	21	7.69	5.36E-22
GO:0007507	heart development	273	31	11.36	3.26E-21
GO:0008015	blood circulation	273	31	11.36	2.70E-20
GO:0009888	tissue development	273	45	16.48	4.71E-16
GO:0060047	heart contraction	273	18	6.59	3.41E-14
GO:0003007	heart morphogenesis	273	16	5.86	1.31E-13
GO:0008016	regulation of heart contraction	273	16	5.86	4.23E-13
GO:0006941	striated muscle contraction	273	14	5.13	1.58E-11
GO:0009887	organ morphogenesis	273	38	13.92	2.04E-11
GO:0051146	striated muscle cell differentiation	273	12	4.4	6.13E-11
GO:0035051	cardiac cell differentiation	273	10	3.66	7.72E-11

DAY 11

Myh7 (figure 3.3.3.13), Actc1 (figure 3.3.3.14) and Tnnt2 (figure 3.3.3.15) are involved in basically every biological process related to heart morphology and development, participating in all processes that are showed in the table 3.3.3.7. These are well known markers for mature cardiac, indicating that the heart development and morphogenesis is coming to an end.

These biomarkers are very important, because they help us understand in the cells that we are looking for, in this case cardiac cells, differentiated [3, 51] or transdifferentiated [4, 52, 53] correctly and if presents a similar expression compared to normal cardiac cells [4, 5].

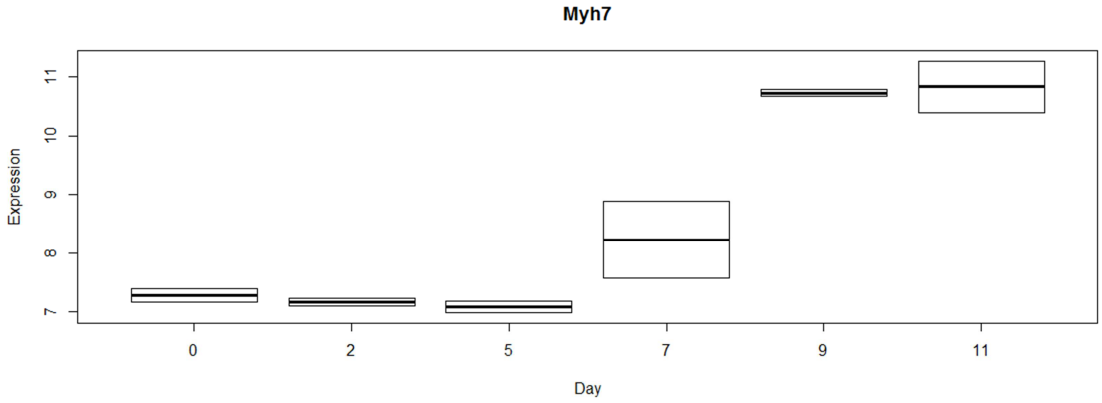


Figure 3.3.3.13. Expression of myosin, heavy chain 7, cardiac muscle, beta on Uosaki e*t al*., 2011 data set.

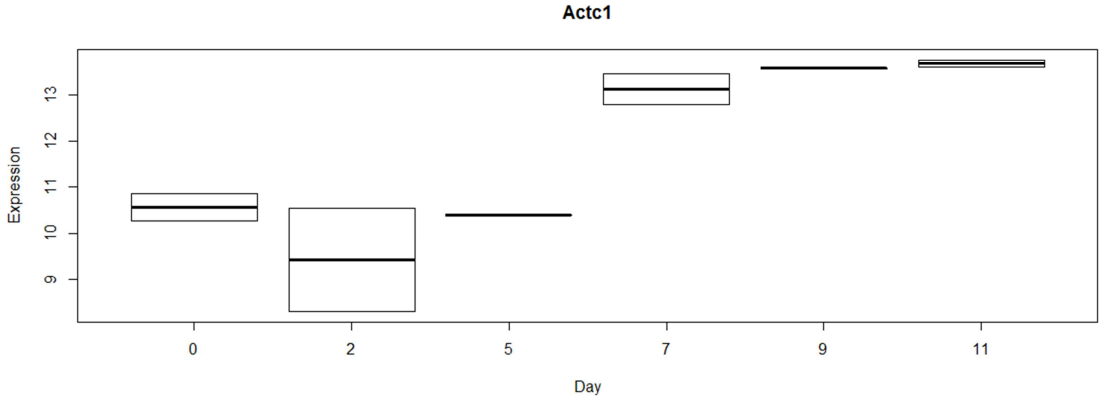


Figure 3.3.3.14. Expression of actin, alpha, cardiac muscle 1 on Uosaki e*t al*., 2011 data set.

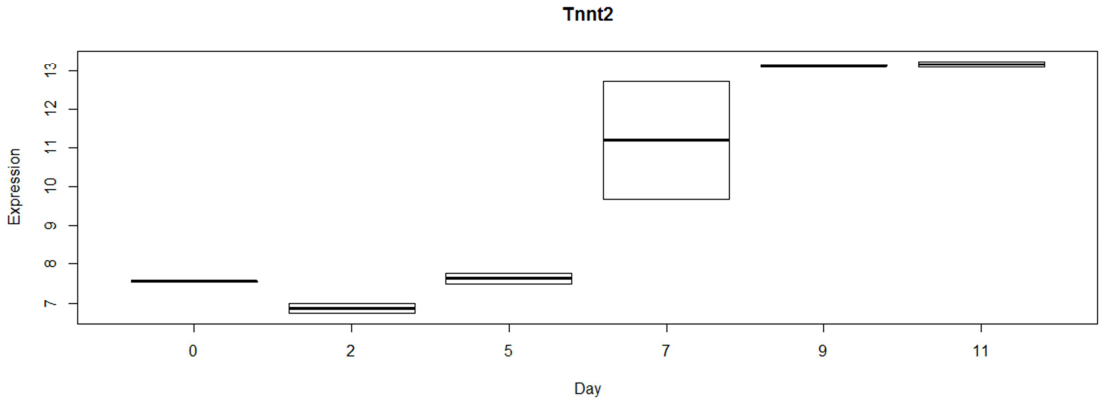


Figure 3.3.3.15. Expression of troponin T type 2 (cardiac) on Uosaki e*t al*., 2011 data set.

The expression of mature cardiomyocyte genes represents the end of heart morphogenesis. In the next days would be expected to see a less important biological function in heart development and morphogenesis. Moreover, it is possible to compare with table 3.1.9, but since these cells were directed to only go forward a heart cell fate, they will not express any other type of biological functions.

Table 3.3.3.7. Biological functional analysis in Day 11 on Uosaki e*t al*., 2011 data set.

GO	Description	Input genes	Positives	%Perc	Adj p-value
GO:0006936	muscle contraction	177	29	16.38	4.06E-25
GO:0007517	muscle organ development	177	30	16.95	4.28E-22
GO:0007507	heart development	177	25	14.12	4.74E-19
GO:0048738	cardiac muscle tissue development	177	17	9.6	4.74E-19
GO:0008015	blood circulation	177	23	12.99	4.90E-16
GO:0009888	tissue development	177	35	19.77	1.12E-14
GO:0060047	heart contraction	177	15	8.47	2.94E-13
GO:0006941	striated muscle contraction	177	13	7.34	1.40E-12
GO:0003007	heart morphogenesis	177	13	7.34	4.60E-12
GO:0008016	regulation of heart contraction	177	13	7.34	9.74E-12
GO:0009887	organ morphogenesis	177	29	16.38	3.67E-10
GO:0042692	muscle cell differentiation	177	12	6.78	8.43E-09
GO:0051146	striated muscle cell differentiation	177	9	5.08	1.67E-08

3.4. COMPARISON OF BIOLOGICAL PROCESS, MOLECULAR FUNCTIONS AND KEGG PATHWAYS BETWEEN CARDIAC DATA SETS AND MOUSE ESC

For this analysis, we used Gaspar (2012) [21], Song (2012) [5], Ieda (2010) [4] and Hideki (2011) [3] data sets (table 3.4.1.). The Gaspar (2012) [21] dataset was used as control to determine which processes are only present in cardiac differentiation (Gaspar *et* al., 2012 VS all other data sets) and cell transdifferentiation into induced cardiomyocytes-like cell (Gaspar *et* al., 2012 VS Hideki *et al.,* 2011) and which are present in all data sets except for the Gaspar *et al*., 2012 dataset.

This approach was used to verify which processes/functions/pathways are exclusive to the heart development and morphology with transdifferentiated cells and confirm if the filtered genes are responsible for obtaining these processes/functions/pathways.

Table 3.4.1. Data sets used for the comparison of the multiple processes

Organism	Authors	Description
Mus musculus	Song *et al*.,2012	Heart repair by reprogramming non-myocytes with cardiac transcription factors
Mus musculus	Ieda *et al.,* 2012	Direct reprogramming of fibroblast into functional cardiomyocytes by defined factors
Homo sapiens	Hideki *et al*., 2011	Time course expression of hiPSc differentiation towards induced cardiomyocytes-like cells
Mus musculus	Gaspar *et al*., 2012	Mouse embryonic stem cell differentiation

3.4.1. BIOLOGICAL PROCESS

Through this heat map it is possible to inspect which biological processes are similar between reprogrammed fibroblasts and hiPSc differentiation toward cardiomyocytes and which biological processes are similar between direct reprogramming of fibroblast. Furthermore, there are many processes that are present in transdifferentiating cells and cells differentiating towards cardiomyocytes cell fate, such as, heart contraction, striate muscle contraction, muscle organ development, heart morphogenesis, and cardiac cell differentiation (figure 3.4.1.1).

One of the main points is the biological processes that are only common to transdifferentiation of cardiac fibroblast into induced cardiomyocytes-like cells, such as, heart morphogenesis and development and cardiac muscle contraction, which all could be linked to the transdifferentiation process. Using Gaspar *et al.,* 2012 data set, it is possible to reject the most common and normal biological processes and make sure that we are only getting the most important biological processes.

Regarding this results, it is possible to infer that there are even few processes common between iCM and differentiation of hiPSc into cardiac cells (only heart morphogenesis).

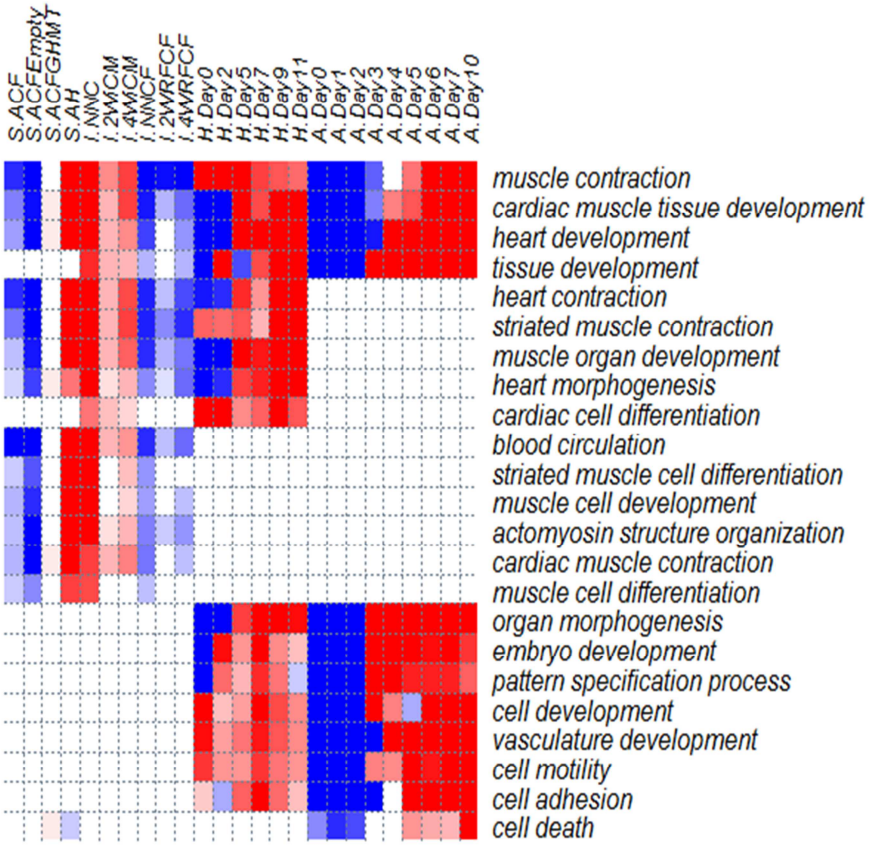


Figure 3.4.1.1. Heat map for biological processes. S – Song *et al*., 2012; I – Ieda *et al*., 2012; H – Hideki *et al*., 2011; A – Gaspar *et al*., 2012.

3.4.2. MOLECULAR FUNCTIONS

The heat map shown in figure 3.4.2.1. represents some selected molecular functions present in the different data sets. Another heat map was obtained (Supplementary figure 1, Annex IV), with a more comprehensive set of molecular functions.

One of the striking things is that it is possible to notice in these molecular functions is that transdifferentiated cells and the normal cells present the same molecular functions, looking like the normal cells. This can suggest that after the transdifferentiation, the cells present a normal molecular function, when compared to the normal cells.

All the molecular binding functions are not exclusive of the transdifferentiated cells, but common processes between transdifferentiated cells are troponin I binding, calcium ion biding, actin filament biding, actin binding, cytoskeletal protein biding, excluding structural molecule activity.

Nevertheless, there are few processes that are not the same between transdifferentiated cells, such as, enzyme inhibitor activity, protein homodimerization, protein dimerization and igG binding (Supplementary figure 1, Annex IV).

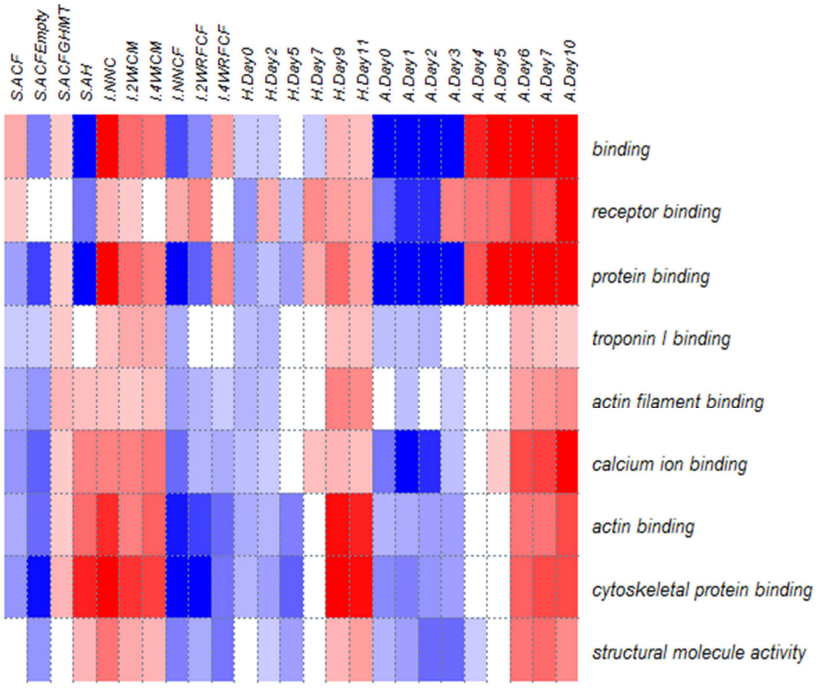


Figure 3.4.2.1. Heat map for molecular function. S – Song *et al*., 2012; I – Ieda *et al*., 2012; H – Hideki *et al*., 2011; A – Gaspar *et al*., 2012.

3.4.3. KEGG PATHWAYS

This heat map (figure 3.4.3.1.) represents some of the main KEGG pathways present in the different data sets. In contrast of what it was observed in the biological process, it is possible to see that some of the KEGG pathways ”up regulated” in the transdifferentiated cell are also present in the cardiac fibroblast, such pathways as, focal adhesion, has this could still be a side effect of the transformation of the cardiac fibroblast into iCM.

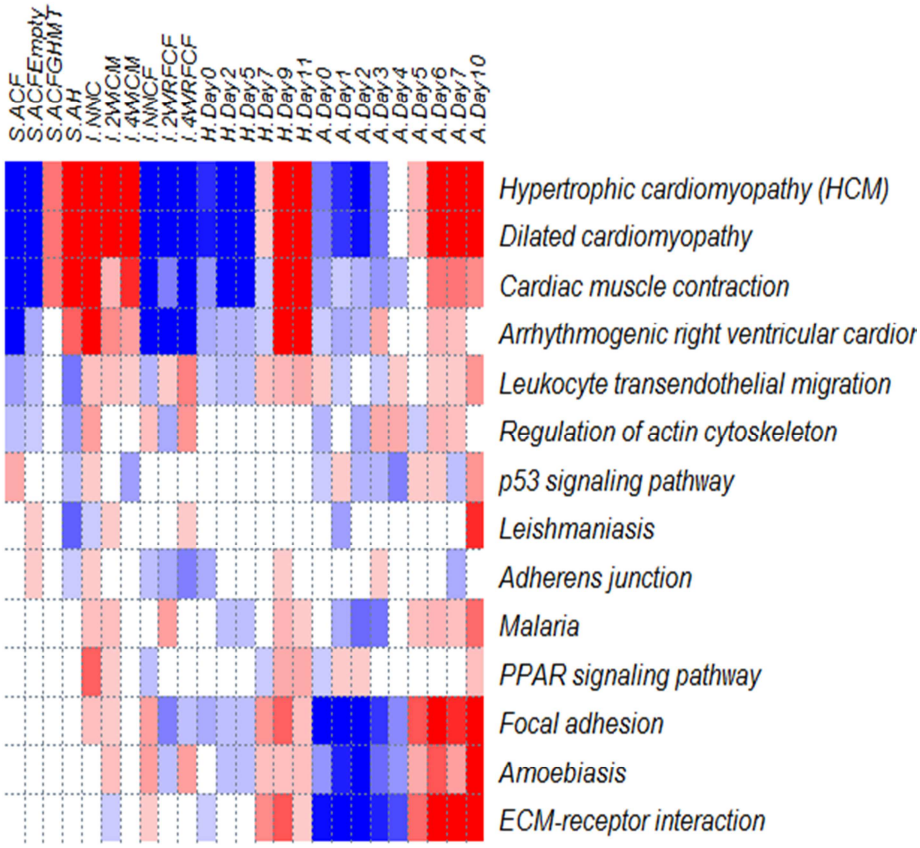


Figure 3.4.3.1. Heat map for KEGG pathways. S – Song *et al*., 2012; I – Ieda *et al*., 2012; H – Hideki *et al*., 2011; A – Gaspar *et al*., 2012.

There are other KEGG pathways that it was not possible to understand why they were coming up, such as, malaria, amoebiasis and leishmaniasis. It is interesting to see that in one experiment (4WiCM), p53 signalling pathway is shut down and that allows the cells to maintain alive even after the transdifferentiation. All normal cells (NNC, AH) and transdifferentiated cells (ACFGHMT, 2WiCM, 4WiCM) show the normal KEGG pathways for cardiac “fate”. PPAR signalling is involved in several processes, such as, regulation expression in liver and skeletal muscle or cell proliferation.

3.5. COMPARISON OF INDIVIDUAL GENE PROFILES BETWEEN CARDIAC DATA SETS AND MOUSE ESC

It was done a comparative differential gene expression analysis for the 4 data set that were used so far for the study of heart development and morphogenesis. They include: differentiation of mouse embryonic stem cells [21]; hiPSc differentiation towards cardiomyocytes [3]; Reprogramming non-cardiomyocytes with cardiac transcription factors [5]; Reprogramming mouse fibroblast into functional cardiomyocytes by defined factors [4](table 3.5.1).

Table 3.5.1. Gene lists used for comparative analysis regarding heart gene expression in early, intermediate and late stage.

Description	Abbreviation
Gaspar *et al*., 2012	mESCDiff / G
Uosaki *et al*., 2011	iPS into iCM / H
Song *et al*., 2012	CF into iCM / S
Ieda *et al*., 2010	Rep into iCM / I

This created a broad basis, which is essential to distinguish between genes that show conserved expression patterns across conditions and genes whose expression changes are specific to certain conditions. This type of study can be of great help to identify some new genes that possibly have not been yet associated to heart development and transformation of cells towards a cardiac cell fate, providing new possible pointers to a future experimental laboratorial analysis.

In these next sub-chapters only the positively expressed genes in the different time stages were analysed, showing the most interesting result. Regarding the negatively expressed gene, none striking outcome was observe during the analysis, so it will be not included in the thesis.

3.5.1. POSITIVE EARLY GENE EXPRESSION

For the comparison of genes positively regulated in the early phase of differentiation, it was only used the only two data sets that have time points (differentiation of mouse embryonic stem cells [21] and hiPSc differentiation towards cardiomyocytes [3]). The experimental conditions used in this comparison were mESc Differentiation assessed at Day0 and Day1, and at Day0 for the iPS differentiation into iCM. It was only used one time point for the iPS into iCM, because the second time point is Day2 and that was not regarded to fall into the early gene expression, since T brachyury is already being expressed.

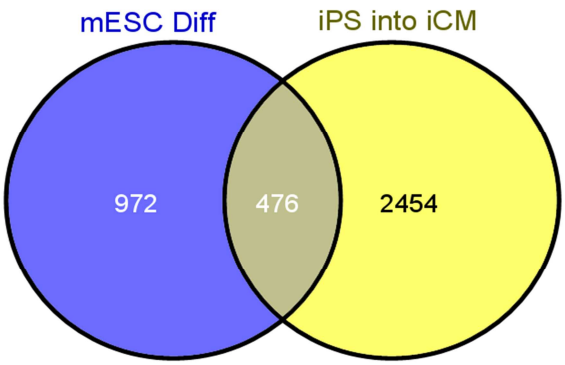


Figure 3.5.1.1. Venn diagram with number of shared genes for the comparative analysis for heart expression.

