TRANSCRIPTIONAL PROFILING OF PLURIPOTENT AND MULTIPOTENT STEM CELLS TO DECIPHER PLURIPOTENCY AND LINEAGE SPECIFICATION

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

SHIKHA SHARMA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

WEI-SHOU HU, Advisor

April, 2012

Acknowledgements

With heartfelt gratitude, I would like to acknowledge the support of my advisor Dr. Wei-Shou Hu. His patience and encouragement has been invaluable in shaping up this work. He has been a true mentor to me in every way. I would also like to thank Dr. Catherine Verfaillie for her guidance at every step of my graduate life. I feel fortunate to have had the opportunity of interacting with her and the Verfaillie lab over the course of my PhD.

I would also like to thank the other members of my thesis committee, Dr. Robert Tranquillo and Dr. Friedrich Srienc in the Department of Chemical Engineering and Materials Science, and Dr. Nobuaki Kikyo in the Department of Medicine for their help and useful critique.

I would like to thank present and former members of Hu group: Kartik Subramanian, Ravali Raju, Yonsil Park, Jason Owens, Andrew Yongky, Dong Seong Cho, Salim Charaniya, Anne Kantardjieff, Siguang Sui, Nitya Jacob, Bhanu Chandra Mulukulta, Huong Le and Kathryn Johnson. I would also like to thank my colleagues at the Stem Cell Institute: Lucas Greder, James Dutton, Ananya Banga, Natarajan Bhanu, Jessie Browers, Beth Lindborg, Lauri Anderson and Genya Gekker. Their help and camaraderie have made the last few years very enjoyable and defined my graduate experience. I would also like to thank the following people in the Verfaillie lab for their intellectual inputs: Antonio Lo Nigro, Tineke Notelaers and Philip Roelandt.

The following friends in Minnesota deserve a very special thank you: Natarajan Bhanu, Rajiv Ranjan, Ravali Raju, Kartik Subramanian, Lindsay Swanson, Kat Volzing, Kathy Mahan, Bhanu Chandra Mulukutla, Nitya Jacob, Vivek Kalihari and Disha Kalihari,

My family was so far away but their support and love was always wrapped around me. I fall short of words in acknowledging their contribution here. Finally, I can't say enough about the strength and encouragement that I have received from my husband, Salim Charaniya. He is my biggest strength, my most fierce critic and my best friend, all in one.

Dedication

To my parents-Ravidutt and Sharda.

Abstract

Stem cells hold great promise for the fields of regenerative medicine, gene therapy and disease modeling. Understanding the transcriptional machinery involved in their maintenance is critical to their successful isolation and experimentation. Careful statistical analysis of high throughput transcriptome data can provide novel insights into the gene networks and patterns active in these cells. Public repositories are a source of gene expression data from various studies involving stem cells. This expression data can be overlaid on functional interactions maps of the genome to predict functional association. Further on, comparison of stem cells of different potencies can help identify key genes involved in the maintenance of pluripotency. The hypotheses derived from such transcriptional profiling can be tested experimentally to confirm expression and interactions.

Transcriptome data from studies involving human and mouse pluripotent stem cells was collected from repositories such as GEO and Arrayexpress (EBI). Non-negative matrix factorization was used as a dimensionality reduction tool to detect biological patterns and clusters in the data. Following the classification of data into biologically meaningful classes, a 'metagene' profile characteristics of pluripotent stem cells was determined in both species. Reverse engineering was performed for predicting gene networks and signaling 'hubs'. An algorithm was also developed to overlay this predicted gene signature onto functional networks that combine a large amount of genetic and genomic data from various sources, for detecting small subnetworks that are conserved in expression in the pluripotent stem cells of both species.

Embryonic and induced pluripotent stem cells are considered as in vitro counterparts of pluripotent cells seen in the early embryo, namely the inner cell mass (ICM) and the epiblast. Multipotent adult progenitor cells (MAPCs), although isolated from the bone marrow of an adult rat, bear a striking similarity with another cell types in early embryonic development-the primitive endoderm or nascent hypoblast cells. Due to the developmental proximity of the pluripotent cells and the primitive endoderm cells in the early embryo, MAPCs have been used as a model system for probing gene interactions in pluripotent cells. On the basis of comparative transcriptome analysis, as well as, experimental studies, a model of gene regulation in MAPCs has been developed. The study of dynamics of this gene network provides novel insights into the transcriptional regulation of key pluripotency-associated genes.

Also, using the concepts of cellular reprogramming, these multipotent stem cells (MAPCs) have been reprogrammed to a pluripotent state. This represents a unique reprogramming system where an Oct4 expressing extraembryonic cell has been transformed to an embryonic stem cells (ESC) like state. While this was performed using the traditional reprogramming cocktail consisting of Oct4, Sox2, Klf4 and c-Myc, future studies are likely to narrow down this number to one or two genes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Transcriptional profiling of pluripotent and multipotent stem cells

Stem cells hold tremendous promise for regenerative therapy, disease modeling and gene therapy. The last few decades have seen an exponential increase in research involving both embryonic and adult stem cells. The isolation of induced pluripotent stem cells (iPSCs) has caused this level of interest to further escalate. In additions to the applications in medicine that are mentioned above, these cells are a novel resource for understanding development and lineage specifications. Pluripotent stem cells are often seen as *in vitro* counterparts of early embryonic cell types, or lineages, such as the ICM and the epiblast. Thus, these cells can be used as a model system to understand the gene expression changes that happen as cells in an embryo progressively differentiate. With the derivation of iPSCs particularly, it is now possible to capture these gene expression changes in either direction: loss, as well as, gain of pluripotency.