467 genes have a common positive expression in these two dataset. To narrow this list of genes down, I performed a functional analysis to determine whether any of those genes have already been linked to any type of biological function. Amongst these genes, 140 genes have described biological functions on GO annotations, the remaining 337 genes have no clear function.

From the subgroup of gene with determined function, at least 59 genes participate in regulation of transcription, regulation of gene expression and DNA binding, in which 6 of those participate also in stem cell differentiation and stem cell maintenance (**Nanog**, **Pou5f1**, **Sox2**, **Tet1**, **Rbpj** and **Sall4).**

Table 3.5.1.1. Number of genes present in the overlap with determined biological functions in an early stage

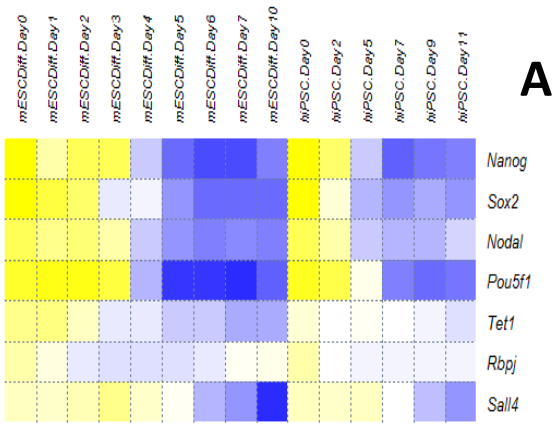
Biological Function	Number of Genes
Regulation of transcription	100
Regulation of gene expression	111
DNA binding	78
Stem cell maintenance	11
Stem cell differentiation	9

Rbpj is a known gene which is linked to the notch signalling and more specifically in heart development. It is also linked to positive regulation of cell proliferation in heart development [58]. The knock out of this gene, triggers the aortic valve degenerative disease, suggesting that this gene mediates Notch signalling, being critically involved in heart homeostasis and disease [58].

Sall4 interacts with Tbx5 to regulate and modulate morphogenesis a patterning in heart [59]. These two genes interact in regulation of heart morphology. When Sall4 is not expressed correctly, this results in abnormal formation of the interventricular groove and disorganized myocardium [59].

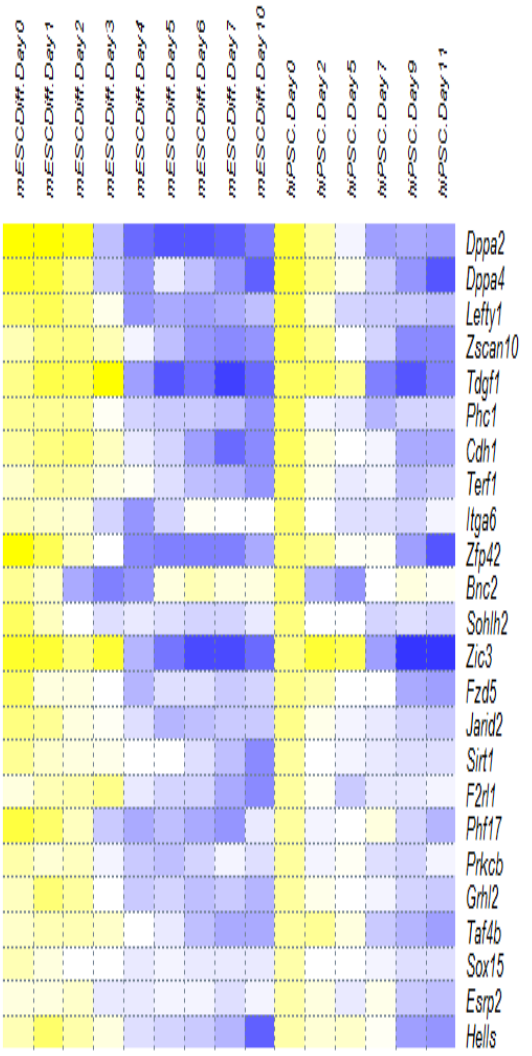
The pattern of expression and biological functions for the most important genes in the early differential stage is presented (figure 3.6.1.1.A and figure 3.6.1.1.B). Notably, a large number of the genes included in figure 3.6.1.1.B have not been associated with stem cell differentiation and stem cell maintenance at least in Gene Ontology. These genes have similar expression, indicating that they are involved in the early differentiation process (early trigger for expression of genes related to heart development) and are involved in maintenance of pluripotency. Thus, the set of genes in figure 3.6.1.1.B provide numerous candidates for stem cell biology. The candidate list provide a rational basis for further experimental studies.

RT	RGE	DNAB	StemCD	StemCM	Gene ID	Gene
■	■	■	■	■	71950	Nanog
■	■	■	■	■	20674	Sox2
■	■		■	■	18119	Nodal
■	■	■	■	■	18999	Pou5f1
■	■	■	■	■	52463	Tet1
■	■	■	■	■	19664	Rbpj
■	■	■	■	■	99377	Sall4



A

RT	RGE	DNAB	StemCD	StemCM	Gene.ID	Gene
■	■		■	■	73703	Dppa2
■	■				73693	Dppa4
■	■				13590	Lefty1
■	■	■	■		332221	Zscan10
■	■				21667	Tdgf1
		■			13619	Phc1
■	■				12550	Cdh1
		■			21749	Terf1
■	■				16403	Itga6
■	■	■			22702	Zfp42
■	■	■			242509	Bnc2
■	■	■			74434	Sohlh2
■	■	■			22773	Zic3
■	■				14367	Fzd5
■	■	■	■		16468	Jarid2
■	■				93759	Sirt1
■	■				14063	F2rl1
■	■				269424	Phf17
■	■				18751	Prkcb
■	■	■			252973	Grhl2
■	■	■			72504	Taf4b
■	■	■			20670	Sox15
■	■	■			15201	Hells



B

Figure 3.5.1.2. Biological function (on the right) and expression (on the left) for some of the obtained genes in the overlap for positive early gene expression. RT – Regulation of Transcription; RGE – Regulation of gene Expression; DNAB – DNA Binding; StemCD – Stem Cell Differentiation; StemCM – Stem Cell Maintenance. A – Well known genes linked to pluripotency; B – Other possible genes linked to maintenance of pluripotency or linked to premature heart gene expression.

3.5.2. POSITIVE INTERMEDIATE GENE EXPRESSION

For this analysis, we used the same data set as the previous study, in which the only modification was in the experimental conditions. The experimental conditions used in this study for the mESC Diff was Day3 and Day4Gaspar (2012)data set [21] and for the iPSc into iCM was Day2 and Day5 Uosaki (2011) data set [3]. This time points were chosen based on the expression of T brachyury that peaks in the transition of stem cells into the different types of tissue and in this case is the transformation into mesoderm.

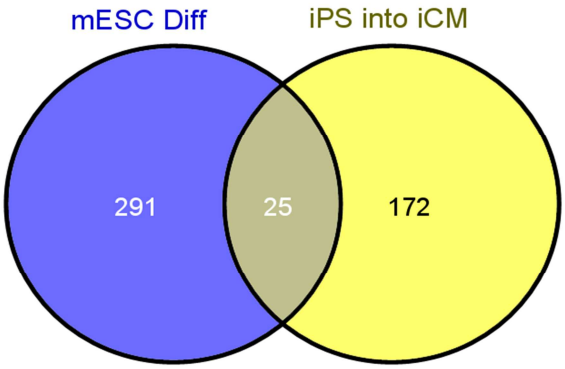


Figure 3.5.2.1. Venn diagram with number of shared genes for the comparative analysis for heart expression.

25 genes were obtained in the overlap between these two dataset. A functional analysis was performed to establish if any of those genes have already been linked to any type of biological function. There are 11 genes with described biological functions, according to GO annotations, in this overlap; the remaining 14 genes have not been described with any type of biological functions regarding these five biological functions (table 3.5.2.1.). From this 11 genes that present described biological functions, 3 of them are associated with stem cell maintenance and differentiation, which are **Eomes, Lin28a** and **Sall4**.

Table 3.5.2.1. Number of genes present in the overlap with determined biological functions in an intermediate stage.

Biological Function	Number of Genes
Regulation of transcription	8
Regulation of gene expression	10
DNA binding	9
Stem cell maintenance	5
Stem cell differentiation	3

T brachyury is considered to be one of the best markers of early mesoderm and is greatly used for keep track in the development of this germ layer, because this gene is present in all mesoderm and down-regulated as these cells are submitted to patterning and specification into the derivate tissues, among them cardiac muscle [44]. This shows that brachyury expression is very brief and cells undergo differentiation do to his expression.

Eomes role in cardiac development is unknown, but this gene is required for early mesoderm pattering and differentiation, some studies that Eomes is expressed in the developing heart, with a great presence in the myocardium [45, 60].

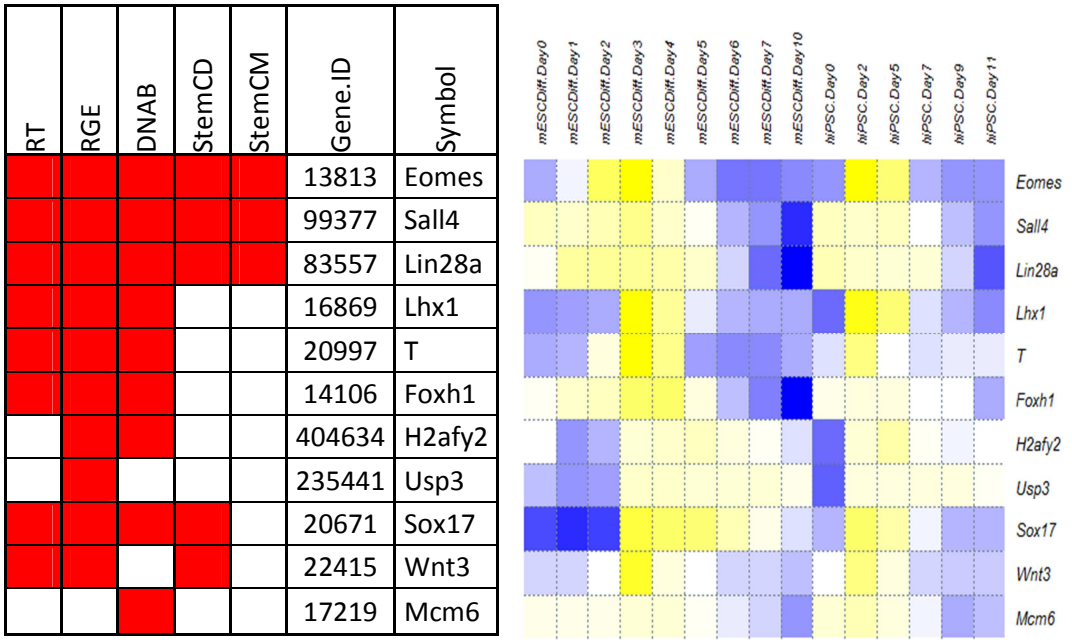


Figure 3.5.2.2. Biological function (on the right) and expression (on the left) for the obtained genes in the overlap for positive intermediate gene expression. RT – Regulation of Transcription; RGE – Regulation of gene Expression; DNAB – DNA Binding; StemCD – Stem Cell Differentiation; StemCM – Stem Cell Maintenance.

Sox17 has established roles in endoderm formation and also participates in mesoderm patterning but not for mesoderm formation. It has an important role in cardiac specification [61]. Studies indicate that the controlled expression of Sox17 during early differentiation affects cardiovascular and more specifically cardiomyocytes differentiation through other gene [62].

In figure 3.5.2.2. is presented the biological functions and expressions for the 11 genes in intermediate stage. Some of the other represented genes could be linked to an intermediate stage of transition or activation of the genes linked to heart development and. Once again this could be proven through experimental procedures.

3.5.3. POSITIVE LATE GENE EXPRESSION

For the positive late gene expression, all microarray studies were used, since all of them had time pointed included for the late phase of differentiation(table 4.5.3.1.). The experimental conditions included were for the mESC Diff Day6 and Day7, (Gaspar 2012) data set [21]; for iPSc into iCM Day7 and Day9, (Uosaki 2011) data set [3]; for CF into iCM the ACFGHMT Sample (Song 2012) data set [5]; Rep into iCM was 2WiCM and 4WiCM (Ieda 2010) data set [4]. These time points were used, because they show significant expression in genes that are known to have high expression in heart development such as **Gata4**, **Tbx5**, **Mef2c** and **Hand2** (figure 3.5.3.2.A) [56].

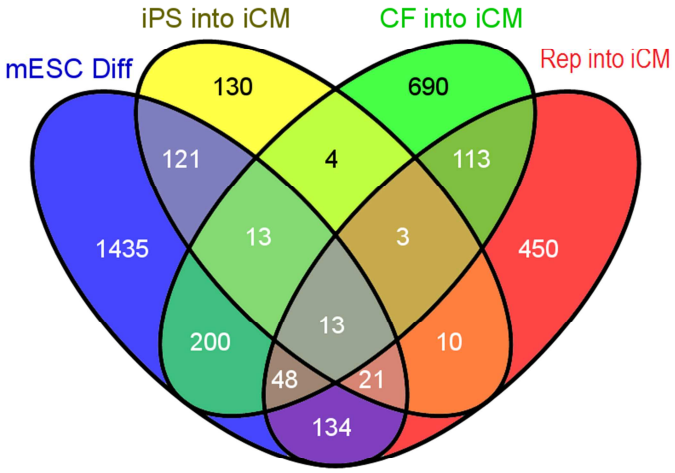


Figure 3.5.3.1. Venn diagram with number of shared genes for the comparative analysis for heart expression.

There are 13 genes present in the overlap of all data sets and 661 genes are present at least in two overlapping data sets. To get a smaller number of genes, it was performed a functional analysis regarding their biological function resorting to GO annotations. So from the 674 only 147 had described biological functions (table 3.5.3.1.). From those 147 genes, 42 presented regulation of transcription, regulation of gene expression and DNA binding functions. Also from those 147 genes, 3 have stem cell differentiation and stem cell

maintenance functions (Tbx3, Bmpr1a and Fzd7) and 3 have only a stem cell differentiation function (Sox5, Pdgfra and Megf10) from the biological processes.

Table 3.5.3.1. Number of genes present in the overlap with determined biological functions in a late stage

Biological Function	Number of Genes
Regulation of transcription	100
Regulation of gene expression	111
DNA binding	78
Stem cell maintenance	11
Stem cell differentiation	9

This four well known genes (figure 3.5.3.2.A) have been extensively studied linked to heart development and cardiac cell fate [4, 5, 11, 12, 59, 63-65]. Since these genes are well known linked to heart, this analysis focus in the identification of possibly new players in heart development.

From the 13 genes that are present in the overlap of all data sets (figure 3.5.3.2.B), only 3 have been associated with selected biological functions: Ankrd1, Igfbp5 and Tgfb2, but only Ankrd1 is specifically linked to heart development.

Ankrd1, also known as Carp, is an early transcription cofactor and early differentiation marker of cardiac myogenesis. It is suggested to have an important role in cardiac muscle function in physiological and pathological conditions [54] and to be involved in biomechanic sensing and regulates cardiac stress responses [55]. In pathological conditions, this gene is up-regulated in the adult heart at end-stage heart failure [54]. Torrado (2005) study suggested that in cardiac Purkinje cells the augmented level of Ankrd1 is due to the high Nkx2-5 gene expression [55].

For the remaining 130 genes, a few genes that could be linked to heart development (figure 3.5.3.2.C), but have not been widely studied linked to heart cell fate. Some of those genes will be focused next.

Smyd1, also known as BOP, is capable of cardiomyocyte differentiation and chamber-specific differentiation in human, and is expressed specifically in cardiac and skeletal muscle precursors and in cardiomyocytes throughout chick and mouse development, beginning

before cardiac differentiation and can also function as a transcriptional repressor [48] This gene is an essential downstream effector of Mef2c in heart development and is also part of a transcriptional cascade involved in development of the anterior heart field [66]. Another remarking observation is mouse embryos that lack Smyd1 gene, which encodes for a transcription repressor and putative histone methyltransferase, die from cardiac abnormalities similar to embryos that have mutations in genes like Isl1, Mef2c, and Hand2 [48, 64, 66]

Hand1 is expressed at a cardiac crescent stage and throughout cardiac development. It exhibits a sided expression pattern, where hand1 is expressed on the left ventricle and hand2 is expressed on the right ventricle [47].This transcription factor regulates cell cycle of cardiomyocytes balancing between cell proliferation and differentiation. It can also promote proliferation is sub-regions of the developing heart, while allowing adjacent myocardial precursors to differentiate. Down-regulation of this gene results in the hypertrophy of cardiomyocytes [47]. Nevertheless, it is still necessary to understand the molecular pathways involved in directing cardiomyocytes proliferation and differentiation regarding this gene and how to generate bigger numbers of functional cardiomyocytes [46]. Hand1 can induce proliferation of cardiomyocytes precursors during heart development, directing progenitor cells towards integrated myocardium [46].

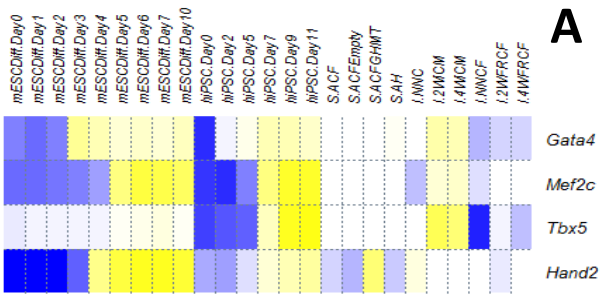
Myocd is a strong co-activator cardiac transcription factor [53] and enhance the cardio-induced effect of established genes related to heart development and cardiac cell fate [57]. This gene alone activates genes involved in several cardiac related processes and inhibits genes involved in non-cardiac processes [53]. The combination of this gene with Tbx5 and Gata4 contributed differently to cardiac gene activation and non-cardiac gene suppression, being the most effective genes for the activation of genes associated to cardiac-related processes [53]. Recent findings by Zhou (2013), the combination of Tbx5, Gata4 and Myocd enhance cardiac gene expression. Combination of Myocd with SFR or Tbx5 can lead to muscle differentiation or cardiomyocyte differentiation [46, 53].

Zfpm2, also known has Fog2, is a co-factor of genes linked to heart development, such as Gata4, modulating transcriptional activity *in vitro* and *in vivo* during heart development and morphogenesis [67, 68]. This gene could be a strong candidate for heart development and his absence is lethal due to consistent defects in heart morphogenesis [67]. Tevosian

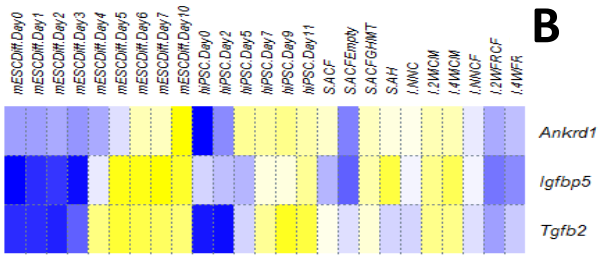
(2000) study established that this gene has an essential role in myocardium and his absence causes multiple morphological abnormalities [67].

Recent studies regarding **Meis1** revealed that the deletion of this gene in mouse cardiomyocytes could lead to an extension of the postnatal cardiomyocytes proliferation window and re-activation of cardiomyocyte mitosis in adult heart with no deleterious effect on cardiac function [69]. Meis1 overexpression in cardiomyocytes decrease neonatal myocyte proliferation and inhibited neonatal heart regeneration [69]. Mahmoud and co-workers (2013) concluded that this gene is a critical transcription regulator of cardiomyocyte proliferation and greatly regulates postnatal cardiomyocytes cell cycle arrest. Taking these results in to account, this gene should be analysed more thoroughly, because it could be an important therapeutic target for heart regenerative medicine.

RGP	RT	DNAB	StemCD	StemCM	Gene	Gene.ID
■	■	■			Gata4	14463
■	■	■			Mef2c	17260
■	■	■			Tbx5	21388
■	■	■			Hand2	15111



RGP	RT	DNAB	StemCD	StemCM	Gene	Gene.ID
■	■	■			Ankrd1	107765
■					Igfbp5	16011
■					Tgfb2	21808



RGP	RT	DNAB	StemCD	StemCM	Gene	Gene.ID
■	■	■			Smyd1	12180
■	■		■	■	Bmpr1a	12166
■	■	■			Hand1	15110
■	■	■			Meis1	17268
■	■	■			Meis2	17536
■	■	■			Myocd	214384
■	■	■			Zfpm2	22762
■	■	■			Snai2	20583
■	■	■	■		Sox5	20678
■	■	■	■	■	Tbx3	21386
■	■	■			Tead1	21676
■	■	■			Lef1	16842
■					Usp3	235441
■	■				Smarcd3	66993
■	■	■			Tbx20	57246
■	■				Myo6	17920

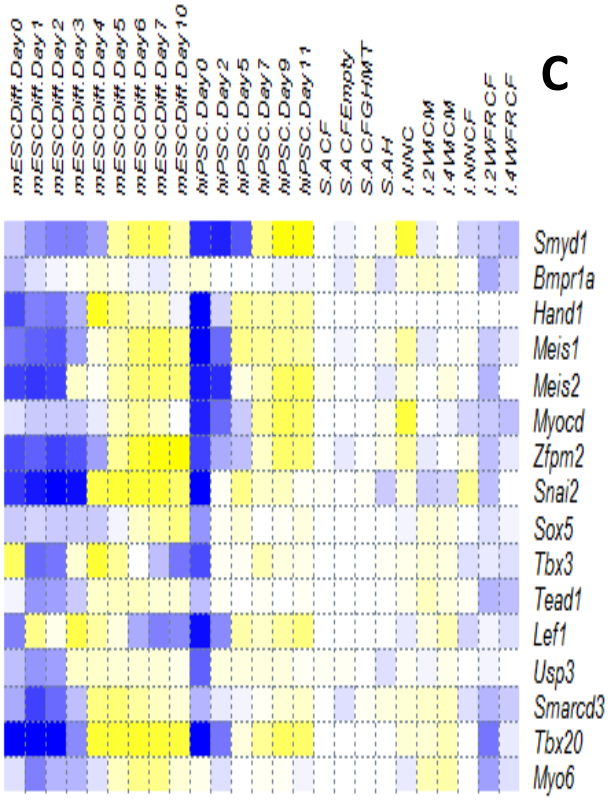


Figure 3.5.3.2. Biological function (on the right) and expression (on the left) for the obtained genes in the overlap for positive late gene expression. RT – Regulation of Transcription; RGE – Regulation of gene Expression; DNAB – DNA Binding; StemCD – Stem Cell Differentiation; Stems – Stem Cell Maintenance. A – Genes greatly related to heart development; B – Genes related to heart development present in the overlap of all data sets with known biological functions; C – Genes that could be linked to heart development, because there are little or none studies linked to heart.

3.6. NETWORK STRUTURAL ANALYSIS

I performed two distinctive types of analysis in this part of work: (i) Correlation network, to link genes according to their expression temporal pattern. In this way, it is possible to obtain sub-cluster of similarly expressed genes; (ii) Interaction Data Network was created based on publically available published protein-protein interaction data. This later allows to analyse how genes from different or similar temporal expression interact, regarding the positive and negative correlation.

3.6.1. CORRELATION NETWORK

The Correlation Network was made from the 296 genes obtained through the comparative differential gene expression analysis of the 4 different data sets that were used in the previous section. To reduce and prioritize the number of genes, I applied a filter using a Pearson correlation coefficient >0.75, so that all linked genes had a correlation higher than 0.75. After applying that filter, the resulting number of genes was still too large (209 genes) to permit a good visual representation in Cytoscape. Thus, for a better interpretation of the results and expression correlation between genes, I have chosen the path in which genes from all temporal expression are represented (figure 3.6.1.1).

The most important genes for each temporal expression here highlighted, so it is possible to see with which direct partners they are highly correlated. Almost of all genes represented in figure 3.6.1.1 are transcription factor, with the exception of the H2afy2, that is not a transcription factor, but it is a regulator of gene expression. Notably, we can observe some sub-clusters, for example, all the first interacting partners of the highlighted genes have a very similar biological function. So we can assume that the first interacting partners of, for example, Meis1 have a similar function to it and of course a very similar temporal expression.

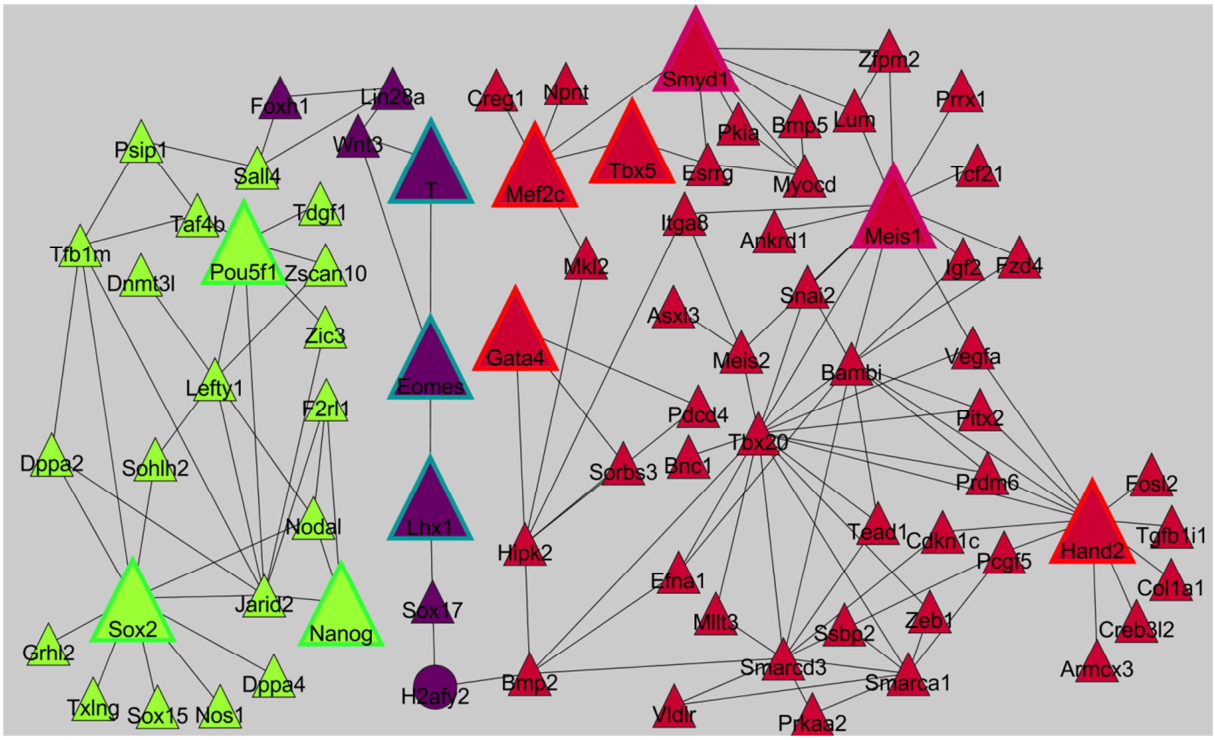


Figure 3.6.1.1. Correlation Network with main gene markers for each temporal expression stage, with a correlation >0.75 in cytoscape. Triangles mean transcription factors; Circle mean non-transcription factors. Green represents early gene expression stage (from day0 to day 2); Purple represents intermediate expression stage (from day 3 to day 5); Red represents late gene expression stage (from day 6, day 7 and day 11).

Although these genes are highly correlated, it is also interesting to see their temporal expression, where these genes show their expression through the different days. Basically, in figure 3.6.1.2, is represented the expression of day 0 in Gaspar (2012) data set [21].

So as the days pass by, it is possible to see the expression variation in the different genes. In the first days, the expression appears in the genes linked to maintenance of pluripotency [1, 2, 13, 41], in the intermediate days it appears in the genes that are responsible for mesoderm cell fate/formation [44, 45] and in a later stage expression shifts to genes linked to heart development and morphogenesis [4, 47, 48, 69].

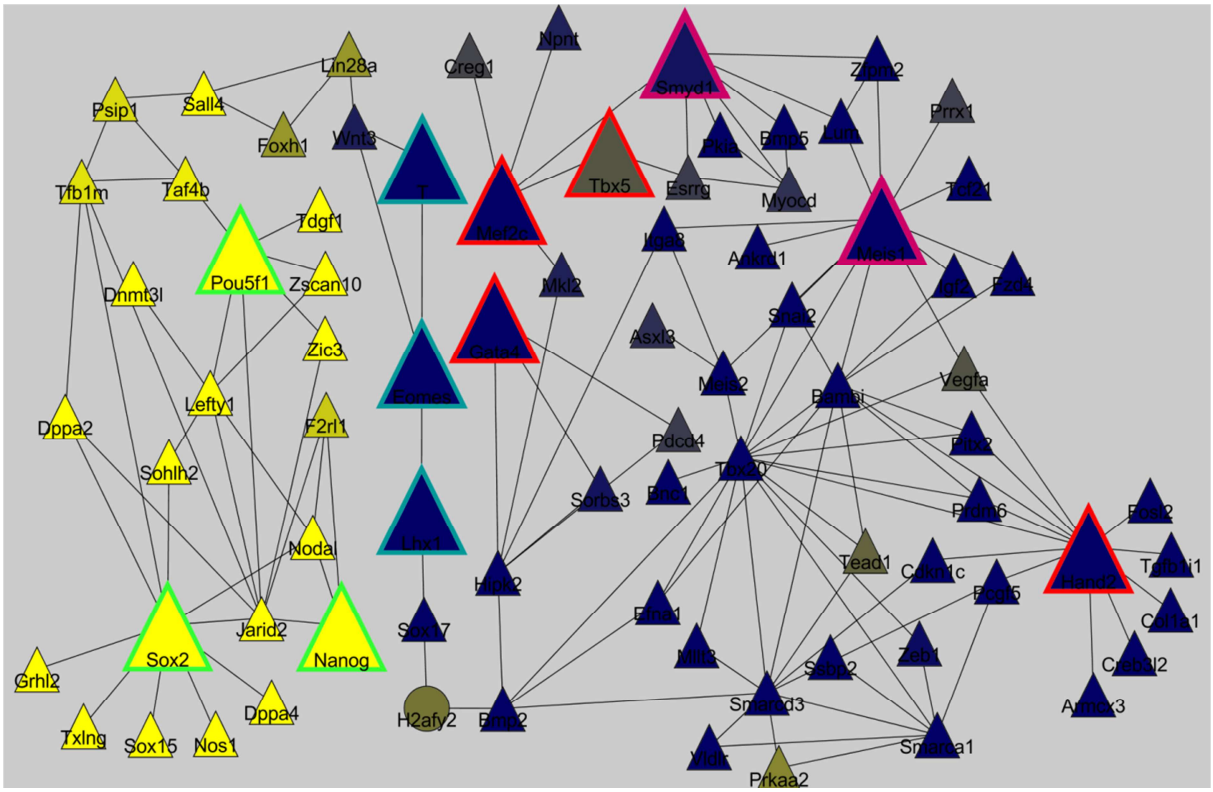


Figure 3.6.1.2. Gene expression network with correlation >0.75 in cytoscape for Day 0 expression values in Gaspar (2012) data set [21]. Triangles mean transcription factors; Circle mean non-transcription factors. Yellow represents genes that are expressed; Blue represents genes that are not expressed.

For a more interactive analysis of the temporal gene expression for these genes, a video (digital attachment I) as created demonstrating gene expression for all days from the Gaspar (2012) data set [21], this video is attached in the digital support. This video allow us to get a better idea of how gene expression correlated at the different days and how expression shifts from the early to late gene expression. Figure 3.6.2.2 gives a little preview of day 0 in the video. As it was referred earlier, all first interacting partners with the major transcription factors, are co-expressing in the different temporal stages, so this sub-clusters appear not only to have the same temporal expression, but they possible can also have similar biological functions.

3.6.2. PPI DATA NETWORK

For a second type of analysis, the objective was to see what type of protein-protein interactions are occurring between the obtained genes from the previous analysis. Protein-Protein interactions information was obtained through Unihi website [70], and it was obtained 232 PPI and 136 genes. For a better visualization, most of the correlation interactions were removed and it was only considered the literature-based protein interactions, since computationally predicted interactions do not provide us much information of how they interact.

Considering the literature-based interactions only, it was obtained a total of 117 genes and 183 PPI. To narrow down even further the number of genes, it was filtered only for the transcription factors, since they are the major responsible for initiate gene expression. Taking into account all these parameters the total number of obtained genes has 52 genes and 66PPI based on literature interactions.

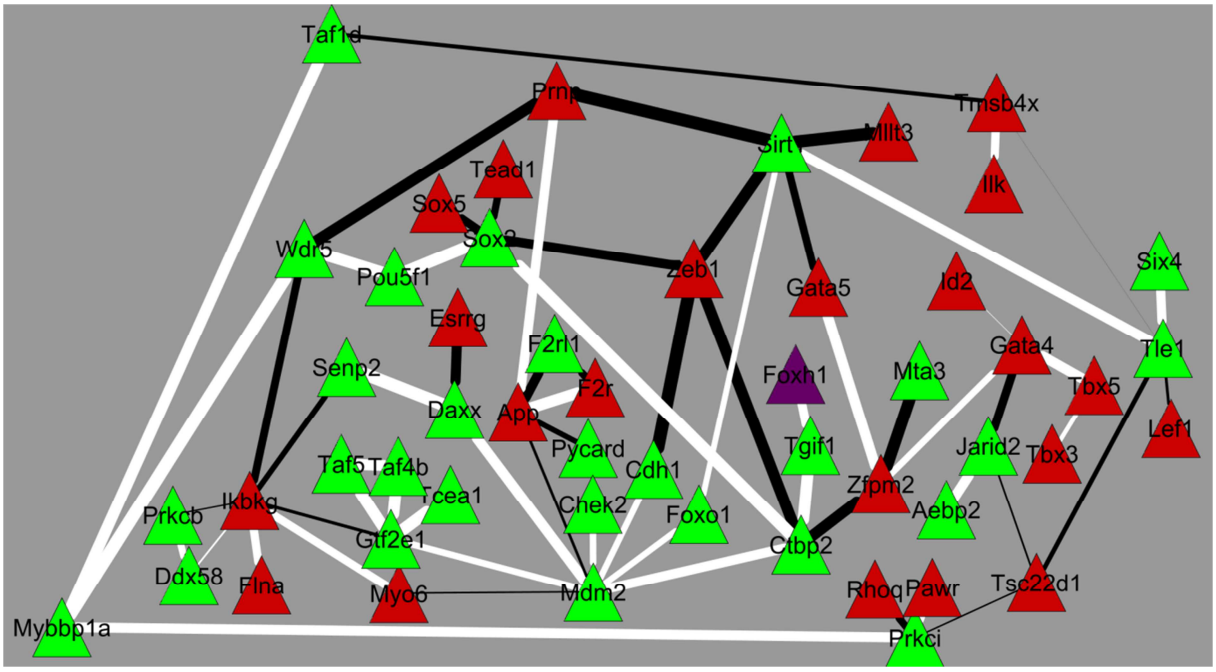


Figure 3.6.2.1. Protein-Protein Interaction Network filtered from publically available interaction studies. Triangles mean transcription factor. Green represents early gene expression stage; Purple represents intermediate gene expression stage; Red represents late gene expression stage. White stands for positive correlation interaction; Black stands for negative correlation interaction.

From these 66 PPI, some interesting interactions came up, particularly regarding inhibitory or repressive regulation, which are represented by black edges. The thicker the edge is, the higher is correlation of expression. One striking outcome is that genes from different temporal expression (early stage represented in green and late stage represented in red) are interacting. Regarding such behaviour it was important to understand what was happening between some genes that are major factors for pluripotency maintenance and major factors for heart development, so we focused in genes such as Jarid1, Zfpm2 and Gata4 that showed more promising results.

Zfpm2 (also known as Fog2) has been under analysis since the importance of this transcript regarding cardiac development is still not very well understood [71]. The protein of Fog-2 is capable of mediating Gata4 biding and repression and it was also a binding site for the transcriptional co-repressor Ctbp2 (C-terminal binding protein 2) [71]. Also the combination of Zfpm2 and Mta factors, maximize the repressive activity over Gata4 [72], so the knock down of the MTA protein expression impairs the ability of Zfpm2 to repress Gata4 activity [72].

Gata4 is a well-known transcription factor highly expressed in cardiac progenitors cells and plays a critical role in regulation of cardiogenesis [4, 51, 52]. After analysing some literature regarding the interactions of Gata4, a very interesting hit was obtain, about the Gata4-Id2 interaction. This protein ectopic or endogenous expression has the capability of inhibiting the transdifferentiation of murine embryonic carcinoma cells into myocytes [73]. The binding of this protein to Gata4 results in the inhibition of this transcription factor, not allowing him to bind with the DNA, thereby blocking the transdifferentiation of cancer cells into cardiac myocytes [73].

Jarid2 (also known as Jmj, Jumonji) is a transcription factor that plays a relevant role in the heart development [74, 75]. This protein contains domains of powerful transcriptional repression, DNA binding and interacts physically with cardiac transcription, such as, Nkx2-5 and Gata4 and represses their abilities to activate target gene expression [74]. Studies have shown that Jmj interacts with the N-terminal region of Gata4, which is the transactivation domain of Gata4. Therefore this protein-protein interaction may lead to inhibition of the transcriptional activity of Gata4, and it may also interfere in the interaction of Gata4 with Nkx2-5 [74].

3.7. GENE COMPARATIVE ANALYSIS WITH STEM CELLS AND CANCER CELLS

To assess the extent and molecular basis of the putative connection between cancer and stem cells, I compared several gene lists associated with stem cells and cancer. The analysis showed that both stem and cancer cells can be share a significant number of common genes, which points to the activation of a common molecular program in both types of cells.

All possible combinations of comparison between the different gene lists were carried out. For each detected overlap, I recorded which genes are present in each data set and which genes are overlapping in the different combinations of these data sets. Genes that are shared between the lists are documented in the supplementary tables. Up- or down-regulated genes from Gaspar 2012 and whether they can also be found in the other gene lists (Supplementary Table 2 and 3, Annex V).

Down Regulated Genes

In this study, genes detected in common with either the list for up-or down-regulated genes collected in the Supplementary table 1 (Annex V). First, list for the down-regulated genes (i.e. those that are down-regulated during differentiation) were compared with the gene lists from the Oct4 RNAi screen, the Cancer Census and the Genetic Association database.

In the supplementary table 2 (Annex V) are genes shared with other lists are represented by 1 and highlighted by green colour green, while genes that are absent in the overlap are represented by 0 and have shown in red colour red. Note that a separate column shows the association with transcriptional regulator activity of gene as determined in Gaspar *et al.,* 2012.

The statistical assessment of the number of shared genes demonstrated that the down-regulated genes are significantly enriched in the two cancer gene sets (Cancer I p= 0.008, N= 9; CancerII: p= 0.002, N= 40) as well as in the Oct4 gene set (p=0.02, N= 8). With the exception of the triple comparison (Down-regulated/Genetic Association Database/Oct4), all comparisons showed significance demonstrating that the number of shared genes were clearly larger than we would expect by chance (see table 3.7.2. and figure 3.7.1. below).

Table 3.7.1. Gene lists used for comparative analysis

Description	Abbreviation
Down	D/Down
Genetic Association Database	CII/CancerII
Oct4 Screen Hits	O/Oct4
Cancer gene Census	CI/CancerI

Table 3.7.2. Significance of common genes in different combinations of gene lists.

Combinations	Hypergeometric test (p-value)
D/CII	0.002
D/O	0.0192
D/CI	0.0077
CII/O	0.0135
CII/CI	0
O/CI	0.023
D/CII/O	0.2659
D/CII/CI	0.0184
D/O/CI	0.0008
CII/O/CI	0.0008
D/CII/O/CI	0.0019

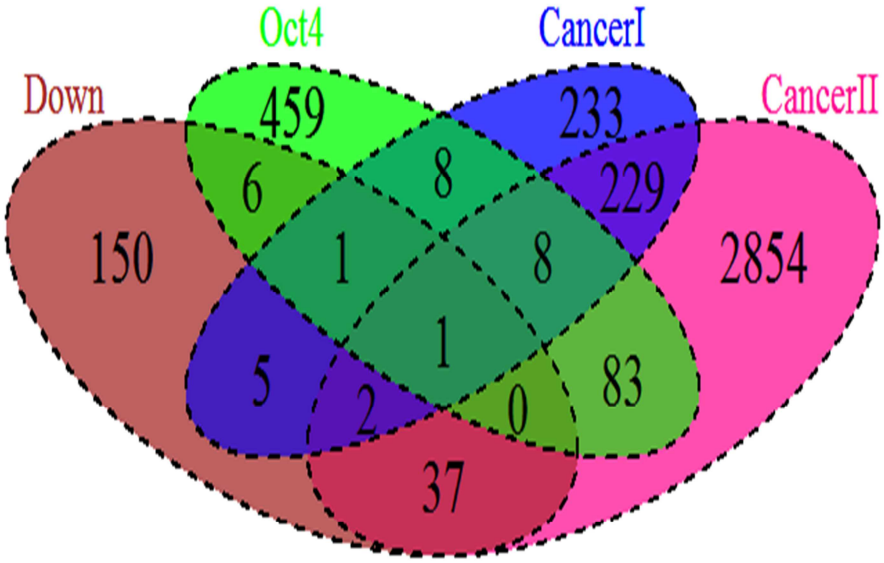


Figure 3.7.1. Venn diagram with number of shared genes for the comparative analysis for genes down-regulated during ESc differentiation.

Several notable genes are shared, but one of the most striking and surprising gene in the intersection of all 4 data set was **TCL1A** (T-cell leukaemia/lymphoma 1A). A preliminary literature review revealed that most of the studies conducted about this gene were linked to tumor associated gene, especially to lymphocytic leukaemia and that overexpression of the **TCL1A** gene in humans has been implicated in the development of mature T-cell leukaemia. Interestingly, however, this genes seems not be linked to ESc so far [76]. In the intersection of the down-regulated genes with CI and CII, we have found two genes: (i) **NF2** is a tumor suppressor that encodes for Merlin, which inhibits mitogenic signalling at or near the plasma membrane. With the accumulation of Merlin in the nucleus, it binds with an ubiquitin ligase and supresses mitotic activity [77]. In the observation by Li (2010) group, many pathogenic NF2 mutations disrupt the capability of Merlin supress tumorgenesis [77]; (ii) **TET1** is highly expressed in undifferentiated ES cells and the expression level of this gene decreases upon ES cell differentiation [78]. Their study not only establishes a role of Tet1 in modulating DNA methylation, but also promotes transcription of pluripotency factors as well as participating in the repression of developmental regulators [78].

Up Regulated Genes

Second list is for the up-regulated genes (i.e. those that are up-regulated during differentiation) were compared with the gene lists from the Oct4 RNAi screen, the Cancer Census and the Genetic Association database. In the supplementary table 3 (Annex I), genes that are present in the overlap are represented by 1 and are marked by green colour, and the genes that are absent in the overlap are represented by 0 and have the red colour. Note that a separate column shows the association with transcriptional regulator activity of gene as determined in Gaspar *et al.*

It was found that the overlap of the up regulated gene list with the other lists (even having statistical significance) tend to be smaller compared to those found for the down regulated genes (table 3.7.4 and figure 3.7.2). The overlap with cancer associative genes is marginally significant (Cancer I p=0.15; Cancer II p=0.04, N=31). No significant overlap with Oct4 screen was detected with only two genes in common (TPM1 & ADAMTS1). This can be expected, since Oct4 hits should be important for stem cell maintenance and thus expressed in ESc, whereas the list of up regulated genes are likely not to be expressed in ESc.