Similarly, multipotent and adult stem cells are seen as precursors to the somatic cell types of a tissue or lineage. Studies involving maintenance and differentiation of these cells provide useful details about their *in vivo* molecular niche and differentiation cues. Bone marrow-derived, multipotent stem cells called MAPCs were derived by the Verfaillie lab [1, 2]. These cells have a unique gene expression profile in that they express markers of pluripotent stem cells, as well as, the extraembryonic endoderm lineage. Similar cells have also been derived from the ICM of the developing embryo [3] and there is evidence pointing to the presence of such cells *in vivo* during early development [4]. Recently, Verfaillie and colleagues have shown that MAPC culture conditions allow successful isolation of MAPC-like extraembryonic endoderm precursor cells from the rat embryo (unpublished work). All these reports suggest that MAPCs are highly similar to the nascent hypoblast of the developing embryo. The nascent hypoblast or primitive endoderm precursor cells emerge from the ICM and are closely related to the pluripotent ICM. The pluripotent ICM and the epiblast are, in turn, highly similar to mouse and human ESCs, respectively.

1.2 Dissertation Focus

This dissertation will outline work performed towards deciphering a gene expression signature for pluripotent human and mouse stem cells by performing a meta-analysis using transcriptome data available in public repositories. In broad terms, studies selected for meta-analysis involved differentiation of ESCs, reprogramming of somatic cells to pluripotency, and in some cases, adult stem cells. A unique methodology was developed to assimilate and process this data to arrive at representative gene signatures. The methodology primarily hinges on performing dimensionality reduction on normalized microarray data to discover trends and biological classes in the data. This processed data has further been used for predicting signaling ‘hubs’ and developing genetic networks through reverse engineering. Further on, the application of functional ‘interactomes’ to processed data for asking biologically relevant questions such as conserved signaling between species has been discussed.

As discussed in section 1.1, MAPCs are closely related to pluripotent ESCs in their gene expression and embryonic developmental context. This developmental proximity of MAPCs to the pluripotent cells of early embryo has been used to gain further insights into the establishment and maintenance of pluripotency. Through comparison of the transcriptome of ESCs and MAPCs, a gene network has been developed and some components of this network have been probed through tools such as shRNA knockdown. Lastly, it has been demonstrated that MAPCs can be reprogrammed to a pluripotent state by inducing expression of exogenous genes Oct4, Sox2, Klf4 and c-Myc. This represents the first report of switching fate of an extraembryonic cell to that of a pluripotent cell of early embryo.

1.3 Thesis Organization

Chapter 2 of this thesis offers a literature review of important scientific work performed in the field over the last few decades. It has been written with the goal of providing a broad overview of the current state of the field while also offering a more specific focus on topics more relevant to the research discussed in this dissertation.

Chapter 3 presents a detailed list of all materials and methods used in the work conducted.

Chapter 4 discusses the meta-analysis performed on transcriptome data of human and mouse pluripotent stem cells. While published statistical tools used in this analysis have been discussed in Chapter 2, the methodology of application and the results obtained have been discussed here. Chapter 5 is an extension of this work and discusses the application of functional interaction maps to the transcriptome data for discovering conserved expression in human and mouse pluripotent stem cells.

Chapter 6 elucidates the primitive endoderm precursor-phenotype of MAPCs through comparison of their transcriptome with that of pluripotent rat ESCs, as well as, embryo-derived nascent hypoblast-like cells. Chapter 7 further delves into the signaling involved in maintaining MAPCs and draws up a potential signaling network responsible for their maintenance. A particular feature of this network, the gene expression switch between two important genes, Nanog and Gata6, is the focus of further analysis. Chapter 8 finally shows how fate of a seemingly extraembryonic cell type can be switched by reprogramming by induction of exogenous factors.

Finally, Chapter 9 summarizes the main conclusion of this thesis and discusses questions raised by this work that can be answered by further research.

CHAPTER 2
Literature Review

2.1 Early Mammalian Development

Stem cells are tied very intricately with mammalian development and fate determination. The understanding about how extrinsic signals, combined with internal transcription program of the cells, determine cell fate during development has contributed significantly to isolation and propagation of many progenitor populations and stem cells. After fertilization, a series of cleavage events happen, resulting into a 16-cell morula (Figure 1).