Remarkable, number of genes common to all three gene lists (Up-regulated gene, CancerI and CancerII) is highly significant (p0.0076, n=3).

Table 3.7.3. Gene lists compared.

Description	Abbreviation
Up	U/Up regulated
Genetic Association Database	CII/CancerII
Oct4 Screen Hits	O/Oct4
Cancer gene Census	CI/CancerI

Table 3.7.4. Significance of common genes in different combinations of gene lists.

Combinations	Hypergeometric test (p-value)
U/CII	0.0368
U/O	0.6913
U/CI	0.152
CII/O	0.0135
CII/CI	0
O/CI	0.023
U/CII/O	0.5457
U/CII/CI	0.0076
U/O/CI	0.0503
CII/O/CI	0.0008
U/CII/O/CI	0.0684

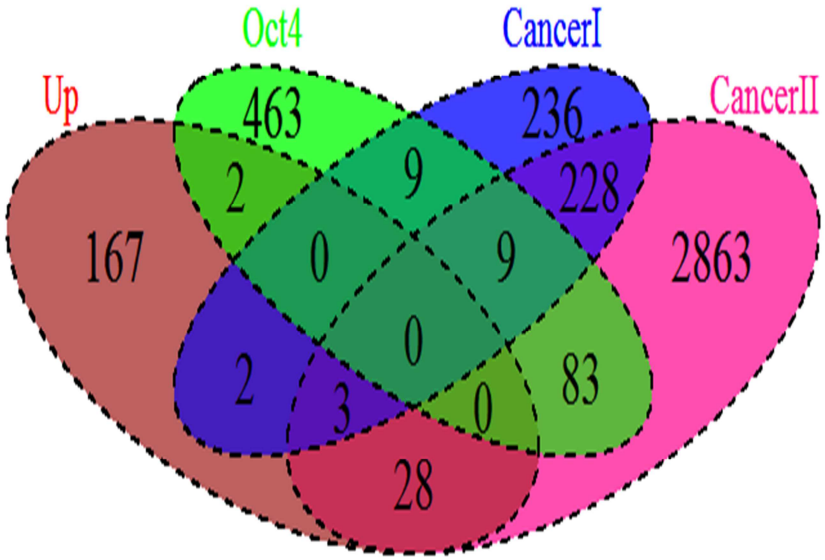


Figure 3.7.2. Venn diagram with number of shared genes for the comparative analysis for genes up-regulated during ESc differentiation.

For this gene list, it was not found any overlap between the 4 different data sets, although it was found 3 genes in the overlap of the Up/CancerI/CancerII: (i) **Ccnd2**, which the encoded protein by this gene belongs to the highly conserved cyclin family, whose members are characterized by a periodicity in protein abundance through the cell cycle. Cyclins function as regulators of CDK kinases, in which these regulate cell cycle; (ii) **Klf6** is a transcriptional factor and functions as a tumor suppressor and it was found that isoforms of this gene are implicated in carcinogenesis and is interestingly an inhibitor of Jun [79]. KLF6 overexpression down regulated c-Jun-dependent transcription and a physical interaction between c-Jun and KLF6 was detected [79]; (iii) **Jun** is a very well-known transcription factor and proto-oncogene in early responsive gene product with unique properties that positively regulate cell proliferation and interacts directly with specific target DNA sequences to regulate gene expression. This gene is mapped to a chromosomal region that is responsible for translocation and deletions in human malignancies [79]. Last examples shows that both tumour suppressors as well as oncogenes can be found to be shared between ESc differentiations associated genes and cancer associated genes. To gain more insight into this putative connection, it would be interesting to check for the single genes whether the recorded genetic mutations lead to a loss or gain of function in cancer.

3.8. "HEARTEXPRESS" INTERACTIVE PLATFORM

Additionally to this work, it was created a web based program (Heart Express: http://heartexpress.sysbiolab.eu/) that provides researchers a direct access to the collected data and enables independent investigations. HeartEXpress is a friendly user and straight forward tool to use, which enables interactive exploration of genome-wide expression data from heart development and morphogenesis, where the user can input their genes of interest and see their correlation with other genes. It also enables visualization of changes in gene expression in the different conditions of the experiments.

The data present in Heart Express was retrieved from public repositories: Gene Expression Omnibus and ArrayExpress, it comprises 38 microarrays from 7 different published papers. It enables visualization of changes in gene expression in the different conditions of the experiments. Currently, it is possible to analyse two types of that, which consist in: (i) Only genes included in all experiments (No missing data) and;(ii) All genes included in more than 40% of microarrays.

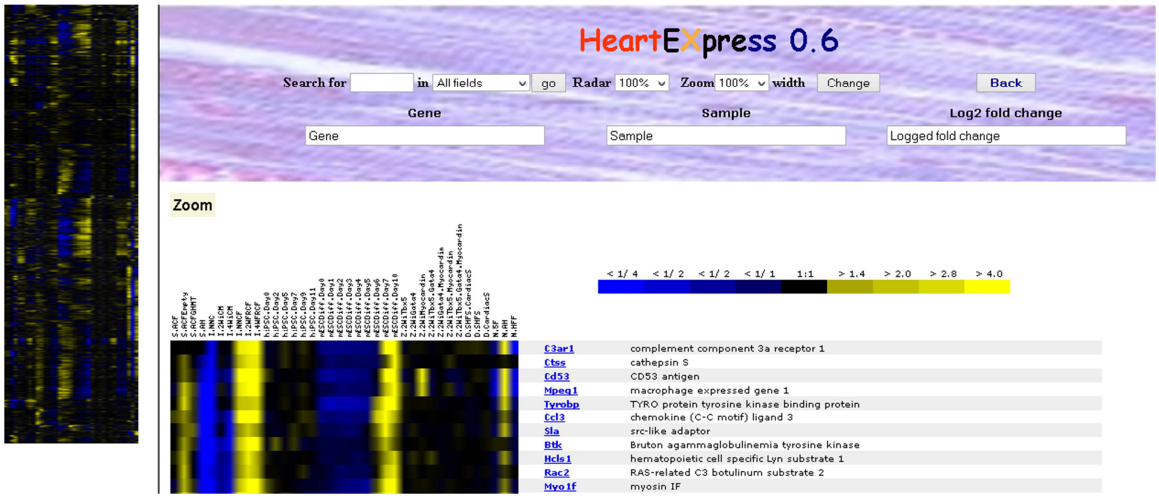


Figure 3.8.1. HeartEXpress user interface [34].

Genes were clustered regarding similarities in the expression changes in order to help in the identification of expression patterns across different conditions. HeartEXpress can be used to examine transcription response of co-expressed genes in order to identify potential regulatory associations.

The user can choose a set of genes, whose expression changes will be displayed. This can be done either by selecting a gene from the miniature heat map, or by inputting a gene symbol or name. Expression changes for each gene are coloured by a blue/yellow gradient, in which decreasing (blue) and increasing (yellow). Black squares represent no differentiation in expression. Graphical view is given as a matrix, for by rows (genes) and columns (experiment conditions). For a more detailed information about the used microarrays the integrated data sets page.

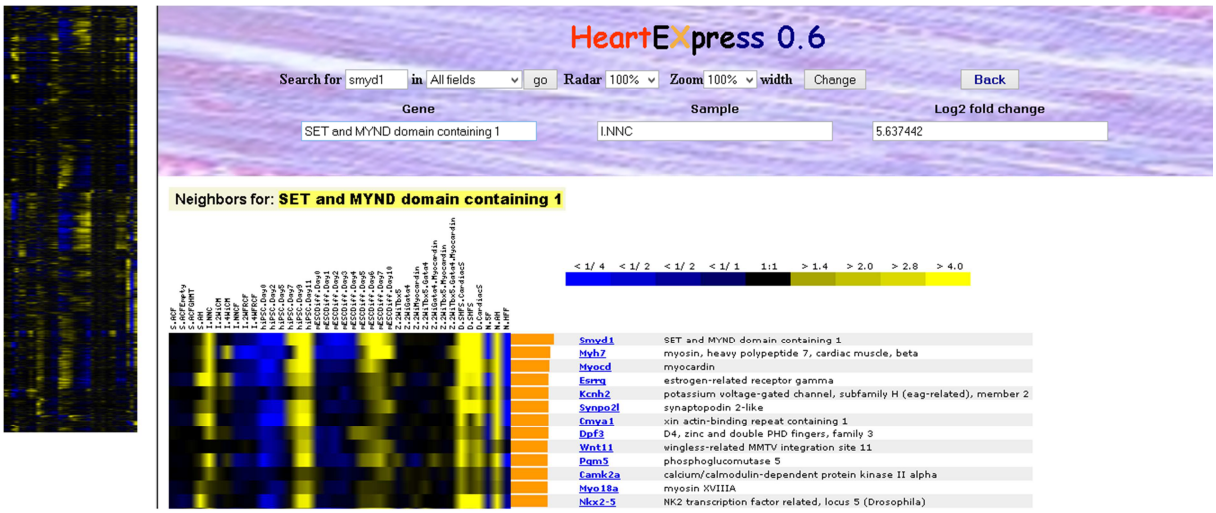


Figure 3.8.2. A mouse click on the displayed expression of a gene will open another page displaying all genes that are co-expressed with the chosen gene [34]. The orange bar indicates spearman correlation coefficient regarding the top gene, which in this case is Smyd1.

Overall, this is an available database online that allows the user an independent and interactive exploration of genes related to heart development. The development of this visualization tool also has the aim of sharing all collected and treated data with the scientific community and assist researchers in the characterization and functional annotation in heart development.

4. CONCLUSION

Since the heart development is a very complex and intricate process with precise spatial and temporal events it was important to perform such type of study, because this was the first study that involves the meta-analysis of different published works regarding heart development and their molecular and genetic events. This study allowed us to do a cross comparison between the different data sets regarding several genes, but mainly the ones responsible for heart development and potential new candidates.

In sub-chapters 3.1, 3.3, 3.5 and 3.7, it was made an extensive analysis of the microarrays used for the study. This study consisted to verify the different gene expression in the different microarrays; (i) General genetic expression of mouse embryonic stem cells – although this microarray is not directly linked to heart development and morphogenesis, it is possible to observe that several of the main transcription factor related to heart development are being expression from day 4, allowing us to determine in a normal development to establish when heart-related genes start to express; (ii) Genetic expression between cardiac fibroblasts and induced cardiomyocytes and differentiation into cardiomyocytes; (iii) A thorough analysis of genes present in each stage of development towards the cardiac cell fate; (iv) Comparative analysis of gene lists associated with stem cells and cancer cells– This is also not directly related to heart development, but it helped us to understand if any of the genes that are associated to cancer or stem cells are also linked to heart development, so in that way it was possible to prevent the use of such genes to avoid the possible formation of teratomas in the development of cardiac cells. The comparative analysis of these data allowed us to get a general view of the different gene expression related to heart development, but most the most important of all, it allowed to confirm that the gene expression presented in the 3.1, 3.3, 3.5 chapters are accurate since they presented the expected expression according to the different papers [3-5, 21] and the biological functions in chapter 3.4 showed processes linked to heart development and morphogenesis.

To get a better insight of the heart development, it was analysed the biological functions, molecular functions and KEGG pathways. To get a smaller list, the genes were filtered for an adjusted p-value of 0.1, to have a higher confidence of the obtained results. For the biological processes, most of the results were linked to heart development. We

observed that heart development starts to occur in day 4-5 in the iPSc and mESc and, the transdifferentiated cells also present heart development and morphogenesis, and additionally, cardiac muscle tissue development and cardiac muscle contraction. This shows that the genes responsible for the development of the heart are present in the transdifferentiated cells, showing that not only the known transcription factors for heart development are present but also new possible candidate genes.

From those filtered genes, they were divided according to their temporal expression and analysed between the possible data sets: (i) Early stage – differentiation of mouse embryonic stem cells [21], day 0 and day 1 vs. hiPSc differentiation towards cardiomyocytes day 0 [3]); (ii) Intermediate stage – differentiation of mouse embryonic stem cells [21], day 3 and 4 vs. hiPSc differentiation towards cardiomyocytes day 2 and day 5 [3]); (iii) Late stage – Mouse embryonic stem cell Differentiation [21], Day 6 and Day 7 vs. iPSc into iCM [3], Day7 and Day9 vs. CF into iCM [5], ACFGHMT vs. Rep into iCM [4], 2WiCM and 4WiCM. With this type of analysis it was possible to see the integral expression of the gene in each temporal stage, allowing to classify the genes into genes that are responsible for maintenance and differentiation of stem cells, genes that are responsible of establishing the cell fate of the differentiating cells and genes that are responsible for inducing the transformation of these cells into the heart cell fate. With the analysis of the late gene expression stage, we could find some interesting genes that could be important for heart development and new pointers for heart regenerative medicine. These new pointers are quite recent and there are few studies of them related to heart development, such is the case of Meis1, Smyd1 and Myocd. There are already some studies regarding Meis1 and Myocd [53, 57, 69], but they are from late 2012 and early 2013, showing to be promising pointers for heart development and regenerative medicine.

For a complete analysis of all obtained results it was created a network structural analysis, to visually examine how the filtered genes correlated between themselves and how strong and relevant were their protein-protein interactions. For the correlation network we could verify that all genes obtained were highly correlated and to diminish the number of present genes, the correlation threshold was raise to ≈ 0.75, reducing from 296 to 209 genes. The temporal gene expression changes showed that they can not only be highly correlated, but they could also be interacting at a protein level, since some sub-clusters present in the correlation network fluctuates in a similar time frame. In the PPI network, it

was required more filtering, since there was still present a lot of genes, so we looked specifically to all transcription factors that had protein-protein interactions already described in literature. With this analysis we were able to find some interesting interactions, which could majorly guide cells to maintain in a progenitor state and proliferate and/or make somatic cells transdifferentiate into cardiac cell fate. The main hits on this study were Jarid2 and Id2, that appear to regulate and represses the activity of Gata4. So the regulation or repression of these 2 genes could also be a main focus in the heart development, allowing for the cardiac cells to proliferate.

The many approaches taken in this work helped to find new pointers for heart development, such as, Meis1 and Smyd, and/or regulate other genes that can influence the expression of well-known transcription factors that are strongly linked to heart development, such as, Gata4. Since this work used microarrays from different studies, the obtained overlapping genes are robust candidates for what we have been looking for, i.e., heart regenerating medicine and cellular therapy.

It would be important to keep performing the same type of work, that was made so far, since the technologies and new studies are coming up every day. Even from the beginning of this year and late 2012, about 3 studies with microarrays were published [6, 52, 53, 57], so it would be important to continue to analyse new data and strengthen the previous obtained data. For a continuous tracking and analysis of the upcoming publications related to heart expression experiments, the results will be threaten and publically displayed in the HeartEXpress website [34]. The continuation of this study would be essential, not only to complement the known information, but also to try to get the best and most efficient combination of genes/transcription factor for an efficient transdifferentiation of somatic cells into iCM and/or for cell therapy to directly transdifferentiate cardiac fibroblasts into cardiomyocytes *in vivo*.

5. REFERENCES

1. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka, S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell, 2007. 131(5): p. 861-72.

2. Yamanaka, S. Strategies and new developments in the generation of patient-specific pluripotent stem cells. Cell Stem Cell, 2007. 1(1): p. 39-49.

3. Uosaki, H., Fukushima, H., Takeuchi, A., Matsuoka, S., Nakatsuji, N., Yamanaka, S., Yamashita, J. K. Efficient and scalable purification of cardiomyocytes from human embryonic and induced pluripotent stem cells by VCAM1 surface expression. PLoS One, 2011. 6(8): p. e23657.

4. Ieda, M., Fu, J. D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B. G., Srivastava, D. Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. Cell, 2010. 142(3): p. 375-386.

5. Song, K., Nam, Y. J., Luo, X., Qi, X., Tan, W., Huang, G. N., Acharya, A., Smith, C. L., Tallquist, M. D., Neilson, E. G., Hill, J. A., Bassel-Duby, R., Olson, E. N.Heart repair by reprogramming non-myocytes with cardiac transcription factors. Nature, 2012. 485(7400): p. 599-604.

6. Inagawa, K. and Ieda, M. Direct Reprogramming of Mouse Fibroblasts into Cardiac Myocytes. J Cardiovasc Transl Res, 2013. 6(1): p. 37-45.

7. Boiani, M. and Scholer, H.R. Regulatory networks in embryo-derived pluripotent stem cells. Nature Reviews Molecular Cell Biology, 2005. 6(11): p. 872-884.

8. Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., Smith, A. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell, 2003. 113(5): p. 643-655.

9. Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., Yamanaka, S. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell, 2003. 113(5): p. 631-642.

10. Ahmed, S. and Khosa, A.N. An Introduction to DNA Technologies and Their Role in Livestock Production: A Review. Journal of Animal and Plant Sciences, 2010. 20(4): p. 305-314.

11. Musunuru, K., Domian, I.J., and Chien, K.R. Stem Cell Models of Cardiac Development and Disease. Annual Review of Cell and Developmental Biology, Vol 26, 2010. 26: p. 667-687.

12. Qian, L., Huang, Y., Spencer, C. I., Foley, A., Vedantham, V., Liu, L., Conway, S. J., Fu, J. D., Srivastava, D. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. Nature, 2012. 485(7400): p. 593-8.

13. Niwa, H. How is pluripotency determined and maintained? Development, 2007. 134(4): p. 635-646.

14. Srivastava, D. Making or breaking the heart: from lineage determination to morphogenesis. Cell, 2006. 126(6): p. 1037-48.

15. Xu, C. Turning cardiac fibroblasts into cardiomyocytes in vivo. Trends Mol Med, 2012. 18(10): p. 575-6.

16. Guimaraes, D.P. and Hainaut, P. TP53: a key gene in human cancer. Biochimie, 2002. 84(1): p. 83-93.

17. Reya, T., Morrison, S. J., Clarke, M. F., Weissman, I. L. Stem cells, cancer, and cancer stem cells. Nature, 2001. 414(6859): p. 105-11.

18. Magee, J.A., Piskounova, E., and Morrison S.J. Cancer stem cells: impact, heterogeneity, and uncertainty. Cancer Cell, 2012. 21(3): p. 283-96.

19. Affymetrix. 2013 [cited 2013 -07-04]; Available from: http://www.affymetrix.com/about_affymetrix/outreach/educator/microarray_curricula.affx.

20. Illumina. 2013 [cited 2013 -07-04]; Available from: http://res.illumina.com/documents/products/datasheets/datasheet_gene_exp_analysis.pdf.

21. Gaspar, J.A., Doss, M. X.,Winkler, J., Wagh, V., Hescheler, J., Kolde, R., Vilo, J., Schulz, H., Sachinidis, A. Gene expression signatures defining fundamental biological processes in pluripotent, early, and late differentiated embryonic stem cells. Stem Cells Dev, 2012. 21(13): p. 2471-84.

22. Johnson, J.MCastle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., Shoemaker, D. D. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science, 2003. 302(5653): p. 2141-4.

23. Russ, J. and Futschik, M.E. Comparison and consolidation of microarray data sets of human tissue expression. BMC Genomics, 2010. 11: p. 305.

24. Schadt, E.E., Edwards, S. W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K. W., Russell, A., Li, G., Cavet, G., Castle, J., McDonagh, P., Kan, Z., Chen, R., Kasarskis, A., Margarint, M., Caceres, R. M., Johnson, J. M., Armour, C. D., Garrett-Engele, P. W., Tsinoremas, N. F., Shoemaker, D. D. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol, 2004. 5(10): p. R73.

25. Shyamsundar, R., Kim, Y. H., Higgins, J. P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, D., Brown, P. O., Pollack, J. R. A DNA microarray survey of gene expression in normal human tissues. Genome Biol, 2005. 6(3): p. R22.

26. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., Hogenesch, J. B. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A, 2004. 101(16): p. 6062-7.

27. Gentleman, R.C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y. C., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., Zhang, J. H. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 2004. 5(10): p. 16.

28. Gautier, L., Cope, L., Bolstad, B. M., Irizarry, R. A. Affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 2004. 20(3): p. 307-15.

29. Smyth, G.K., limma: Linear Models for Microarray Data, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, et al., Editors. 2005, Springer New York. p. 397-420.

30. Cluster-3.0. 2013 [cited 2013 07-04]; Available from: http://bonsai.hgc.jp/~mdehoon/software/cluster/manual/index.html#Top.

31. Saldanha, A.J., Java Treeview--extensible visualization of microarray data. Bioinformatics, 2004. 20(17): p. 3246-8.

32. Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J., FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 2004. 20(4): p. 578-80.

33. Medina, I., Carbonell, J., Pulido, L., Madeira, S. C., Goetz, S., Conesa, A., Tarraga, J., Pascual-Montano, A., Nogales-Cadenas, R., Santoyo, J., Garcia, F., Marba, M., Montaner, D., Dopazo, J. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res, 2010. 38(Web Server issue): p. W210-3.

34. HeartExpress0.6. 2013 [cited 2013 08-28]; Available from: http://heartexpress.sysbiolab.eu/.

35. Cytoscape. 2013 [cited 2013 07-11]; Available from: http://www.cytoscape.org/.

36. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 2003. 13(11): p. 2498-504.

37. Chia, N.Y., Chan, Y. S., Feng, B., Lu, X. Y., Orlov, Y. L., Moreau, D., Kumar, P., Yang, L., Jiang, J. M., Lau, M. S., Huss, M., Soh, B. S., Kraus, P., Li, P., Lufkin, T., Lim, B., Clarke, N. D., Bard, F., Ng, H. H. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature, 2010. 468(7321): p. 316-U207.

38. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M. R. A census of human cancer genes. Nature Reviews Cancer, 2004. 4(3): p. 177-183.

39. Becker, K.G., Barnes, K. C., Bright, T. J., Wang, S. A.The genetic association database. Nat Genet, 2004. 36(5): p. 431-2.

40. Hata, K., Okano, M., Lei, H., Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development, 2002. 129(8): p. 1983-93.

41. Niwa, H., Miyazaki, J., and Smith, A.G., Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet, 2000. 24(4): p. 372-6.

42. Du, J., Chen, T., Zou, X., Xiong, B., Lu, G. Dppa2 knockdown-induced differentiation and repressed proliferation of mouse embryonic stem cells. J Biochem, 2010. 147(2): p. 265-71.

43. Jordanova, A., De Jonghe, P., Boerkoel, C. F., Takashima, H., De Vriendt, E., Ceuterick, C., Martin, J. J., Butler, I. J., Mancias, P., Papasozomenos, SCh., Terespolsky, D., Potocki, L., Brown, C. W., Shy, M., Rita, D. A., Tournev, I., Kremensky, I., Lupski, J. R., Timmerman, V. Mutations in the neurofilament light chain gene (NEFL) cause early onset severe Charcot-Marie-Tooth disease. Brain, 2003. 126(Pt 3): p. 590-7.

44. Doss, M.X., Wagh, V., Schulz, H., Kull, M., Kolde, R., Pfannkuche, K., Nolden, T., Himmelbauer, H., Vilo, J., Hescheler, J., Sachinidis, A. Global transcriptomic analysis of murine embryonic stem cell-derived brachyury (T) cells. Genes Cells, 2010.

45. Russ, A.P., Wattler, S., Colledge, W. H., Aparicio, S. A., Carlton, M. B., Pearce, J. J., Barton, S. C., Surani, M. A., Ryan, K., Nehls, M. C., Wilson, V., Evans, M. J. Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature, 2000. 404(6773): p. 95-9.

46. Risebro, C.A., Smart, N., Dupays, L., Breckenridge, R., Mohun, T. J., Riley, P. R. Hand1 regulates cardiomyocyte proliferation versus differentiation in the developing heart. Development, 2006. 133(22): p. 4595-606.

47. Thattaliyath, B.D., Livi, C. B., Steinhelper, M. E., Toney, G. M., Firulli, A. B. HAND1 and HAND2 are expressed in the adult-rodent heart and are modulated during cardiac hypertrophy. Biochem Biophys Res Commun, 2002. 297(4): p. 870-5.

48. Gottlieb, P.D., Pierce, S. A., Sims, R. J., Yamagishi, H., Weihe, E. K., Harriss, J. V., Maika, S. D., Kuziel, W. A., King, H. L., Olson, E. N., Nakagawa, O., Srivastava, D. Bop encodes a muscle-restricted protein containing MYND and SET domains and is essential for cardiac differentiation and morphogenesis. Nat Genet, 2002. 31(1): p. 25-32.

49. Facucho-Oliveira, J., Bento, M., and Belo, J.A. Ccbe1 expression marks the cardiac and lymphatic progenitor lineages during early stages of mouse development. Int J Dev Biol, 2011. 55(10-12): p. 1007-14.

50. Macdonald, S.T., Bamforth, S. D., Braganca, J., Chen, C. M., Broadbent, C., Schneider, J. E., Schwartz, R. J., Bhattacharya, S. A cell-autonomous role of Cited2 in controlling myocardial and coronary vascular development. Eur Heart J, 2012.

51. Dixon, J.E., Dick, E., Rajamohan, D., Shakesheff, K. M., Denning, C. Directed Differentiation of Human Embryonic Stem Cells to Interrogate the Cardiac Gene Regulatory Network. Molecular Therapy, 2011. 19(9): p. 1695-1703.

52. Nam, Y.J., Song, K., Luo, X., Daniel, E., Lambeth, K., West, K., Hill, J. A., DiMaio, J. M., Baker, L. A., Bassel-Duby, R., Olson, E. N.Reprogramming of human fibroblasts toward a cardiac fate. Proc Natl Acad Sci U S A, 2013. 110(14): p. 5588-93.

53. Zhou, L., Liu, Y., Lu, L., Lu, X., Dixon, R. A.Cardiac gene activation analysis in mammalian non-myoblasic cells by Nkx2-5, Tbx5, Gata4 and Myocd. PLoS One, 2012. 7(10): p. e48028.

54. Arimura, T., Bos, J. M., Sato, A., Kubo, T., Okamoto, H., Nishi, H., Harada, H., Koga, Y., Moulik, M., Doi, Y. L. Towbin, J. A., Ackerman, M. J., Kimura, A. Cardiac ankyrin repeat protein gene (ANKRD1) mutations in hypertrophic cardiomyopathy. J Am Coll Cardiol, 2009. 54(4): p. 334-42.

55. Torrado, M., Nespereira, B., Lopez, E., Centeno, A., Castro-Beiras, A., Mikhailov, A. T. ANKRD1 specifically binds CASQ2 in heart extracts and both proteins are co-enriched in piglet cardiac Purkinje cells. J Mol Cell Cardiol, 2005. 38(2): p. 353-65.

56. Friedrich, F.W., Dilanian, G., Khattar, P., Juhr, D., Gueneau, L., Charron, P., Fressart, V., Vilquin, J. T., Isnard, R., Gouya, L., Richard, P., Hammoudi, N., Komajda, M., Bonne, G., Eschenhagen, T., Dubourg, O., Villard, E., Carrier, L. A novel genetic variant in the transcription factor Islet-1 exerts gain of function on myocyte enhancer factor 2C promoter activity. Eur J Heart Fail, 2013. 15(3): p. 267-76.

57. Christoforou, N., Chellappan, M., Adler, A. F., Kirkton, R. D., Wu, T., Addis, R., Bursac, N., Leong, K. W. Transcription Factors MYOCD, SRF, Mesp1 and SMARCD3 Enhance the Cardio-Inducing Effect of GATA4, TBX5, and MEF2C during Direct Cellular Reprogramming. PLoS One, 2013. 8(5): p. e63577.

58. Li, Z., Feng, L., Wang, C. M., Zheng, Q. J., Zhao, B. J., Yi, W., Zhang, J. Z., Wang, Y. M., Guo, H. T., Yi, D. H., Han, H. Deletion of RBP-J in adult mice leads to the onset of aortic valve degenerative diseases. Mol Biol Rep, 2012. 39(4): p. 3837-45.

59. Koshiba-Takeuchi, K., Takeuchi, J. K., Arruda, E. P., Kathiriya, I. S., Mo, R., Hui, C. C., Srivastava, D., Bruneau, B. G. Cooperative and antagonistic interactions between Sall4 and Tbx5 pattern the mouse limb and heart. Nat Genet, 2006. 38(2): p. 175-83.

60. Ryan, K., Russ, A. P., Levy, R. J., Wehr, D. J., You, J., Easterday, M. C. Modulation of eomes activity alters the size of the developing heart: implications for in utero cardiac gene therapy. Hum Gene Ther, 2004. 15(9): p. 842-55.

61. Liu, Y., Asakura, M., Inoue, H., Nakamura, T., Sano, M., Niu, Z., Chen, M., Schwartz, R. J., Schneider, M. D. Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells. Proc Natl Acad Sci U S A, 2007. 104(10): p. 3859-64.

62. Hoxha, E., Lambers, E., Wasserstrom, J. A., Mackie, A., Ramirez, V., Abramova, T., Verma, S. K., Krishnamurthy, P., Kishore, R. Elucidation of a novel pathway through which HDAC1 controls cardiomyocyte differentiation through expression of SOX-17 and BMP2. PLoS One, 2012. 7(9): p. e45046.

63. Dierickx, P., Doevendans, P. A., Geijsen, N., van Laake, L. W. Embryonic template-based generation and purification of pluripotent stem cell-derived cardiomyocytes for heart repair. J Cardiovasc Transl Res, 2012. 5(5): p. 566-80.

64. Protze, S., Khattak, S., Poulet, C., Lindemann, D., Tanaka, E. M., Ravens, U. A new approach to transcription factor screening for reprogramming of fibroblasts to cardiomyocyte-like cells. J Mol Cell Cardiol, 2012. 53(3): p. 323-32.

65. Olson, E.N. Gene regulatory networks in the evolution and development of the heart. Science, 2006. 313(5795): p. 1922-1927.

66. Phan, D., Rasmussen, T. L., Nakagawa, O., McAnally, J., Gottlieb, P. D., Tucker, P. W., Richardson, J. A., Bassel-Duby, R., Olson, E. N. BOP, a regulator of right ventricular heart development, is a direct transcriptional target of MEF2C in the developing heart. Development, 2005. 132(11): p. 2669-78.

67. Tevosian, S.G., Deconinck, A. E., Tanaka, M., Schinke, M., Litovsky, S. H., Izumo, S., Fujiwara, Y., Orkin, S. H. FOG-2, a cofactor for GATA transcription factors, is essential for heart morphogenesis and development of coronary vessels from epicardium. Cell, 2000. 101(7): p. 729-39.

68. Di Felice, V. and Zummo, G., Tetralogy of fallot as a model to study cardiac progenitor cell migration and differentiation during heart development. Trends Cardiovasc Med, 2009. 19(4): p. 130-5.

69. Mahmoud, A.I., Kocabas, F., Muralidhar, S. A., Kimura, W., Koura, A. S., Thet, S., Porrello, E. R., Sadek, H. A. Meis1 regulates postnatal cardiomyocyte cell cycle arrest. Nature, 2013. 497(7448): p. 249-53.

70. Kalathur, R.K.R. and Futschik, M.E. Unihi. 2013 [cited 2013 07-24]; Available from: http://unihi.org/.

71. Svensson, E.C., Wilk, J., Dale, R., Modrell, M. The role of the transcriptional co-repressor FOG-2 in cardiac development, in Cardiovascular Development and Congenital Malformations, M. Artman, et al., Editors. 2005, Blackwell Publishing.

72. Roche, A.E., Bassett, B. J., Samant, S. A., Hong, W., Blobel, G. A., Svensson, E. C. The zinc finger and C-terminal domains of MTA proteins are required for FOG-2-mediated transcriptional repression via the NuRD complex. J Mol Cell Cardiol, 2008. 44(2): p. 352-360.

73. Ding, B., Liu, C. J., Huang, Y., Yu, J., Kong, W. H., Lengyel, P. p204 protein overcomes the inhibition of the differentiation of P19 murine embryonal carcinoma cells to beating cardiac myocytes by id proteins. Journal of Biological Chemistry, 2006. 281(21): p. 14893-14906.

74. Kim, T., Chen, J. Q., Sadoshima, J., Lee, Y. Jumonji represses atrial natriuretic factor gene expression by inhibiting transcriptional activities of cardiac transcription factors. Molecular and Cellular Biology, 2004. 24(23): p. 10151-10160.

75. Lee, Y., Song, A. J., Baker, R., Micales, B., Conway, S. J., Lyons, G. E. Jumonji, a nuclear protein that is necessary for normal heart development. Circ Res, 2000. 86(9): p. 932-938.

76. Virgilio, L., Lazzeri, C., Bichi, R., Nibu, K., Narducci, M. G., Russo, G., Rothstein, J. L., Croce, C. M. Deregulated expression of TCL1 causes T cell leukemia in mice. Proc Natl Acad Sci U S A, 1998. 95(7): p. 3885-9.

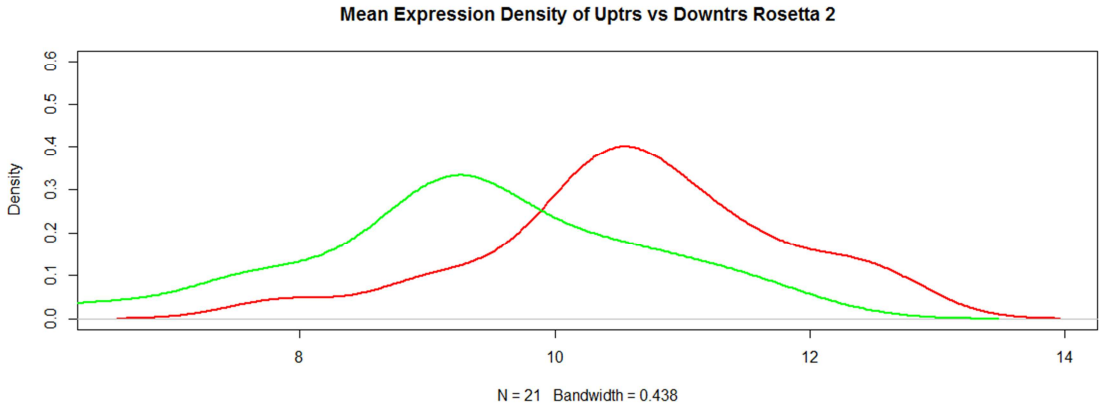
77. Li, W., You, L., Cooper, J., Schiavon, G., Pepe-Caprio, A., Zhou, L., Ishii, R., Giovannini, M., Hanemann, C. O., Long, S. B., Erdjument-Bromage, H., Zhou, P., Tempst, P., Giancotti, F. G. Merlin/NF2 suppresses tumorigenesis by inhibiting the E3 ubiquitin ligase CRL4(DCAF1) in the nucleus. Cell, 2010. 140(4): p. 477-90.

78. Wu, H., D'Alessio, A. C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y. E., Zhang, Y. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature, 2011. 473(7347): p. 389-93.

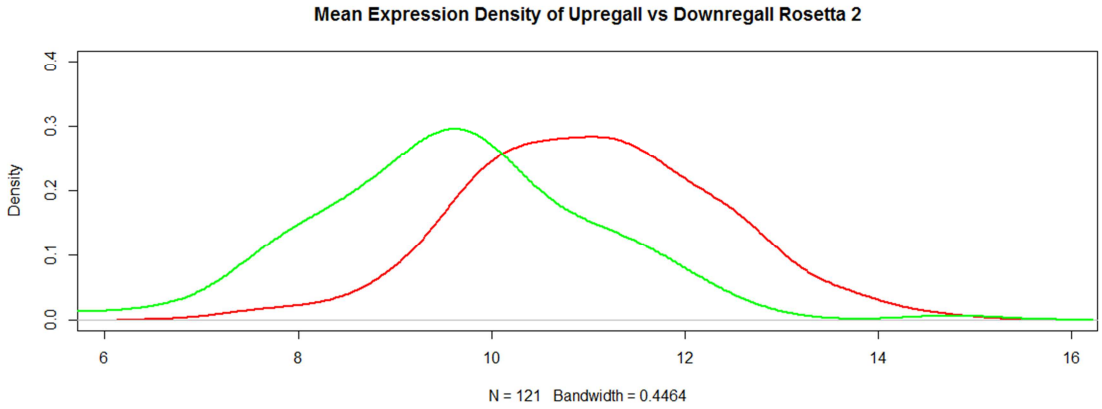
79. Slavin, D.A., Koritschoner, N. P., Prieto, C. C., Lopez-Diaz, F. J., Chatton, B., Bocco, J. L. A new role for the Kruppel-like transcription factor KLF6 as an inhibitor of c-Jun proto-oncoprotein function. Oncogene, 2004. 23(50): p. 8196-205.

ANNEX

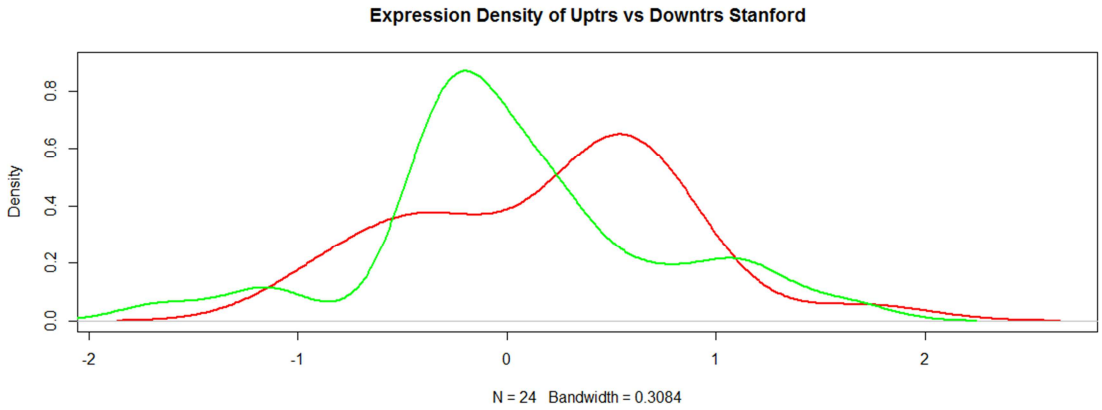
ANNEX I



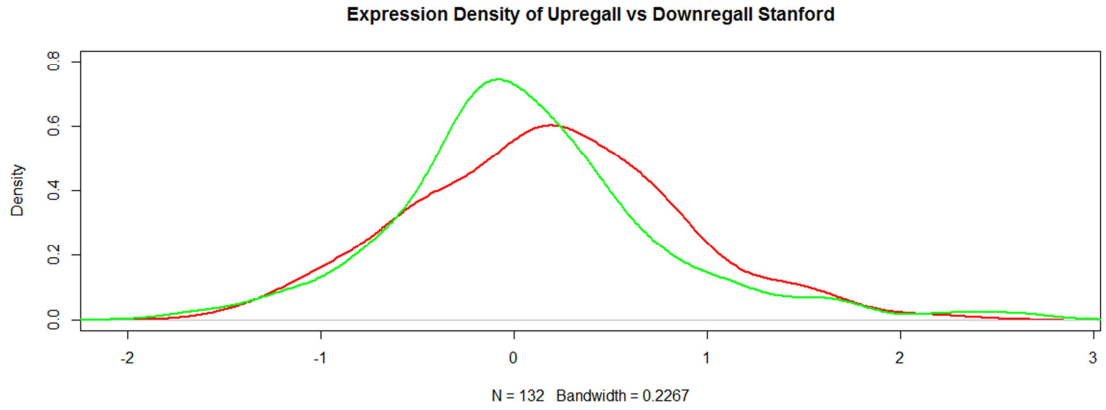
Supplementary figure 1. Distribution of mean expression i of transcription factor that are up regulated vs. down regulated Rosetta 2.



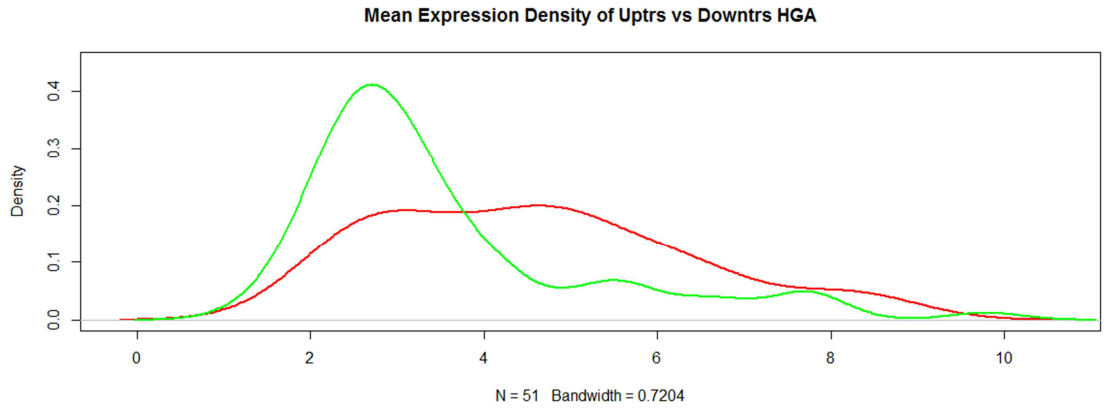
Supplementary figure 2. Mean expression in density of all genes that are up regulated vs. down regulated Rosetta 2.



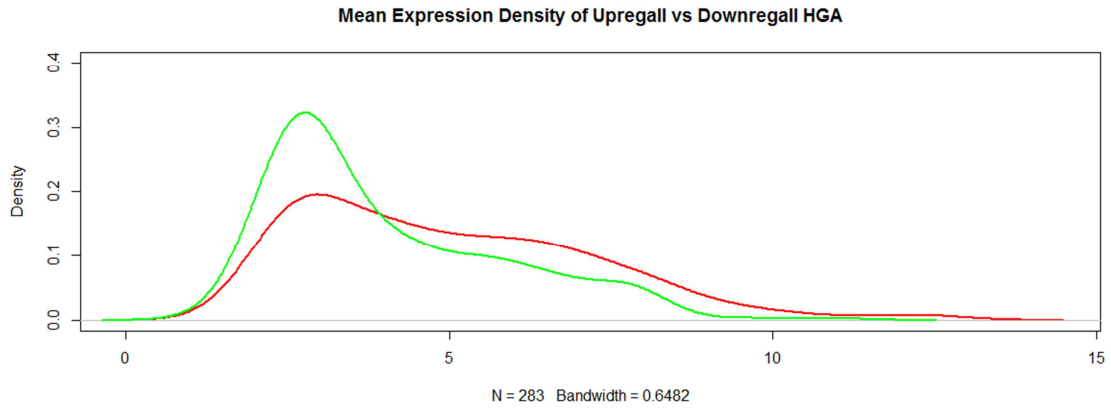
Supplementary figure 3. Mean expression in density of transcription factor that are up regulated vs. down regulated Stanford.