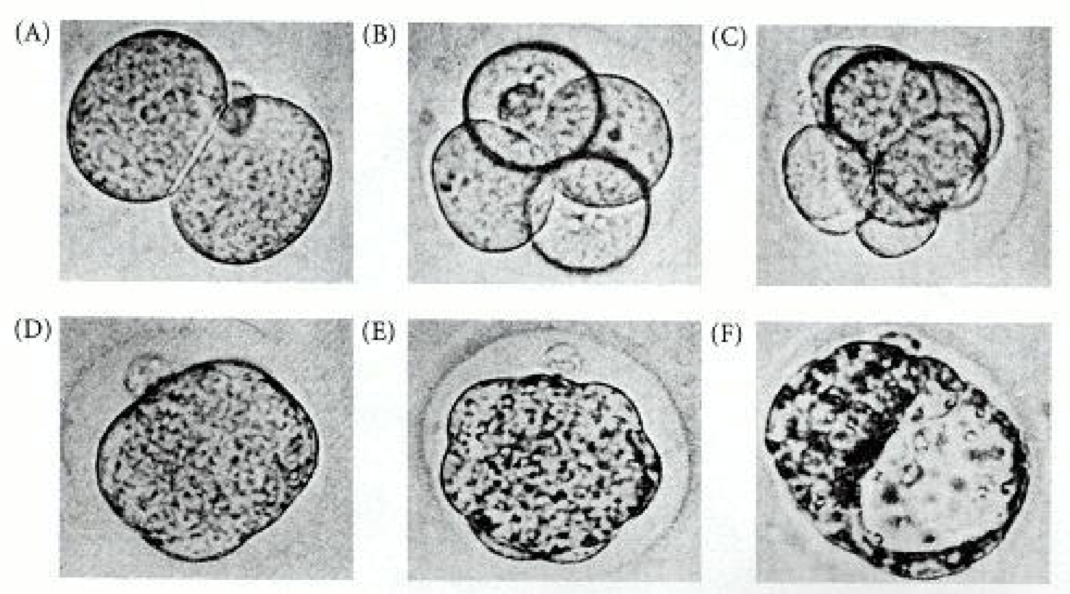


Figure 1. The cleavage of a single mammalian embryo in vitro. (A) 2-cell stage. (B) 4-cell stage. (C) Early 8-cell stage. (D) Compacted 8-cell stage. (E) Morula. (F) Blastocyst.[5]

The outer cells of the morula (Figure 1(E)) predominantly give rise to the large trophoblast cells while the internal cells give rise to the ICM that give rise to the embryo proper. During a process called cavitation, the trophoblast cells secrete a fluid into the morula that creates a cavity and causes the ICM to get positioned on one side of this structure which is called the blastocyst, as shown in Figure 1(F). This represents the first lineage segregation event in mammalian development where, the resulting trophoblast cells do not contribute to the embryo proper but form the chorion and placenta (Figure 2) that help with the imminent implantation of the embryo. The ICM is called pluripotent because of their ability to form every cell type of the embryo proper.

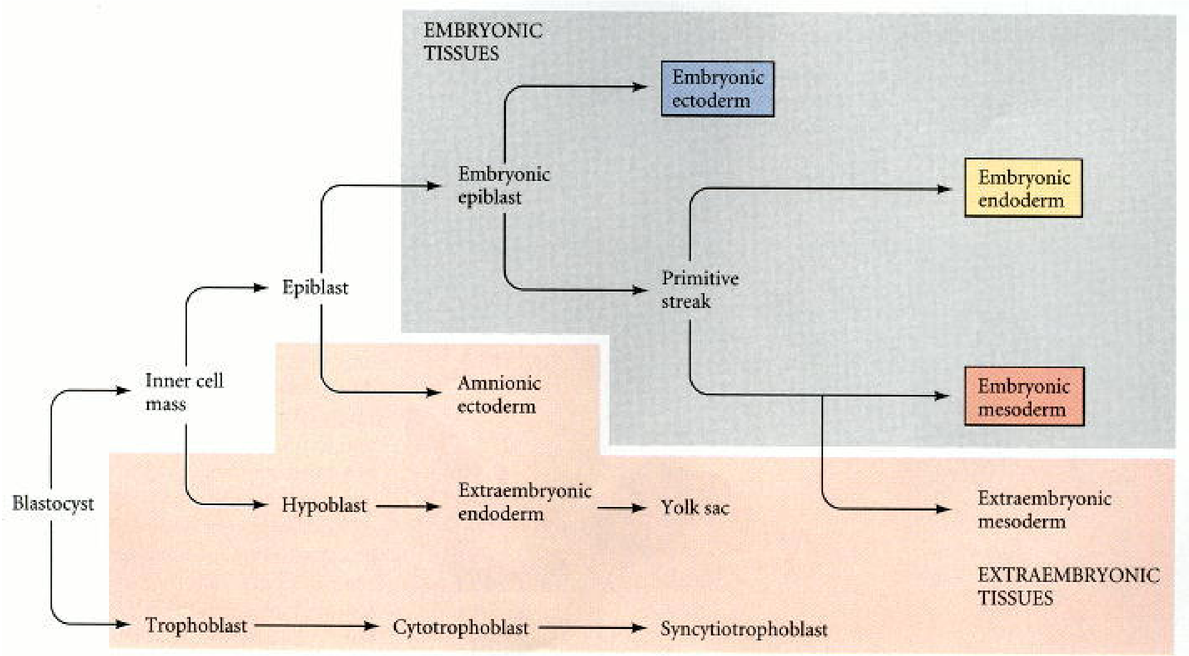


Figure2. Schematic outlining lineage segregation in early mammalian development [5].

Within the ICM, the first segregation of cells happens with differentiation to the hypoblast and the epiblast. The hypoblast, also referred as the primitive endoderm, is another extraembryonic lineage (like Trophoblast) that does not contribute to the embryo proper and plays a supporting role by forming the yolk sac. The embryonic epiblast gives rise to all the three germ lineages: ectoderm, endoderm and mesoderm.

Stem or progenitor cells have been isolated from many different stages of development. Successful derivation and propagation has been done for cells that are characteristic of the trophoblast (trophoblast stem cells), ICM (ESCs), hypoblast (hypoblast stem cells) and epiblast (epiblast stem cells), amongst others. The in vitro derivation and culture of these cells have contributed immensely to our understanding of gene expression at the different stages of mammalian development, as well as, the microenvironment signals that are required to support these cells in vivo.

2.2 Stem Cells

A stem cell is defined as a cell type that has two important defining features: self-renewal and differentiation. Self-renewal is defined as the ability of a cell to divide long term while maintaining its phenotype. Differentiation, on the other hand, is the ability to give rise to other, specialized cell types upon exposure to certain signaling cues. The earliest clinical proof of the existence of such cells came from bone marrow transplants where it was found that hematopoietic stem cells (HSCs) present in the bone marrow of a donor were seen to exhibit long-term, multilineage repopulating activity in the host. Since then, many types of cells exhibiting stem cell-like properties have been discovered in the adult organism and the developing embryo. Stem cells are very commonly classified on the basis of the tissue or cell type of isolation. One such classification is into adult and ESCs. Another property commonly used to describe and classify stem cells is called 'potency'. This refers to the ability of the cell to give rise to one or more multiple, specialized cell types. While a unipotent cell can give rise to just a single differentiated cell type, a pluripotent cell, such as ESCs, can, in principle, give rise to every cell type of an adult organism.

2.2.1 Adult stem cells

Cells that are responsible for tissue homeostasis or repair/regeneration following an injury have been found in many tissues such as bone marrow, digestive tract, heart and kidney, amongst others. These cells often exist in their specific microenvironment or niche, which regulates their self-renewal or differentiation. While their differentiation potential is generally believed to be restricted to giving rise to cells of the particular tissue or organ that they reside in, the ability of some adult stem cells to transdifferentiate to cell types characteristic of other tissues and lineages have also been reported. Successful and reproducible isolation of adult stem cells relies on identification of reliable surface markers that can be used to identify these cells. Hematopoietic stem cells (HSCs) continue to be the best characterized and most clinically relevant adult stem cells to date. Human HSC are usually characterized by the expression of CD-34 surface antigen ($CD34^+$) (Murray, et al., 1995). Other markers that have been shown to be expressed by these cells are Thy-1 and CD-105 [6, 7]. These cells are able to differentiate into multiple, specialized hematopoietic cells such as lymphocytes, nature killer cells, megakaryocytes etc. Hematopoietic stem cells are not the only adult stem cell type that has been isolated from the bone marrow. Mesenchymal stem cells (MSCs), endothelial progenitor cells and other types of progenitor cells have also been isolated. MSCs, first isolated from the bone marrow [8], have now been isolated from multiple tissues such as brain, liver, kidney, muscle etc. [9, 10]. They have been shown to differentiate to cells of not only the mesodermal lineage such as bone and fat, but also, endodermal [11] and neuroectodermal lineage [12, 13]. Additional characteristics of these cells, such as being immunosuppressive, have generated immense interest in the clinical application of these cells for therapy. There are multiple ongoing clinical trials for assessing the regenerative potential of MSCs for conditions such as Graft versus host disease, acute myocardial infarction, and severe limb ischemia, among others [14]

Multiple neural precursor cells have been isolated from the adult nervous systems such as neural stem cells, neuronal progenitor and glial progenitor cells [15]. Neural stem cells, lining the lateral ventricles of the brain and the hypocampus, are multipotent in comparison to the two progenitor cells types mentioned above, and are able to give rise to both neurons and glia. The proliferation, migration and differentiation potential of these cells can be mobilized endogenously upon injury, as well as, exogenously through the use of mitogens like brain-derived neurotrophic factor (BDNF) [16, 17].

2.2.2 Embryonic stem cells

ESCs are derived from the ICM of an embryo. These cells differentiate to all the three germ layers and are thus considered pluripotent. ESCs, commonly referred as ESCs, were first derived by Evans and Kaufman from mouse embryos by culturing mouse blastocysts on layer of STO fibroblasts [18]. The ‘pluripotential’ cells, derived from the outgrowth of the ICM, differentiated in vitro into cell types representing multiple tissues and formed teratomas upon subcutaneous injection into syngeneic male mice. Subsequent studies showed that these cells express pluripotency-associated genes like Oct4, Sox2, Nanog, Rex1 that are also expressed by the ICM. Investigations into the factors responsible for maintain the self-renewal of these cells in vitro showed that a cytokine called Leukemia inhibitory factor (LIF) is able to replace mouse embryonic fibroblasts or feeders in mESC culture [19, 20]. Smith and colleagues later showed that LIF, in collaboration with Bone morphogenetic proteins (BMP2 or BMP4) is able to sustain mESC in absence of feeders and serum. LIF signals through STAT3, while BMP proteins signal through the SMAD pathway and activate Inhibitor of differentiation (Id) genes. Both of these signaling cascades are necessary and sufficient for self-renewal of mESC. Identification of these growth factors has made it possible to use fully defined culture medium that was serum free. The undefined nature of serum had been a cause of concern in the field because it interfered with the identification of critical growth factors for maintenance of self-renewal and prevention of differentiation. These early reports of mESC derivation used the 129 strain of mouse. The non-obese diabetic (NOD) mouse is another strain that has been widely used owing to it being an animal model for human insulin-dependent diabetes mellitus (IDDM). Efforts to derive mESC lines from this strain were surprisingly, largely unsuccessful though and it was believed that this was a ‘non-permissive’ or ‘refractory’ strain for ESC derivation although the genetic factors that resulted in this non-permissiveness were not understood. It was not until much later, when the concept of epiblast stem cells and ‘naïve vs. primed’ state of pluripotency was discovered, that it became possible to derive mESCs from this strain that were able to give germline transmission (as discussed in next section).

Pluripotent stem cells from the human blastocyst, termed human ESCs, were derived much later in 1998 by Thomson et al. [21]. The human cells were similar to their mouse counterparts in many ways such as high nucleus to cytoplasm ratio, expression of genes such as Oct4, Nanog etc., and differentiation to the three germ lineages in vivo and in vitro. But, there were also considerable differences between these pluripotent cells from the two species such as expression of SSEA-3 and SSEA-4 by the human cells but not the mouse ones, flattened morphology of human ESCs in contrast to the dome-shaped morphology of the mouse counterparts and, intolerance of human ESCs to passaging as single cells, amongst others. The most prominent difference though, was that while mouse ESCs require LIF and BMPs to self-renew [20, 22], in human ESCs, STAT3 signaling seemed redundant and 'stemness' was shown to depend entirely on Activin/Nodal and bFGF signaling cascade. Subsequently, it was discovered that these two cell types do not exactly represent the same developmental stage. While mouse ESCs have been shown to express markers characteristic of the ICM stage, such as Rex1 and Stella, human ESCs do not. Human ESCs are more similar to another type pluripotent cell type called mouse epiblast stem cells, than to mouse ESCs.