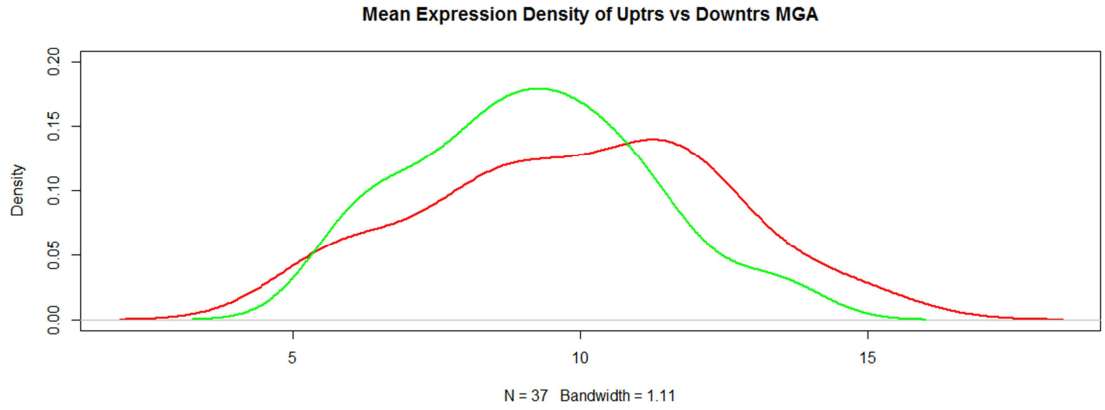
Supplementary figure 4. Mean expression in density of genes that are up regulated vs. down regulated Stanford.



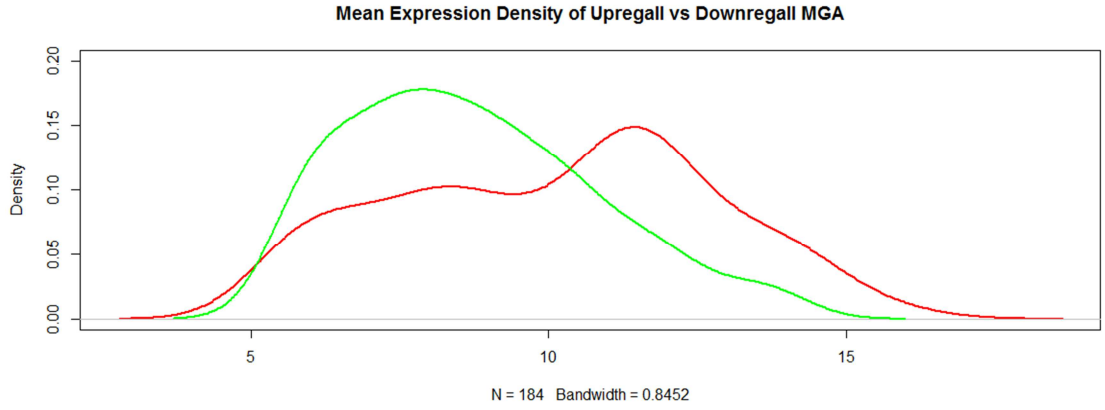
Supplementary figure 5. Mean expression in density of transcription factor that are up vs. down regulated Human Gene Atlas data set



Supplementary figure 6. Mean expression in density of genes that are up vs. down regulated Human Gene Atlas data set

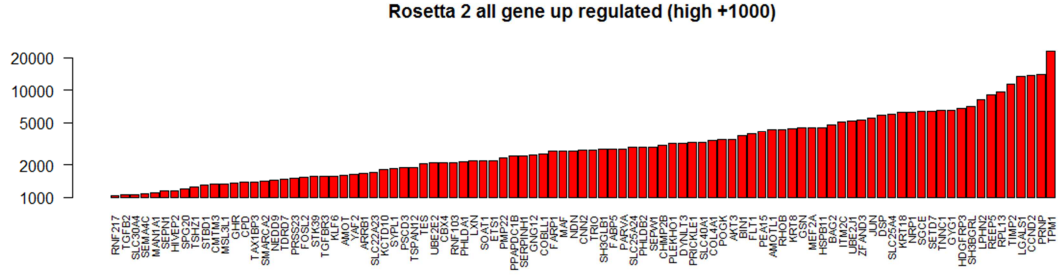


Supplementary figure 7. Mean expression in density of genes that are up vs. down regulated Mouse Gene Atlas data set



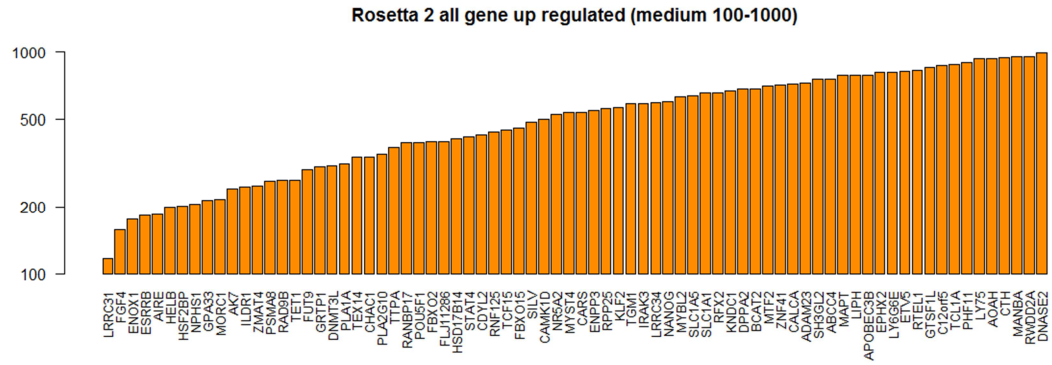
Supplementary figure 8. Mean expression in density of genes that are up vs. down regulated Mouse Gene Atlas data set

This bar plots shows which genes are most expressed in differentiated tissue and we can see the same pattern of expression as in Rosetta 1. In this bar plot Tpm1, Lgals1, Tnnc1, Rpl13, Timp2, Jun seem to be highly up-regulated.



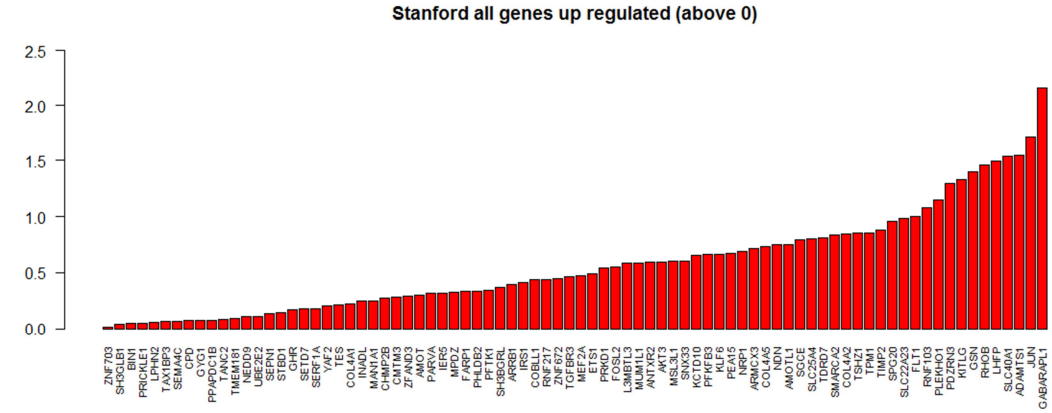
Supplementary figure 9. Genes with the highest averaged expression in all tissue in Rosetta 2 data set

Similar to what happened in Rosetta 1, the genes down regulated in every differentiated tissue are the expected, like Fgf4, Esrrb, Morc1, Dnmt3l, Pou5f1, Nanog. All this are linked to the undifferentiated cell state.



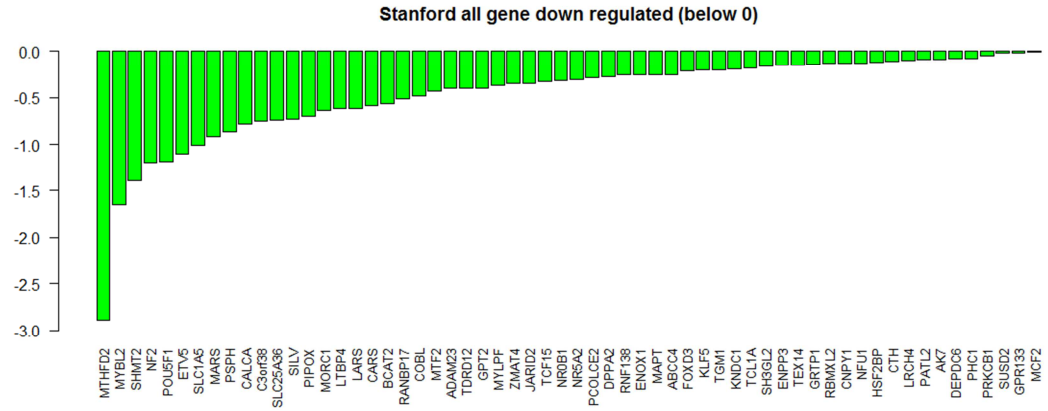
Supplementary figure 10. Genes with the highest averaged expression in all tissue in Rosetta 1 data set

The results of bar plots in Stanford match the previous 2 analysed data sets show again almost the same up and down regulated genes, in the several tissues. The up regulated genes consist in Timp2, Jun, Tpm1, Klf6.



Supplementary figure 11. Genes with the highest expression in all tissue in Stanford data set.

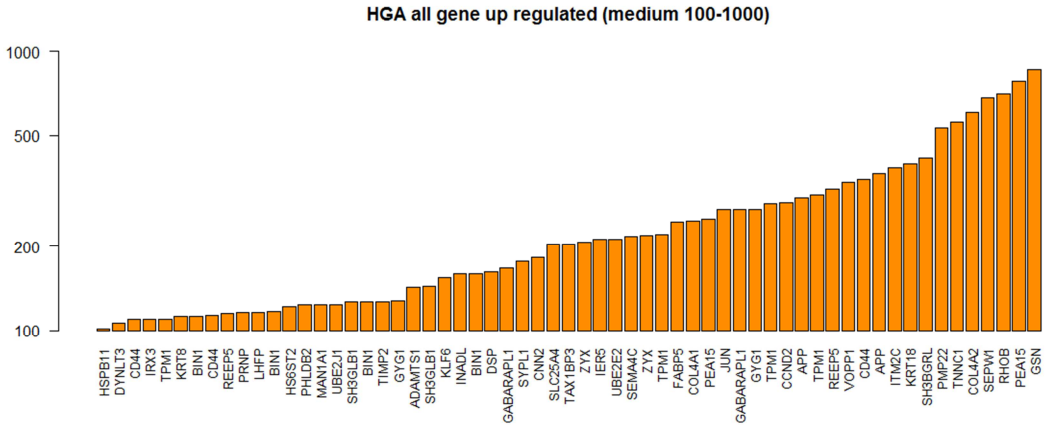
In the down regulated genes we can find genes that linked to pluripotent state, such as, Pou5f1, Klf5, Morc1, Fgf5 and Dppa2.



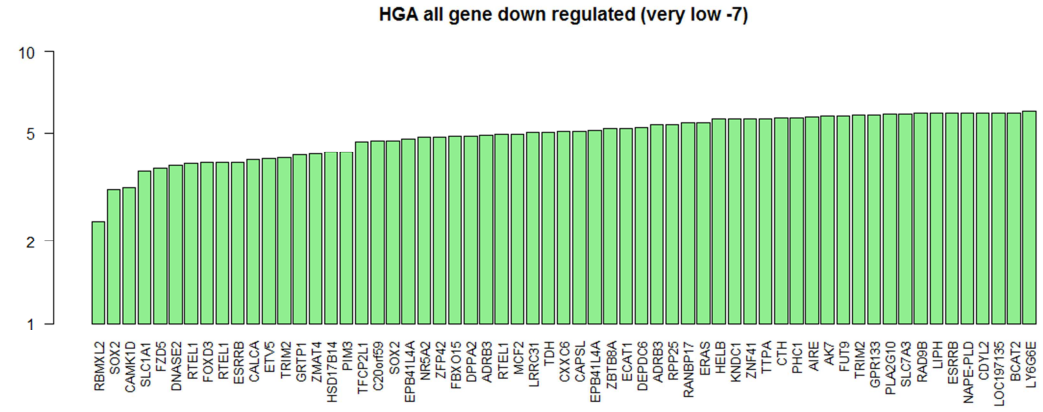
Supplementary figure 12. Genes with the lowest expression in all tissue in Stanford data set.

Could not do clustering for these genes, since the values in the database where between -3 and 3.

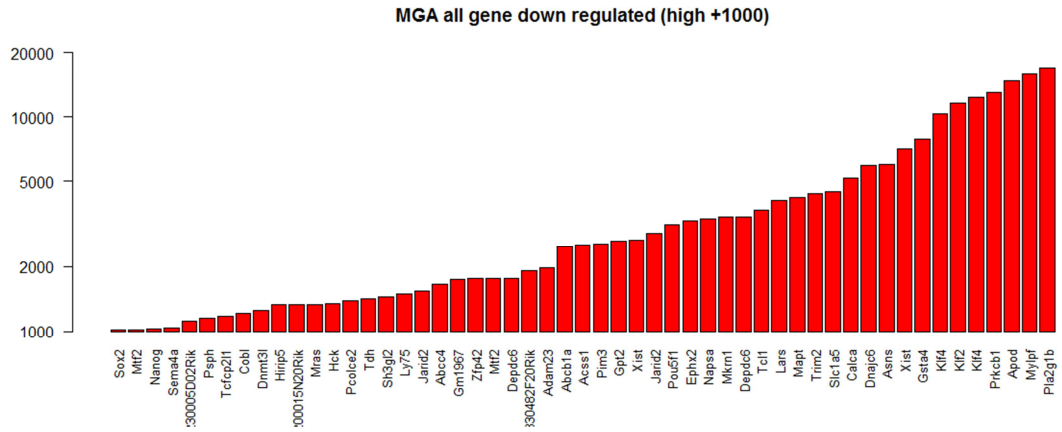
Once again we can see a similar pattern in the expression of genes. Tnnc1, Jun, Tpm1. We can also see that down regulated genes are also similar to the other data set, Sox2, Esrrb, Dppa2. So basically, it is very standard the type of expression in every data set, so it would be safe to assume that in the different tissues these are the most up and down regulated genes.



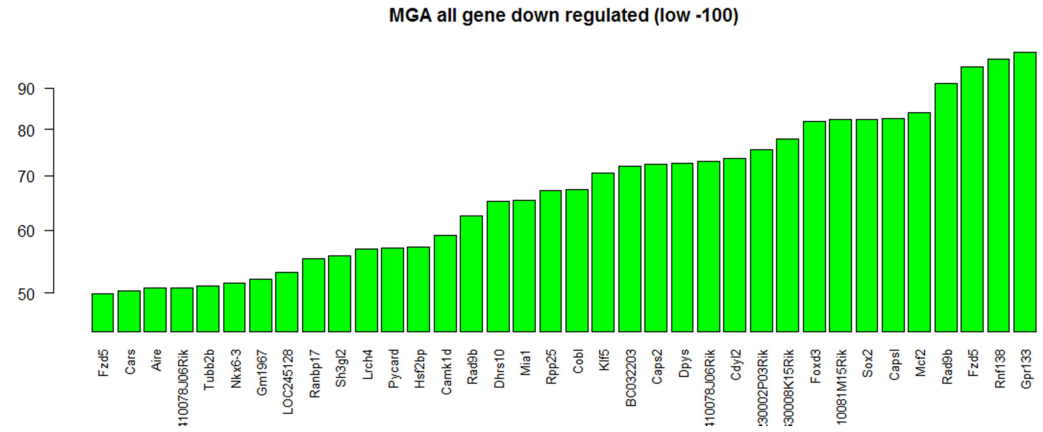
Supplementary figure 13. Genes with the highest averaged expression in all tissue in Human Gene Atlas data set.



Supplementary figure 14. Genes with the lowest averaged expression in all tissue in Human Gene Atlas data set

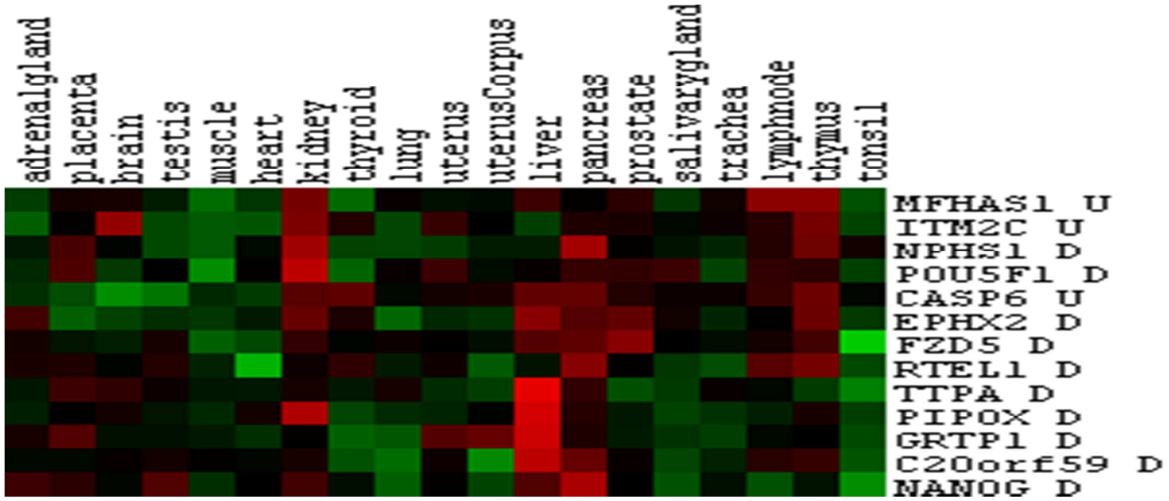


Supplementary figure 15 Genes with the lowest averaged expression in all tissue in Mouse Gene Atlas data set



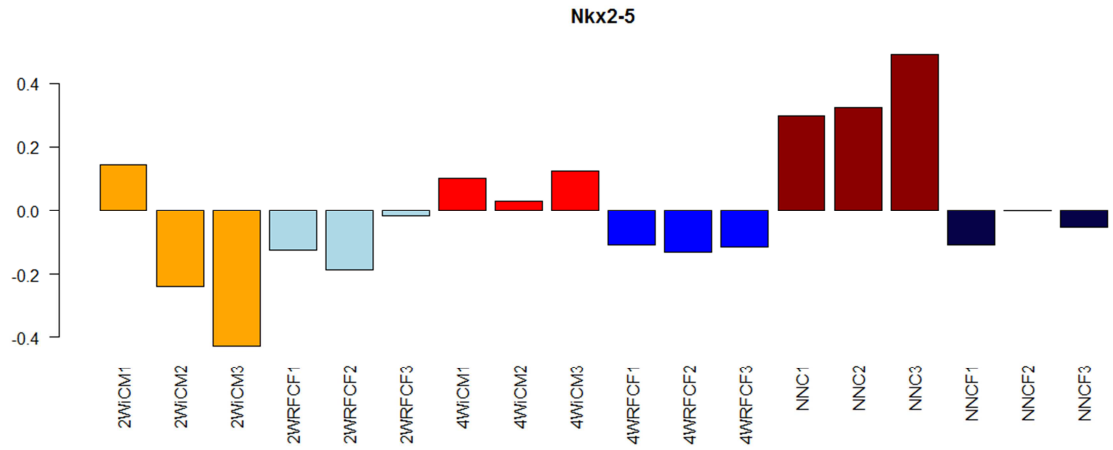
Supplementary figure 16. Genes with the lowest averaged expression in all tissue in Human Gene Atlas data set

In the clustering analysis, once again we can see the Nanog and Pou5f1 clustering together, being this last ones responsible for maintenance of the cells in the undifferentiated state.

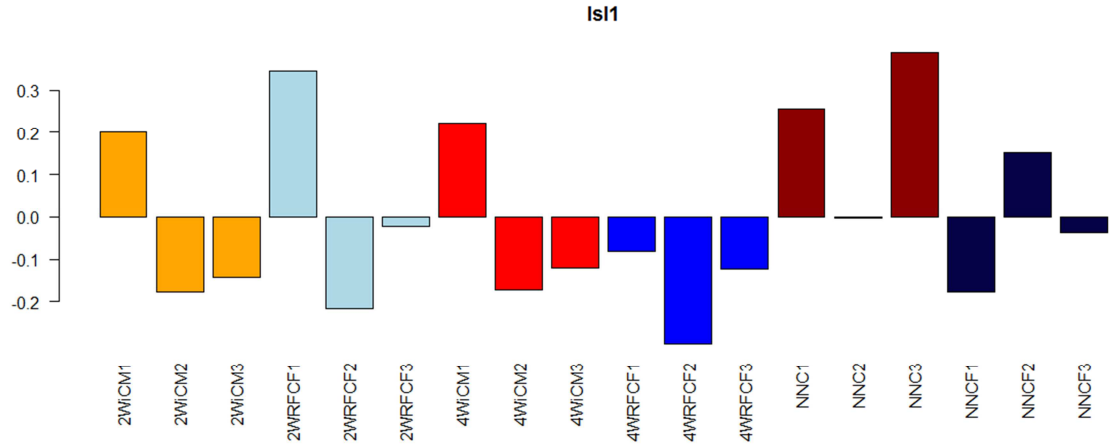


Supplementary figure 17. Group of genes that are clustering together in Rosetta 1.