2.2.2.1 Rat ESCs

After the successful derivation of ESCs from mouse and human embryos, it was expected that the derivation of ESCs from other species, especially rat, would follow suit. Attempts to derive pluripotent cells from the rat embryo using mESC culture conditions resulted in rapid loss of Oct4 within 4-5 days post derivation from E 4.5 rat blastocyst [23]. In this report, the authors were able to derive an Oct4-negative cell line that resembled mESCs morphologically but the cells had a tendency to differentiate spontaneously to the extraembryonic, parietal endoderm and trophectoderm lineages. Another study derived cell lines from rat blastocyst that expressed Oct4 but the level of expression in was much lower than what is seen in mouse ESCs [24]. Also, no differentiation potential or chimera formation was shown for these cells. At this time, a widely accepted notion was that external signaling cues were required to maintain the self-renewal of pluripotent ESCs in vitro. So, it was believed that LIF+BMP (in case of mESCs) and Activin+bFGF (in case of human ESCs) were the all important, extrinsic signaling cues that acted as the stimuli for self-renewal. A seminal study by Smith and Colleagues in 2008 showed that mESCs had an innate self-renewal program and the role of LIF and BMP was to suppress differentiation by acting downstream of MAPK pathway [25]. The authors used a defined, serum-free medium, N2B27, and used a combination of three signaling inhibitors to maintain the cells in absence of LIF and BMP. The three inhibitors used were PD0325901, SU5402 and CHIR99021.

PD0325901 is a small molecule targeting mitogen-activated protein kinase (MAPK/ERK kinase or MEK) with potential antineoplastic activity. PD0325901, a derivative of MEK inhibitor CI-1040, selectively binds to and inhibits MEK, which may result in the inhibition of the phosphorylation and activation of MAPK/ERK. SU5402 Inhibits the tyrosine kinase activity of fibroblast growth factor receptor 1 (FGFR1) and also inhibits aFGF-induced tyrosine phosphorylation of ERK1 and ERK2. CHIR99021 is a selective inhibitor of glycogen synthase kinase 3β (GSK-3β). Unlike other potent inhibitors of GSK-3, CHIR99201 does not exhibit cross-reactivity against cyclin-dependent kinases (CDKs) and shows a 350-fold selectivity toward GSK-3β compared to CDKs.

While the first two (PD+SU) were shown to be necessary for suppressing the differentiation-inducing effect of FGF4-ERK signaling while CHIR, a GSK3β inhibitor, that was shown to promote clonal survival and growth. Based on the idea that the successful derivation of ESCs may require only suppression of differentiation signals rather than inductive self-renewal signals, Smith and Ying were able to derive rat ESCs in N2B27 basal medium supplemented with 3i (PD+CH+SU) and even just 2i (PD+CHIR)[26, 27]. These cells resembled mouse ESCs in their dome-shaped morphology and high nucleus to cytoplasm ratio. The cells successfully gave rise to chimeras and germline transmission was shown. An important observation that the authors made was that these rat ESCs could not be maintained at all in medium with LIF and serum which are commonly used for mESC culture. 3i or 2i was necessary to prevent differentiation of rat ESC even in the presence of LIF. Another important difference between the rat ESCs and their mouse counterparts, that was observed in these studies was, that in serum with the absence of LIF and 3i/2i, rat ESCs differentiated to cells with features of the hypoblast lineage, while mESCs did not show differentiation to the extraembryonic hypoblast (primitive endoderm) lineage. This falls in line with the observed differences in embryogenesis in the two species. The rat epiblast (d6.5) predominantly gives rise to parietal endoderm cells of the hypoblast lineage upon in vitro culture while the developmentally synchronous mouse epiblast (d5.5) seems to have lost the ability to produce parietal endoderm [28]. Another finding that reinforces the idea that, despite the seemingly similar stages of early development in the two species, there are considerable differences in the potency of the cells, showed that the d8 egg cylinders of rat give rise to yolk sac carcinomas upon ectopic transplantation under kidney capsule [29]. The mouse egg cylinders give rise to teratocarcinomas that contain pluripotent embryonic carcinoma cells.

2.2.3 Epiblast stem cells

In 2007, two independent reports described the derivation of pluripotent stem cells from post-implantation mouse embryos in human ESC culture conditions [30, 31]. Thes stem cells, termed EpiSCs, were derived by isolating the post-implantation epiblast layer, instead of the pre-implantation ICM used for the derivation of mouse ESCs. Mouse EpiSCs, resembled human ESCs in their morphology, gene expression and epigenetic signature and, like Human ESCs, relied on Activin A/Nodal signaling to maintain for self-renewal but not LIF/Stat3 signaling (like mouse ESCs). Remarkably, mouse ESC culture conditions could not be used for the derivation and propagation of cells from this stage of embryonic development, suggesting, for the first time, that human ESCs (which these cells resembled so closely) could perhaps represent a later stage in embryonic development. mEpiSCs were shown to be pluripotent from their in vitro differentiation and teratoma formation potential. Interestingly, both studies observed that mEpiSC were unable to integrate into pre-implantation blast cyst or the morula stage of embryo to produce chimaeras efficiently. This was largely attributed to the developmental asynchrony between the later stages of mammalian development that mEpiSC represent versus the earlier, pre-implantation stage that mouse ESCs are believed to represent.

mEpiSCs do not express markers characteristic of ICM/ESCs such as Pecam and Tbx3 but instead express genes characteristic of the late epiblast such as Otx2, Brachyury and Fgf5. Similar to human ESCs, mEpiSCs do not show active histone marks (H3K4me3) around the transcriptional start site of ICM-specific genes such as Stella. Besides their identical dependence on Activin A/Nodal signaling, both human ESCs and mouse EpiSCs, differentiate to trophectoderm upon induction with BMP4, while mouse ESCs show little capacity to differentiate to this lineage. Also, both cell types show inactivation of X chromosome in female cell lines in stark contrast to mouse ESCs or the ICM where the chromosome is active. While mouse EpiSCs and human ESCs have many similarities, as discussed above, they also have some differences in gene expression. Human ESCs express ICM-specific genes like Klf4 and Rex1, not expressed by mouse EpiSCs while they lack expression of epiblast marker FGF5. Also, while the Activin pathway contributes to maintain expression of Nanog in both cell types, the FGF signaling has divergent roles [32]. In human ESCs, the FGF2 signaling cooperates with Smad2/3 to promote self-renewal FGF2 stabilizes the mEpiSC fate by suppressing neuroectoderm differentiation, as well as, preventing reversion to an ESC-like state.

The derivation of EpiSCs reconciled some of the observed differences between human and mouse ESCs by attributing them to the difference in developmental stage that the two cell types are derived from as opposed to species-specificity. This combined with the derivation of induced pluripotent stem cells (next section) reinforced the idea that the successful derivation/identification of a stem cell depends as much on the conducive culture conditions as it does on the parent tissue or cell type of isolation. A natural question that arose at this point was the possibility of inter-conversion between ESCs and EpiSCs and whether the understanding of mammalian development can lend to this. The scientific advancement brought upon by fortuitous, simultaneous discovery of induced pluripotent stem cells (iPSCs) helped to elucidate this (as discussed in next section). As expected, EpiSC do not integrate well into the embryo to give rise to chimeras. This could be due to the developmental asynchrony.

Using their mouse EpiSC derivation protocol, Brons et al. were able to derive EpiSCs from E 7.5-7.75 rat embryos. This was the first report of derivation of pluripotent stem cells from rat embryo that had so far been considered a ‘nonpermissive’ species for ESC derivation. This was due to the fact that all the efforts to derive rat ESCs from rat blastocysts, under mESC culture conditions, had failed to produce bona-fide, Oct4 expressing stem cells. As mentioned above, the eventual derivation of ESCs from rat was achieved using a specific chemical environment with inhibitors of signaling (2i). The authors also used their protocol to successfully derive EpiSCs from the ‘nonpermissive’ non obese diabetic (NOD) strain of mouse. Efforts to derive ESCs from this popular strain of mouse had also been unsuccessful till this point. This brought forth the idea of whether some strains or species are more likely to adapt to one pluripotent state versus the other and that there could be other states of pluripotency that can be observed if the stabilizing determinants could be identified for those states.

2.2.4 Naïve versus Primed state of pluripotency

With the derivation of epiblast stem cells, the idea that there can be two or multiple states of pluripotency in the developing embryo had emerged. Pre implantation epiblast-derived mouse ESCs (mESCs) and post implantation epiblast-derived mouse epiblast stem cells (mEpiSCs) can be imagined to represent a naïve and a primed state of pluripotency, respectively [33]. Nichols and Smith put forth the idea that ESCs represent immortalization of the pre implantation epiblast in development and, like the latter, can incorporate into blastocyst and contribute to development. Once the embryo undergoes implantation, the post implantation, gastrulating epiblast cells are incapable of contributing to blastocyst chimera [34]. Repeated attempts to derive mouse ESCs from this stage of the developing embryo had failed when it was discovered that pluripotent cells that resemble human ESCs could instead be derived from this embryonic stage (as discussed above). While it is true that EpiSCs are derived from the post implantation embryo and the human ESCs are derived from cultured blastocysts, this contradiction can be resolved by understanding that the cells in culture are not bona fide representatives of the stage that they are derived from. One can argue that the cells of the blastocyst continue on a differentiation program post derivation and eventually get stabilized in permissive culture conditions.

Further investigations into the reversibility of two states led Guo et al. to discover that mouse ESCs could readily be converted to EpiSCs in response to culture conditions but the opposite required genetic manipulation. The authors noted that ESCs, after being transferred to EpiSC culture conditions, rapidly downregulated markers of ICM such as

KLF4, Rex1 and Nr0b1. The cells also showed X chromosome inactivation and shift in control of Oct4 expression from distal to proximal enhancer. While this transition occurs spontaneously without any genetic or epigenetic manipulation, the reverse could be achieved only by constitutive expression of Klf4 in cells transferred to mESC culture conditions. However, it is important to note that the efficiency of this transition is still about 1%. So, the reversal of state still only happens in a small number of cells expressing Klf4. Hanna et al. were able to get ESC-like cells from the 'nonpermissive' NOD strain using the same approach [35]. The authors converted derived-EpiSCs to ESCs by constitutive expression of Klf4 or c-Myc. The importance of these two factors will be discussed in more detail in the following section on cellular reprogramming.