ANNEX II

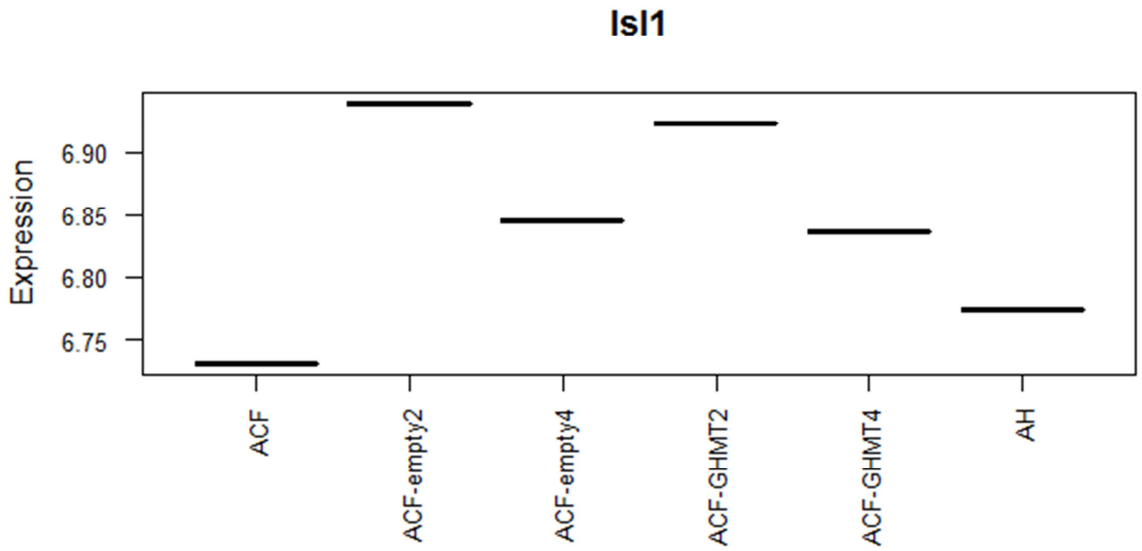


Supplementary figure 1. Expression of NK2 homeobox 5 in several cells types on Ieda e*t al.*, 2010 data set.



Supplementary figure 2. Expression of ISL1 transcription factor, LIM/homeodomain in several cells types on Ieda e*t al.*, 2010 data set.

ANNEX III



Supplementary figure 1. Expression of ISL1 transcription factor, LIM/homeodomain on Song e*t al*., 2012 data set.

ANNEX IV

Supplementary figure 1. Heat map for molecular function with missing values and less important functions.

ANNEX V

Supplementary table 1. Up- and down-regulated genes from Gaspar 2012.

In differentiated cells						
Up regulated genes			Down regulated genes			
2610002M06Rik	Inadl	Samd4	1190003J15Rik	Cth	Klf2	Psma8
Adamts1	Irs1	Sema4c	1700019N12Rik	D14Ertd668e	Klf4	Psph
Akt3	Irx3	Sepn1	1700029I01Rik	D15Ertd55e	Klf5	Pycard
Amot	Itm2c	Sepw1	1700029I01Rik	D1Pas1	Kndc1	Rad9b
Amotl1	Itpripl2	Serf1	1700029P11Rik	Depdc6	Krt42	Ranbp17
Antxr2	Jun	Serpinh1	1700061G19Rik	Dleu7	Lars	Rbpj
App	Kctd10	Setd7	1700097N02Rik	Dnajc6	Liph	Rfx2
Armcx3	Kitl	Sgce	2200001I15Rik	Dnase2a	Rbmxl2	Rinl
Arrb1	Klf6	Sh3bgrl	2210409E12Rik	Dnmt3l	Lrch4	Rnf125
Bag2	Krt18	Sh3glb1	2310043M15Rik	Dppa2	Lrrc2	Rnf138
Bin1	Krt8	Slc22a23	2410004A20Rik	Dpys	Lrrc31	Rpp25
Casp6	L3mbtl3	Slc25a24	2610305D13Rik	Enox1	Lrrc34	Rtel1
Cbx4	Lgals1	Slc25a4	4833420G11Rik	Enpp3	Ltbp4	Rwdd2a
Ccnd2	Lhfp	Slc30a4	4930453N24Rik	Epb4.1l4a	Ly6g6e	Sema4a
Cd44	Lhfpl2	Slc36a4	4930486G11Rik	Ephx2	Ly75	Serpinb6c
Cdk14	LOC100048050	Slc39a8	4930566F21Rik	Eras	Mageb16	Sh3gl2
Chmp2b	Lxn	Slc40a1	5730507C01Rik	Esrrb	Manba	Shmt2
Cmtm3	Maf	Smarca2	9630033F20Rik	Etv5	Mapt	Si
Cnn2	Man1a	Snx33	A230050P20Rik	Fam13c	Mars	Slc17a9
Cobll1	Mef2a	Soat1	Aard	Fbxo15	Mcf2	Slc1a1
Col4a1	Mfhas1	Sox4	Abcb1a	Fbxo2	Mia1	Slc1a5
Col4a2	Mpdz	Speg	Abcc4	Fgf4	Mkrn1	Slc25a36
Col4a5	Msl3	Spg20	Accsl	Foxd3	Mlh3	Slc7a3
Cpd	Mum1l1	Stbd1	Acss1	Frrs1	Morc1	Sox2
Cyth3	Ndn	Stk39	Adam23	Fut9	Mras	Stat4
Dsp	Nedd9	Sypl	Adrb3	Fzd5	Mtf2	Sult6b1
Dynlt3	Nrp1	Tanc2	Aire	Gdf3	Mthfd2	Susd2
Efna5	Parva	Tax1bp3	Ak7	Gli1	Mybl2	Tcea3
Ets1	Pdzrn3	Tdrd7	Ano9	Gm13138	Mylpf	Tcf15
Fabp5	Pea15a	Tes	Aoah	Gm13139	Myst4	Tcfcp2l1
Fam115a	Pfkfb3	Tgfb2	Apobec3	Gm14430	Nanog	Tcl1
Farp1	Phlda1	Tgfbr3	Apod	Gm5101	Napepld	Tdh
Flt1	Phldb2	Timp2	As3mt	Gm6792	Napsa	Tdrd12
Fndc3c1	Pitx2	Tmem181a	Asns	Gm7325	Nf2	Tet1
Fosl2	Pkdcc	Tnnc1	AU018091	Gpa33	Nfu1	Tex14
Fzd2	Plekho1	Tox3	BC032203	Gpr133	Nkx6-3	Tgm1
Gabarapl1	Pmp22	Tpm1	Bcat2	Gpt2	Nphs1	Trim2
Ghr	Pogk	Trio	Bdh2	Grtp1	Nr0b1	Triml1
Gng12	Ppapdc1b	Tshz1	C330019L16Rik	Gsta4	Nr5a2	Ttpa
Gpsm1	Prickle1	Tspan12	C730049O14Rik	Gtsf1	Ooep	Tubb2b
Gpx8	Prkd1	Ube2e2	Calca	Gtsf1l	Patl2	Xist
Gsn	Prnp	Ube2j1	Camk1d	Hck	Pcolce2	Zbtb8a
Gyg	Prss23	Vopp1	Caps2	Helb	Phc1	Zfp229
H2-T10	Pxdn	Vwa5a	Capsl	Hsd17b14	Pim3	Zfp296
Hdgfrp3	Reep5	Yaf2	Cars	Hsf2bp	Pipox	Zfp42
Hivep2	Rhob	Zfand3	Ccdc160	I1C0022H11Rik	Pla1a	Zfp518b
Hs6st2	Rhou	Zfp386	Cdyl2	Ildr1	Pla2g10	Zfp59
Hspb11	Rnf103	Zfp672	Chac1	Irak3	Pla2g1b	Zfp819
Ier5	Rnf217	Zfp703	Clca4	Jam2	Pou5f1	Zmat4
Igf2	Rpl13	Zyx	Cnpy1	Jarid2	Prdm14	
			Cobl	Kbtbd11	Prkcb	

Supplementary table 2. Overlapping down regulated genes between the different data sets.

Human Entrez	Gene Symbol	Agapios	CancerI	CancerII	Oct4
25	ABL1	0	1	1	0
100	ADA	0	0	1	1
155	ADRB3	1	0	1	0
207	AKT1	0	1	1	0
208	AKT2	0	1	1	0
217	ALDH2	0	1	1	0
238	ALK	0	1	1	0
324	APC	0	1	1	1
326	AIRE	1	0	1	0
330	BIRC3	0	1	1	0
355	FAS	0	1	1	0
405	ARNT	0	1	1	0
466	ATF1	0	1	1	0
471	ATIC	0	1	1	0
472	ATM	0	1	1	0
546	ATRX	0	1	1	0
595	CCND1	0	1	1	0
596	BCL2	0	1	1	0
604	BCL6	0	1	1	0
605	BCL7A	0	1	1	0
608	TNFRSF17	0	1	1	0
613	BCR	0	1	1	0
641	BLM	0	1	1	1
657	BMPR1A	0	1	1	0
668	FOXL2	0	1	1	0
672	BRCA1	0	1	1	0
673	BRAF	0	1	1	0
675	BRCA2	0	1	1	0
688	KLF5	1	0	1	0
701	BUB1B	0	1	1	0
768	CA9	0	0	1	1
796	CALCA	1	0	1	0
824	CAPN2	0	0	1	1
833	CARS	1	1	0	0
835	CASP2	0	0	1	1
861	RUNX1	0	1	1	0
865	CBFB	0	1	1	0
867	CBL	0	1	1	0
894	CCND2	0	1	1	0

896	CCND3	**0**	**1**	**1**	**0**
898	CCNE1	**0**	**1**	**1**	**0**
947	CD34	**0**	**0**	**1**	**1**
966	CD59	**0**	**0**	**1**	**1**
973	CD79A	**0**	**1**	**1**	**0**
974	CD79B	**0**	**1**	**1**	**0**
999	CDH1	**0**	**1**	**1**	**0**
1009	CDH11	**0**	**1**	**1**	**0**
1019	CDK4	**0**	**1**	**1**	**0**
1021	CDK6	**0**	**1**	**1**	**0**
1029	CDKN2A	**0**	**1**	**1**	**0**
1031	CDKN2C	**0**	**1**	**1**	**0**
1045	CDX2	**0**	**1**	**1**	**1**
1050	CEBPA	**0**	**1**	**1**	**0**
1277	COL1A1	**0**	**1**	**1**	**0**
1301	COL11A1	**0**	**0**	**1**	**1**
1316	KLF6	**0**	**1**	**1**	**0**
1345	COX6C	**0**	**1**	**1**	**0**
1376	CPT2	**0**	**0**	**1**	**1**
1387	CREBBP	**0**	**1**	**1**	**0**
1491	CTH	**1**	**0**	**1**	**0**
1499	CTNNB1	**0**	**1**	**1**	**0**
1535	CYBA	**0**	**0**	**1**	**1**
1540	CYLD	**0**	**1**	**1**	**0**
1585	CYP11B2	**0**	**0**	**1**	**1**
1589	CYP21A2	**0**	**0**	**1**	**1**
1615	DARS	**0**	**0**	**1**	**1**
1616	DAXX	**0**	**1**	**1**	**0**
1643	DDB2	**0**	**1**	**1**	**0**
1649	DDIT3	**0**	**1**	**1**	**1**
1769	DNAH8	**0**	**0**	**1**	**1**
1785	DNM2	**0**	**1**	**1**	**0**
1788	DNMT3A	**0**	**1**	**1**	**0**
1807	DPYS	**1**	**0**	**1**	**0**
1956	EGFR	**0**	**1**	**1**	**0**
2014	EMP3	**0**	**0**	**1**	**1**
2033	EP300	**0**	**1**	**1**	**1**
2053	EPHX2	**1**	**0**	**1**	**0**
2064	ERBB2	**0**	**1**	**1**	**0**
2068	ERCC2	**0**	**1**	**1**	**0**
2071	ERCC3	**0**	**1**	**1**	**0**
2072	ERCC4	**0**	**1**	**1**	**0**
2073	ERCC5	**0**	**1**	**1**	**0**

2078	ERG	**0**	**1**	**1**	**0**
2103	ESRRB	**1**	**0**	**1**	**0**
2115	ETV1	**0**	**1**	**1**	**0**
2119	ETV5	**1**	**1**	**0**	**0**
2120	ETV6	**0**	**1**	**1**	**0**
2122	MECOM	**0**	**1**	**1**	**0**
2131	EXT1	**0**	**1**	**1**	**0**
2146	EZH2	**0**	**1**	**1**	**0**
2175	FANCA	**0**	**1**	**1**	**0**
2176	FANCC	**0**	**1**	**1**	**0**
2177	FANCD2	**0**	**1**	**1**	**0**
2178	FANCE	**0**	**1**	**1**	**0**
2188	FANCF	**0**	**1**	**1**	**0**
2189	FANCG	**0**	**1**	**1**	**0**
2206	MS4A2	**0**	**0**	**1**	**1**
2213	FCGR2B	**0**	**1**	**1**	**0**
2249	FGF4	**1**	**0**	**1**	**0**
2260	FGFR1	**0**	**1**	**1**	**0**
2261	FGFR3	**0**	**1**	**1**	**0**
2263	FGFR2	**0**	**1**	**1**	**0**
2267	FGL1	**0**	**0**	**1**	**1**
2271	FH	**0**	**1**	**1**	**0**
2272	FHIT	**0**	**1**	**1**	**0**
2308	FOXO1	**0**	**1**	**1**	**0**
2322	FLT3	**0**	**1**	**1**	**0**
2494	NR5A2	**1**	**0**	**1**	**0**
2512	FTL	**0**	**0**	**1**	**1**
2623	GATA1	**0**	**1**	**1**	**0**
2624	GATA2	**0**	**1**	**1**	**0**
2625	GATA3	**0**	**1**	**1**	**0**
2638	GC	**0**	**0**	**1**	**1**
2705	GJB1	**0**	**0**	**1**	**1**
2735	GLI1	**1**	**0**	**1**	**0**
2739	GLO1	**0**	**0**	**1**	**1**
2767	GNA11	**0**	**1**	**1**	**0**
2776	GNAQ	**0**	**1**	**1**	**0**
2778	GNAS	**0**	**1**	**1**	**0**
2941	GSTA4	**1**	**0**	**1**	**0**
2950	GSTP1	**0**	**0**	**1**	**1**
2953	GSTT2	**0**	**0**	**1**	**1**
2956	MSH6	**0**	**1**	**1**	**0**
3030	HADHA	**0**	**0**	**1**	**1**
3094	HINT1	**0**	**0**	**1**	**1**

3265	HRAS	**0**	**1**	**1**	**0**
3320	HSP90AA1	**0**	**1**	**1**	**0**
3417	IDH1	**0**	**1**	**1**	**0**
3418	IDH2	**0**	**1**	**1**	**0**
3489	IGFBP6	**0**	**0**	**1**	**1**
3492	IGH@	**0**	**1**	**1**	**0**
3558	IL2	**0**	**1**	**1**	**0**
3572	IL6ST	**0**	**1**	**1**	**0**
3662	IRF4	**0**	**1**	**1**	**0**
3716	JAK1	**0**	**1**	**1**	**0**
3717	JAK2	**0**	**1**	**1**	**0**
3718	JAK3	**0**	**1**	**1**	**0**
3720	JARID2	**1**	**0**	**1**	**0**
3725	JUN	**0**	**1**	**1**	**0**
3791	KDR	**0**	**1**	**1**	**0**
3811	KIR3DL1	**0**	**0**	**1**	**1**
3815	KIT	**0**	**1**	**1**	**0**
3817	KLK2	**0**	**1**	**1**	**0**
3845	KRAS	**0**	**1**	**1**	**0**
3977	LIFR	**0**	**1**	**1**	**0**
4004	LMO1	**0**	**1**	**1**	**1**
4005	LMO2	**0**	**1**	**1**	**0**
4089	SMAD4	**0**	**1**	**1**	**0**
4126	MANBA	**1**	**0**	**1**	**0**
4137	MAPT	**1**	**0**	**1**	**0**
4141	MARS	**1**	**0**	**1**	**0**
4193	MDM2	**0**	**1**	**1**	**0**
4194	MDM4	**0**	**1**	**1**	**0**
4214	MAP3K1	**0**	**0**	**1**	**1**
4221	MEN1	**0**	**1**	**1**	**0**
4233	MET	**0**	**1**	**1**	**0**
4261	CIITA	**0**	**1**	**1**	**0**
4286	MITF	**0**	**1**	**1**	**0**
4292	MLH1	**0**	**1**	**1**	**0**
4297	MLL	**0**	**1**	**1**	**0**
4300	MLLT3	**0**	**1**	**1**	**0**
4352	MPL	**0**	**1**	**1**	**0**
4436	MSH2	**0**	**1**	**1**	**0**
4515	MTCP1	**0**	**1**	**1**	**0**
4582	MUC1	**0**	**1**	**1**	**0**
4595	MUTYH	**0**	**1**	**1**	**0**
4605	MYBL2	**1**	**0**	**1**	**0**
4609	MYC	**0**	**1**	**1**	**0**

4610	MYCL1	**0**	**1**	**1**	**0**
4613	MYCN	**0**	**1**	**1**	**0**
4615	MYD88	**0**	**1**	**1**	**0**
4629	MYH11	**0**	**1**	**1**	**0**
4683	NBN	**0**	**1**	**1**	**0**
4761	NEUROD2	**0**	**0**	**1**	**1**
4763	NF1	**0**	**1**	**1**	**0**
4771	NF2	**1**	**1**	**1**	**0**
4780	NFE2L2	**0**	**1**	**1**	**0**
4790	NFKB1	**0**	**0**	**1**	**1**
4791	NFKB2	**0**	**1**	**1**	**0**
4798	NFRKB	**0**	**0**	**1**	**1**
4851	NOTCH1	**0**	**1**	**1**	**0**
4853	NOTCH2	**0**	**1**	**1**	**0**
4869	NPM1	**0**	**1**	**1**	**0**
4893	NRAS	**0**	**1**	**1**	**0**
4914	NTRK1	**0**	**1**	**1**	**0**
4916	NTRK3	**0**	**1**	**1**	**0**
4926	NUMA1	**0**	**1**	**1**	**0**
4958	OMD	**0**	**1**	**0**	**1**
5055	SERPINB2	**0**	**0**	**1**	**1**
5077	PAX3	**0**	**1**	**1**	**0**
5079	PAX5	**0**	**1**	**1**	**0**
5155	PDGFB	**0**	**1**	**1**	**0**
5156	PDGFRA	**0**	**1**	**1**	**0**
5159	PDGFRB	**0**	**1**	**1**	**0**
5187	PER1	**0**	**1**	**1**	**0**
5243	ABCB1	**1**	**0**	**1**	**0**
5245	PHB	**0**	**0**	**1**	**1**
5273	SERPINB10	**0**	**0**	**1**	**1**
5290	PIK3CA	**0**	**1**	**1**	**0**
5292	PIM1	**0**	**1**	**1**	**0**
5295	PIK3R1	**0**	**1**	**1**	**0**
5334	PLCL1	**0**	**0**	**1**	**1**
5371	PML	**0**	**1**	**1**	**0**
5378	PMS1	**0**	**1**	**1**	**0**
5395	PMS2	**0**	**1**	**1**	**0**
5429	POLH	**0**	**0**	**1**	**1**
5434	POLR2E	**0**	**0**	**1**	**1**
5450	POU2AF1	**0**	**1**	**1**	**0**
5460	POU5F1	**1**	**1**	**0**	**1**
5468	PPARG	**0**	**1**	**1**	**0**
5518	PPP2R1A	**0**	**1**	**1**	**0**

5551	PRF1	**0**	**1**	**1**	**0**
5573	PRKAR1A	**0**	**1**	**1**	**0**
5648	MASP1	**0**	**0**	**1**	**1**
5723	PSPH	**1**	**0**	**0**	**1**
5727	PTCH1	**0**	**1**	**1**	**0**
5728	PTEN	**0**	**1**	**1**	**0**
5781	PTPN11	**0**	**1**	**1**	**0**
5829	PXN	**0**	**0**	**1**	**1**
5884	RAD17	**0**	**0**	**1**	**1**
5890	RAD51B	**0**	**1**	**1**	**0**
5894	RAF1	**0**	**1**	**1**	**0**
5900	RALGDS	**0**	**1**	**0**	**1**
5909	RAP1GAP	**0**	**0**	**1**	**1**
5925	RB1	**0**	**1**	**1**	**0**
5966	REL	**0**	**1**	**1**	**0**
5979	RET	**0**	**1**	**1**	**0**
6098	ROS1	**0**	**1**	**1**	**0**
6184	RPN1	**0**	**1**	**1**	**0**
6347	CCL2	**0**	**0**	**1**	**1**
6390	SDHB	**0**	**1**	**1**	**0**
6391	SDHC	**0**	**1**	**1**	**0**
6392	SDHD	**0**	**1**	**1**	**0**
6416	MAP2K4	**0**	**1**	**1**	**0**
6421	SFPQ	**0**	**1**	**1**	**1**
6428	SRSF3	**0**	**1**	**0**	**1**
6456	SH3GL2	**1**	**0**	**1**	**0**
6472	SHMT2	**1**	**0**	**1**	**0**
6490	PMEL	**1**	**0**	**1**	**0**
6597	SMARCA4	**0**	**1**	**1**	**0**
6598	SMARCB1	**0**	**1**	**1**	**0**
6608	SMO	**0**	**1**	**1**	**0**
6657	SOX2	**1**	**1**	**0**	**0**
6688	SPI1	**0**	**0**	**1**	**1**
6775	STAT4	**1**	**0**	**1**	**0**
6789	STK4	**0**	**0**	**1**	**1**
6794	STK11	**0**	**1**	**1**	**0**
6840	SVIL	**0**	**0**	**1**	**1**
6850	SYK	**0**	**1**	**1**	**0**
6886	TAL1	**0**	**1**	**1**	**0**
6888	TALDO1	**0**	**0**	**1**	**1**
6927	HNF1A	**0**	**1**	**1**	**0**
6934	TCF7L2	**0**	**1**	**1**	**0**
6938	TCF12	**0**	**1**	**1**	**0**