Guo and Smith also conducted a *piggybac* transposition based genetic screen to identify factors that could promote reprogramming of the EpiSC state to the ground state of pluripotency, i.e., the ESC state, and identified two nuclear receptors, Nr5a1 and Nr5a2, whose expression allowed cells to transition from the former state to the latter [36]. The combined expression of Klf4 and the Nr5a factors further increased this frequency of reprogramming. An interesting discovery made by the authors here was that continued expression or over-expression of Oct4 in the EpiSCs does not facilitate this transition despite the fact that Nr5a factors are known to be regulators of Oct4 [37].

Recently, a role of Wnt signaling in the transition between mouse ESCs and mouse EpiSCs was discovered [38]. It was observed that cultured mESCs (on MEFs, in presence of LIF), differentiated to EpiSC phenotype in the presence of the small molecule inhibitor, IWP2, which interferes with the ability of cells to produce Wnt in autocrine fashion. The authors were able to confirm that presence of Wnt3a in EpiSC medium prevented ESCs from differentiating into the former, though Wnt3a alone does not sustain the ESC state and LIF is needed for long term self-renewal. The importance of Wnt signaling for maintenance of ESC self renewal has been known for a long time and GSK3 inhibitor CHIR99021 is an integral part of the 3i/2i medium for ESC derivation but the association of Wnt signaling with the naïve and primed states of pluripotency is a new and interesting finding.

2.3 Cellular Reprogramming and Induced Pluripotent Stem Cells

The technique of nuclear transfer transplantation or Somatic Cell Nuclear transfer (SCNT) was developed by King and Briggs [39, 40] in an attempt to understand the extent of differentiation of nuclei from cells of the late-stage embryos. The authors developed the technique to isolate nuclei from Xenopus cells and transfer them successfully into the enucleated egg without damage. These, and other seminal experiments conducted afterwards, showed that development does not impose any irreversible genetic changes on differentiated cells and that it was possible to somewhat reverse the epigenetic changes that had occurred under correct experimental conditions [41]. In 1996, Ian Wilmut, Keith Campbell and colleagues derived live sheep offsprings by transferring the nucleus from an embryo-derived cell line into an enucleated oocyte [42]. This was the first report that showed successful 'reprogramming' of the nucleus of a mammalian cell by the enucleated host cell, resulting in the completion of term and live birth. Soon after, the authors showed that this procedure could be reproduced by taking the nucleus of a somatic, mammary cell too, highlighting the success of cellular reprogramming[43]. These and other experiments involving SCNT confirmed that the genetic and epigenetic signature of a somatic cell is not set in stone. Upon successfully receiving the right cues from the cytoplasm of an oocyte, this signature can be reset, allowing the cell to enter the development program and give rise to all cell types of an adult organism. Fascinating as this was, the experimental technique itself is very sensitive to factors such as the development stage and cell cycle stage of the donor cells, resulting in very low efficiency [44]. Nevertheless, many labs were able to repeat this is different organisms such as cattle, mice, pigs, rabbits etc. [45-47] and the idea of cellular remodeling to recreate a state of totipotency was promoted.

Almost 10 years later, Shinya Yamanaka and colleagues showed that adult somatic cells from mice can be reprogrammed to a pluripotent, stem cell-like fate by exogenous induction of four transcription factors: Oct4, Sox2, Klf4 and c-Myc [48]. Yamanaka and colleagues screened 24 candidate genes that are expressed at high levels in mouse ESCs and discovered that the combination of these four genes, often referred as OSKM, is necessary and sufficient to kick start a cellular program that remodels the epigenome/genome of the somatic cell. The cells, termed induced pluripotent stem cells or iPSCs, were morphologically similar to ESCs, displayed highly similar gene expression profile and differentiated in vivo, as well as in vitro, into the three germ layers. The experimental procedure involved transducing somatic cells with transgenes expressing the four transcription factors followed by transferring the cells to ESC culture conditions (in mESC medium containing LIF on a layer of mitomycin-treated STO fibroblasts). mESC-like colonies could be observed in 2-3 weeks post transgene transduction. The concept of inducing fate change through ectopic expression of genes had been demonstrated before. As an example, Xie et al. had shown that differentiated beta cells can be reprogrammed through forced expression of C/EBP alpha and C/EBP beta to become macrophages [49]. What was novel about iPSC reprogramming was that it showed that the potency of somatic cells can be changed to produce immature, unspecialized stem cells. It was determined that the transgene expression had been silenced in the colonies and the successfully reprogrammed cells had setup their endogenous signaling network that allowed maintenance of a pluripotent phenotype. This discovery spurred an immediate research interest into the mechanism of reprogramming, as well as, utilization of this approach to create pluripotent stem cell out of different patient/disease-specific cells. The utilization of iPSC technology for cell therapy has entailed investigations into derivation of iPSC cells from different starting populations of cells and tissues and also into efficiently generating transgene-free iPSCs. iPSC discovery has created a new and unprecedented tool for dynamic analysis of the core pluripotency gene network. Just how is a set of four transcription factors able to open up the chromatin structure of a somatic cell and set up a highly complex and interconnected transcriptional network, characteristic of ESCs, in a somatic cell has become the focus of leading scientific research.