6940	ZNF354A	**0**	**0**	**1**	**1**
6957	TRB@	**0**	**1**	**1**	**0**
7029	TFDP2	**0**	**0**	**1**	**1**
7037	TFRC	**0**	**1**	**1**	**0**
7051	TGM1	**1**	**0**	**1**	**0**
7080	NKX2-1	**0**	**1**	**1**	**0**
7113	TMPRSS2	**0**	**1**	**1**	**0**
7128	TNFAIP3	**0**	**1**	**1**	**0**
7157	TP53	**0**	**1**	**1**	**0**
7175	TPR	**0**	**1**	**0**	**1**
7248	TSC1	**0**	**1**	**1**	**0**
7249	TSC2	**0**	**1**	**1**	**0**
7253	TSHR	**0**	**1**	**1**	**0**
7268	TTC4	**0**	**0**	**1**	**1**
7274	TTPA	**1**	**0**	**1**	**0**
7356	SCGB1A1	**0**	**0**	**1**	**1**
7428	VHL	**0**	**1**	**1**	**0**
7430	EZR	**0**	**1**	**1**	**0**
7450	VWF	**0**	**0**	**1**	**1**
7486	WRN	**0**	**1**	**1**	**0**
7490	WT1	**0**	**1**	**1**	**0**
7507	XPA	**0**	**1**	**1**	**0**
7508	XPC	**0**	**1**	**1**	**0**
7515	XRCC1	**0**	**0**	**1**	**1**
7531	YWHAE	**0**	**1**	**1**	**0**
7704	ZBTB16	**0**	**1**	**1**	**0**
7849	PAX8	**0**	**1**	**1**	**0**
7918	GPANK1	**0**	**0**	**1**	**1**
7979	SHFM1	**0**	**0**	**1**	**1**
8028	MLLT10	**0**	**1**	**1**	**0**
8030	CCDC6	**0**	**1**	**1**	**0**
8031	NCOA4	**0**	**1**	**1**	**0**
8091	HMGA2	**0**	**1**	**1**	**0**
8115	TCL1A	**1**	**1**	**1**	**1**
8242	KDM5C	**0**	**1**	**0**	**1**
8314	BAP1	**0**	**1**	**1**	**0**
8425	LTBP4	**1**	**0**	**1**	**0**
8435	SOAT2	**0**	**0**	**1**	**1**
8648	NCOA1	**0**	**1**	**1**	**0**
8651	SOCS1	**0**	**1**	**1**	**0**
8764	TNFRSF14	**0**	**1**	**1**	**0**
8880	FUBP1	**0**	**1**	**0**	**1**
8887	TAX1BP1	**0**	**0**	**1**	**1**

8915	BCL10	**0**	**1**	**1**	**0**
8929	PHOX2B	**0**	**1**	**1**	**0**
9150	CTDP1	**0**	**0**	**1**	**1**
9321	TRIP11	**0**	**1**	**1**	**0**
9338	TCEAL1	**0**	**0**	**1**	**1**
9401	RECQL4	**0**	**1**	**1**	**0**
9817	KEAP1	**0**	**0**	**1**	**1**
9935	MAFB	**0**	**1**	**1**	**0**
9968	MED12	**0**	**1**	**0**	**1**
9969	MED13	**0**	**0**	**1**	**1**
10039	PARP3	**0**	**0**	**1**	**1**
10142	AKAP9	**0**	**1**	**1**	**0**
10257	ABCC4	**1**	**0**	**1**	**0**
10320	IKZF1	**0**	**1**	**1**	**0**
10397	NDRG1	**0**	**1**	**1**	**0**
10413	YAP1	**0**	**0**	**1**	**1**
10499	NCOA2	**0**	**1**	**1**	**0**
10797	MTHFD2	**1**	**0**	**1**	**0**
10892	MALT1	**0**	**1**	**1**	**0**
11064	CNTRL	**0**	**1**	**1**	**0**
11197	WIF1	**0**	**1**	**1**	**0**
11200	CHEK2	**0**	**1**	**1**	**0**
11213	IRAK3	**1**	**0**	**1**	**0**
23092	ARHGAP26	**0**	**1**	**1**	**0**
23242	COBL	**1**	**0**	**1**	**0**
23305	ACSL6	**0**	**1**	**1**	**0**
23332	CLASP1	**0**	**0**	**1**	**1**
23405	DICER1	**0**	**1**	**1**	**0**
23522	KAT6B	**1**	**1**	**0**	**0**
25818	KLK5	**0**	**0**	**1**	**1**
27030	MLH3	**1**	**0**	**1**	**0**
27063	ANKRD1	**0**	**0**	**1**	**1**
27436	EML4	**0**	**1**	**1**	**0**
29844	TFPT	**0**	**1**	**0**	**1**
29947	DNMT3L	**1**	**0**	**1**	**0**
29998	GLTSCR1	**0**	**0**	**1**	**1**
50802	IGK@	**0**	**1**	**1**	**0**
50939	IMPG2	**0**	**0**	**1**	**1**
51094	ADIPOR1	**0**	**0**	**1**	**1**
51131	PHF11	**1**	**0**	**1**	**0**
51199	NIN	**0**	**1**	**1**	**0**
51378	ANGPT4	**0**	**0**	**1**	**1**
51520	LARS	**1**	**0**	**0**	**1**

51684	SUFU	**0**	**1**	**1**	**0**
51750	RTEL1	**1**	**0**	**1**	**0**
51755	CDK12	**0**	**1**	**1**	**1**
54790	TET2	**0**	**1**	**1**	**0**
55234	SMU1	**0**	**0**	**1**	**1**
55259	CASC1	**0**	**0**	**1**	**1**
55294	FBXW7	**0**	**1**	**1**	**0**
55862	ECHDC1	**0**	**0**	**1**	**1**
55971	BAIAP2L1	**0**	**0**	**1**	**1**
56155	TEX14	**1**	**0**	**1**	**0**
57103	C12orf5	**1**	**0**	**1**	**0**
57118	CAMK1D	**1**	**0**	**1**	**0**
57412	AS3MT	**1**	**0**	**1**	**0**
63910	SLC17A9	**1**	**0**	**0**	**1**
63976	PRDM16	**0**	**1**	**1**	**0**
63978	PRDM14	**1**	**0**	**0**	**1**
64241	ABCG8	**0**	**0**	**1**	**1**
64746	ACBD3	**0**	**0**	**1**	**1**
64805	P2RY12	**0**	**0**	**1**	**1**
64901	RANBP17	**1**	**1**	**0**	**0**
79723	SUV39H2	**0**	**0**	**1**	**1**
79728	PALB2	**0**	**1**	**1**	**0**
80312	TET1	**1**	**1**	**1**	**0**
80380	PDCD1LG2	**0**	**1**	**1**	**0**
81623	DEFB126	**0**	**0**	**1**	**1**
83990	BRIP1	**0**	**1**	**1**	**0**
84441	MAML2	**0**	**1**	**1**	**0**
84532	ACSS1	**1**	**0**	**0**	**1**
84651	SPINK7	**0**	**0**	**1**	**1**
92689	FAM114A1	**0**	**0**	**1**	**1**
121227	LRIG3	**0**	**1**	**1**	**0**
139285	FAM123B	**0**	**1**	**1**	**0**
140767	NRSN1	**0**	**0**	**1**	**1**
143187	VTI1A	**0**	**1**	**1**	**0**
144715	RAD9B	**1**	**0**	**0**	**1**
151393	FAM82A1	**0**	**0**	**1**	**1**
171023	ASXL1	**0**	**1**	**1**	**0**
221895	JAZF1	**0**	**1**	**1**	**0**
285733		**0**	**0**	**1**	**1**
374378	GALNTL4	**0**	**0**	**1**	**1**

Supplementary table 2. Overlapping up regulated genes between the different data sets.

Human Entrez	Symbol	Up Agapios	CancerII	Oct4	CancerI
25	ABL1	0	1	0	1
100	ADA	0	1	1	0
207	AKT1	0	1	0	1
208	AKT2	0	1	0	1
217	ALDH2	0	1	0	1
238	ALK	0	1	0	1
324	APC	0	1	1	1
330	BIRC3	0	1	0	1
355	FAS	0	1	0	1
405	ARNT	0	1	0	1
466	ATF1	0	1	0	1
471	ATIC	0	1	0	1
472	ATM	0	1	0	1
546	ATRX	0	1	0	1
595	CCND1	0	1	0	1
596	BCL2	0	1	0	1
604	BCL6	0	1	0	1
605	BCL7A	0	1	0	1
608	TNFRSF17	0	1	0	1
613	BCR	0	1	0	1
641	BLM	0	1	1	1
657	BMPR1A	0	1	0	1
668	FOXL2	0	1	0	1
672	BRCA1	0	1	0	1
673	BRAF	0	1	0	1
675	BRCA2	0	1	0	1
701	BUB1B	0	1	0	1
768	CA9	0	1	1	0
824	CAPN2	0	1	1	0
835	CASP2	0	1	1	0
839	CASP6	1	1	0	0
861	RUNX1	0	1	0	1
865	CBFB	0	1	0	1
867	CBL	0	1	0	1
894	CCND2	1	1	0	1
896	CCND3	0	1	0	1
898	CCNE1	0	1	0	1
947	CD34	0	1	1	0
960	CD44	1	1	0	0
966	CD59	0	1	1	0
973	CD79A	0	1	0	1

974	CD79B	0	1	0	1
999	CDH1	0	1	0	1
1009	CDH11	0	1	0	1
1019	CDK4	0	1	0	1
1021	CDK6	0	1	0	1
1029	CDKN2A	0	1	0	1
1031	CDKN2C	0	1	0	1
1045	CDX2	0	1	1	1
1050	CEBPA	0	1	0	1
1277	COL1A1	0	1	0	1
1284	COL4A2	1	1	0	0
1301	COL11A1	0	1	1	0
1316	KLF6	1	1	0	1
1345	COX6C	0	1	0	1
1376	CPT2	0	1	1	0
1387	CREBBP	0	1	0	1
1499	CTNNB1	0	1	0	1
1535	CYBA	0	1	1	0
1540	CYLD	0	1	0	1
1585	CYP11B2	0	1	1	0
1589	CYP21A2	0	1	1	0
1615	DARS	0	1	1	0
1616	DAXX	0	1	0	1
1643	DDB2	0	1	0	1
1649	DDIT3	0	1	1	1
1769	DNAH8	0	1	1	0
1785	DNM2	0	1	0	1
1788	DNMT3A	0	1	0	1
1956	EGFR	0	1	0	1
2014	EMP3	0	1	1	0
2033	EP300	0	1	1	1
2064	ERBB2	0	1	0	1
2068	ERCC2	0	1	0	1
2071	ERCC3	0	1	0	1
2072	ERCC4	0	1	0	1
2073	ERCC5	0	1	0	1
2078	ERG	0	1	0	1
2115	ETV1	0	1	0	1
2120	ETV6	0	1	0	1
2122	MECOM	0	1	0	1
2131	EXT1	0	1	0	1
2146	EZH2	0	1	0	1
2175	FANCA	0	1	0	1
2176	FANCC	0	1	0	1
2177	FANCD2	0	1	0	1

2178	FANCE	0	1	0	1
2188	FANCF	0	1	0	1
2189	FANCG	0	1	0	1
2206	MS4A2	0	1	1	0
2213	FCGR2B	0	1	0	1
2260	FGFR1	0	1	0	1
2261	FGFR3	0	1	0	1
2263	FGFR2	0	1	0	1
2267	FGL1	0	1	1	0
2271	FH	0	1	0	1
2272	FHIT	0	1	0	1
2308	FOXO1	0	1	0	1
2321	FLT1	1	1	0	0
2322	FLT3	0	1	0	1
2512	FTL	0	1	1	0
2623	GATA1	0	1	0	1
2624	GATA2	0	1	0	1
2625	GATA3	0	1	0	1
2638	GC	0	1	1	0
2690	GHR	1	1	0	0
2705	GJB1	0	1	1	0
2739	GLO1	0	1	1	0
2767	GNA11	0	1	0	1
2776	GNAQ	0	1	0	1
2778	GNAS	0	1	0	1
2950	GSTP1	0	1	1	0
2953	GSTT2	0	1	1	0
2956	MSH6	0	1	0	1
3030	HADHA	0	1	1	0
3094	HINT1	0	1	1	0
3265	HRAS	0	1	0	1
3320	HSP90AA1	0	1	0	1
3417	IDH1	0	1	0	1
3418	IDH2	0	1	0	1
3481	IGF2	1	1	0	0
3489	IGFBP6	0	1	1	0
3492	IGH@	0	1	0	1
3558	IL2	0	1	0	1
3572	IL6ST	0	1	0	1
3662	IRF4	0	1	0	1
3667	IRS1	1	1	0	0
3716	JAK1	0	1	0	1
3717	JAK2	0	1	0	1
3718	JAK3	0	1	0	1
3725	JUN	1	1	0	1

3791	KDR	0	1	0	1
3811	KIR3DL1	0	1	1	0
3815	KIT	0	1	0	1
3817	KLK2	0	1	0	1
3845	KRAS	0	1	0	1
3956	LGALS1	1	1	0	0
3977	LIFR	0	1	0	1
4004	LMO1	0	1	1	1
4005	LMO2	0	1	0	1
4089	SMAD4	0	1	0	1
4094	MAF	1	0	0	1
4193	MDM2	0	1	0	1
4194	MDM4	0	1	0	1
4214	MAP3K1	0	1	1	0
4221	MEN1	0	1	0	1
4233	MET	0	1	0	1
4254	KITLG	1	1	0	0
4261	CIITA	0	1	0	1
4286	MITF	0	1	0	1
4292	MLH1	0	1	0	1
4297	MLL	0	1	0	1
4300	MLLT3	0	1	0	1
4352	MPL	0	1	0	1
4436	MSH2	0	1	0	1
4515	MTCP1	0	1	0	1
4582	MUC1	0	1	0	1
4595	MUTYH	0	1	0	1
4609	MYC	0	1	0	1
4610	MYCL1	0	1	0	1
4613	MYCN	0	1	0	1
4615	MYD88	0	1	0	1
4629	MYH11	0	1	0	1
4683	NBN	0	1	0	1
4692	NDN	1	1	0	0
4761	NEUROD2	0	1	1	0
4763	NF1	0	1	0	1
4771	NF2	0	1	0	1
4780	NFE2L2	0	1	0	1
4790	NFKB1	0	1	1	0
4791	NFKB2	0	1	0	1
4798	NFRKB	0	1	1	0
4851	NOTCH1	0	1	0	1
4853	NOTCH2	0	1	0	1
4869	NPM1	0	1	0	1
4893	NRAS	0	1	0	1

4914	NTRK1	0	1	0	1
4916	NTRK3	0	1	0	1
4926	NUMA1	0	1	0	1
4958	OMD	0	0	1	1
5055	SERPINB2	0	1	1	0
5077	PAX3	0	1	0	1
5079	PAX5	0	1	0	1
5155	PDGFB	0	1	0	1
5156	PDGFRA	0	1	0	1
5159	PDGFRB	0	1	0	1
5187	PER1	0	1	0	1
5245	PHB	0	1	1	0
5273	SERPINB10	0	1	1	0
5290	PIK3CA	0	1	0	1
5292	PIM1	0	1	0	1
5295	PIK3R1	0	1	0	1
5334	PLCL1	0	1	1	0
5371	PML	0	1	0	1
5378	PMS1	0	1	0	1
5395	PMS2	0	1	0	1
5429	POLH	0	1	1	0
5434	POLR2E	0	1	1	0
5450	POU2AF1	0	1	0	1
5460	POU5F1	0	0	1	1
5468	PPARG	0	1	0	1
5518	PPP2R1A	0	1	0	1
5551	PRF1	0	1	0	1
5573	PRKAR1A	0	1	0	1
5621	PRNP	1	1	0	0
5648	MASP1	0	1	1	0
5727	PTCH1	0	1	0	1
5728	PTEN	0	1	0	1
5781	PTPN11	0	1	0	1
5829	PXN	0	1	1	0
5884	RAD17	0	1	1	0
5890	RAD51B	0	1	0	1
5894	RAF1	0	1	0	1
5900	RALGDS	0	0	1	1
5909	RAP1GAP	0	1	1	0
5925	RB1	0	1	0	1
5966	REL	0	1	0	1
5979	RET	0	1	0	1
6098	ROS1	0	1	0	1
6184	RPN1	0	1	0	1
6347	CCL2	0	1	1	0

6390	SDHB	0	1	0	1
6391	SDHC	0	1	0	1
6392	SDHD	0	1	0	1
6416	MAP2K4	0	1	0	1
6421	SFPQ	0	1	1	1
6428	SRSF3	0	0	1	1
6595	SMARCA2	1	1	0	0
6597	SMARCA4	0	1	0	1
6598	SMARCB1	0	1	0	1
6608	SMO	0	1	0	1
6688	SPI1	0	1	1	0
6789	STK4	0	1	1	0
6794	STK11	0	1	0	1
6840	SVIL	0	1	1	0
6850	SYK	0	1	0	1
6886	TAL1	0	1	0	1
6888	TALDO1	0	1	1	0
6927	HNF1A	0	1	0	1
6934	TCF7L2	0	1	0	1
6938	TCF12	0	1	0	1
6940	ZNF354A	0	1	1	0
6957	TRB@	0	1	0	1
7029	TFDP2	0	1	1	0
7037	TFRC	0	1	0	1
7042	TGFB2	1	1	0	0
7049	TGFBR3	1	1	0	0
7077	TIMP2	1	1	0	0
7080	NKX2-1	0	1	0	1
7113	TMPRSS2	0	1	0	1
7128	TNFAIP3	0	1	0	1
7157	TP53	0	1	0	1
7168	TPM1	1	0	1	0
7175	TPR	0	0	1	1
7248	TSC1	0	1	0	1
7249	TSC2	0	1	0	1
7253	TSHR	0	1	0	1
7268	TTC4	0	1	1	0
7325	UBE2E2	1	1	0	0
7356	SCGB1A1	0	1	1	0
7428	VHL	0	1	0	1
7430	EZR	0	1	0	1
7450	VWF	0	1	1	0
7486	WRN	0	1	0	1
7490	WT1	0	1	0	1
7507	XPA	0	1	0	1

7508	XPC	0	1	0	1
7515	XRCC1	0	1	1	0
7531	YWHAE	0	1	0	1
7704	ZBTB16	0	1	0	1
7782	SLC30A4	1	1	0	0
7849	PAX8	0	1	0	1
7918	GPANK1	0	1	1	0
7979	SHFM1	0	1	1	0
8028	MLLT10	0	1	0	1
8030	CCDC6	0	1	0	1
8031	NCOA4	0	1	0	1
8091	HMGA2	0	1	0	1
8115	TCL1A	0	1	1	1
8242	KDM5C	0	0	1	1
8314	BAP1	0	1	0	1
8435	SOAT2	0	1	1	0
8648	NCOA1	0	1	0	1
8651	SOCS1	0	1	0	1
8764	TNFRSF14	0	1	0	1
8829	NRP1	1	1	0	0
8880	FUBP1	0	0	1	1
8887	TAX1BP1	0	1	1	0
8915	BCL10	0	1	0	1
8929	PHOX2B	0	1	0	1
9150	CTDP1	0	1	1	0
9321	TRIP11	0	1	0	1
9338	TCEAL1	0	1	1	0
9401	RECQL4	0	1	0	1
9510	ADAMTS1	1	0	1	0
9817	KEAP1	0	1	1	0
9935	MAFB	0	1	0	1
9968	MED12	0	0	1	1
9969	MED13	0	1	1	0
10000	AKT3	1	1	0	0
10039	PARP3	0	1	1	0
10142	AKAP9	0	1	0	1
10186	LHFP	1	0	0	1
10320	IKZF1	0	1	0	1
10397	NDRG1	0	1	0	1
10413	YAP1	0	1	1	0
10499	NCOA2	0	1	0	1
10892	MALT1	0	1	0	1
11064	CNTRL	0	1	0	1
11197	WIF1	0	1	0	1
11200	CHEK2	0	1	0	1

23092	ARHGAP26	0	1	0	1
23305	ACSL6	0	1	0	1
23332	CLASP1	0	1	1	0
23405	DICER1	0	1	0	1
23710	GABARAPL1	1	1	0	0
25818	KLK5	0	1	1	0
26136	TES	1	1	0	0
27063	ANKRD1	0	1	1	0
27324	TOX3	1	1	0	0
27347	STK39	1	1	0	0
27436	EML4	0	1	0	1
29844	TFPT	0	0	1	1
29998	GLTSCR1	0	1	1	0
30061	SLC40A1	1	1	0	0
50802	IGK@	0	1	0	1
50939	IMPG2	0	1	1	0
51094	ADIPOR1	0	1	1	0
51100	SH3GLB1	1	1	0	0
51199	NIN	0	1	0	1
51378	ANGPT4	0	1	1	0
51684	SUFU	0	1	0	1
51755	CDK12	0	1	1	1
54790	TET2	0	1	0	1
55234	SMU1	0	1	1	0
55259	CASC1	0	1	1	0
55294	FBXW7	0	1	0	1
55862	ECHDC1	0	1	1	0
55971	BAIAP2L1	0	1	1	0
58480	RHOU	1	1	0	0
63976	PRDM16	0	1	0	1
64116	SLC39A8	1	1	0	0
64241	ABCG8	0	1	1	0
64746	ACBD3	0	1	1	0
64805	P2RY12	0	1	1	0
79723	SUV39H2	0	1	1	0
79728	PALB2	0	1	0	1
80312	TET1	0	1	0	1
80380	PDCD1LG2	0	1	0	1
81623	DEFB126	0	1	1	0
83892	KCTD10	1	1	0	0
83990	BRIP1	0	1	0	1
84441	MAML2	0	1	0	1
84651	SPINK7	0	1	1	0
92689	FAM114A1	0	1	1	0
121227	LRIG3	0	1	0	1

139285	FAM123B	0	1	0	1
140767	NRSN1	0	1	1	0
143187	VTI1A	0	1	0	1
151393	FAM82A1	0	1	1	0
171023	ASXL1	0	1	0	1
221895	JAZF1	0	1	0	1
285733		0	1	1	0
374378	GALNTL4	0	1	1	0