2.3.1 Reprogramming factors and small molecules

Following the derivation of mouse iPSCs in 2006, several labs published reports of cellular reprogramming. Two different types of transgene combinations emerged initially, the traditional Oct4, Sox2, Klf4 and c-Myc combination, as well as, another combination where the last two genes had been replaced by Nanog and Lin-28 [50-55]. The observation that two different sets of genes could be used for reprogramming identical starting population of cells to seemingly identical ESC-like phenotype suggested the existence of multiple trajectories between the two states in the sense of Waddington's epigenetic landscape [56]. The two common factors here, Oct4 and Sox2, were already very well known transcription factors known to be crucial for maintenance of pluripotency. These two genes, along with another transcription factor called Nanog, are believed to be at the core of the regulatory circuit governing ESC pluripotency and have been shown to be expressed in the early developing embryo [57-60]. Oct4, also known as Pou5f1, is a transcription factor belonging to the POU family and was first identified in the P19 mouse embryonal carcinoma cells [61]. Oct4 was identified early on as a master regulator for maintenance of ICM in the developing embryo, as well as, the in vitro culture of mouse ESCs. During murine development, Oct4 null embryos do not survive beyond the formation of the blastocyst. Nichols and colleagues showed that the these cells differentiate to the extraembryonic trophectoderm lineage and are incapable of producing any cell types that a normal Oct4-expressing ICM would normally give rise to [62]. The Oct4 expression also needs to be tightly regulated in ESCs to maintain their pluripotent phenotype. Niwa et al. used conditional expression and repression in mouse ES cells to show that a 50% increase or decrease in the Oct4 levels in mouse ESCs leads to differentiation of cells to primitive endoderm/ mesoderm and trophectoderm lineages, respectively [63]. Chromatin immunoprecipitation analysis to determine Oct4 binding sites in pluripotent cells have shown that the transcription factor has numerous binding sites across the genome, many of which are co-occupied by other important transcription factors such as Sox2 and Nanog [64]. More recently, two studies used improved affinity purification technique to determine the Oct4 'interactome' [65, 66] The authors were able to confirm a number of previously identified interactions of Oct4 with pluripotency-associated genes such as Sall4, Esrrb and Dax1.

Sox2 is a member of the Sox gene family and, like Oct4, is expressed in the ICM, epiblast and germ cells of the embryo. Beyond the epiblast stage, unlike Oct4, the expression persist in the presumptive neuroectoderm and the homozygous mutant embryos only give rise to trophoblast giant cells and extraembryonic endoderm [67]. In their seminal study, Guo et al. profiled single cells of the growing mouse embryo to understand lineage segregation and the role played by transcription factors such as Oct4, Sox2, trophectoderm marker Cdx2, primitive endoderm marker Gata6 etc. in it [68]. The authors observed a strong upregulation of Sox2 in the inner cells of the 16 cell embryo. At this stage, no change can be observed in the levels of Oct4, Nanog and Klf4 from 8-cell to the 10-cell stage. The timing of the increase in Sox2 also correlated with cell fate. The cells where Sox2 increased early predominantly formed the epiblast, presumably by upregulating Fgf4, while the ones where Sox2 increase occurred later, predominantly gave rise to the primitive endoderm. These findings underscore the importance of Sox2 in early embryonic development and lineage segregation. Sox2 is a key component of the pluripotency circuitry in stem cells. Many important genes in this circuit, such as Nanog, Lefty1 and Fgf4, contain enhancer elements having Oct4 and Sox2 binding motifs [69-71]. However, Sox2 was shown to be dispensable for maintaining mouse ESCs by forced expression of Oct4 [72]. The authors showed that while Sox2-null mouse ES cells differentiate to trophectoderm-like cells, the activity of the Oct-Sox enhancers was maintained, perhaps due to redundancy from Sox15. They further confirmed that the differentiation of Sox2 null cells was mediated through reduction in Oct4 expression and the forced expression of Oct4 transgene can alleviate this effect.

Klf4 belongs to the family of Kruppel-like factors, along with other members such as Klf2 and Klf5. As mentioned above, these factors are expressed in early embryo and are important in transcription regulation and lineage segregation during development. As an example, Klf5 null embryos arrest at blastocyst stage and the genes has been shown to be critical for formation of trophectoderm and epiblast, while suppressing the primitive endoderm lineage [73]. Klf4 has been shown to be involved in the regulation of Oct4, Nanog and other pluripotency-associated genes [70, 74] The different Klf genes exhibit a certain level of redundancy in function. Jiang et al. showed that Klf2, Klf4 and Klf5 have redundant binding to different loci and thus, one can compensate for the reduced expression of the other [75]. The authors found that mouse ES cells differentiated only when all the three genes were knocked down simultaneously using RNAi. With the derivation of epiblast stem cells, the importance of Klf4 was further emphasized when it was found that constitutive expression of Klf4 could allow epiblast stem cells to revert to the ground state of pluripotency, i.e., a state resembling ICM in vivo (section 2.3.4 above)[76].

The role of c-Myc in reprogramming immediately came under investigation owing to it being a proto-oncogene. The use of c-Myc or other cancer-associated genes in cellular reprogramming was an immediate concern for their use in therapy. An investigation into the ability of family proteins of the 4 factors (OSKM) led Nakagawa et al. to discover that mouse iPSCs could be generated without using the c-Myc factor. The authors reported that upon using the 3 remaining factors, while the overall number of colonies were less, they were able to get higher quality iPSCs with minimal non-iPSC background [77]. Similar investigations by other groups showed that 3 factor reprogramming is substantially delayed in comparison to the 4 factor one and also exhibits a much lower overall efficiency of reprogramming [78] but the quality of iPSCs, assessed by endogenous gene activation, chimera formation and germline transmission, was not compromised. This suggests that c-Myc promotes reprogramming by increasing cell proliferation, especially during early stages of reprogramming. Nakagawa, Yamanaka and colleagues replaced c-Myc with L-Myc and N-Myc in their reprogramming experiments and found that L-Myc, that lacks the transformation ability that c-Myc has, is able to efficiently give rise to iPSCs [79]. The authors concluded that c-Myc promotes reprogramming in two different ways- by suppressing expression of genes in fibroblasts that are not expressed by ESCs, and through its transformation capability where it activates proliferative genes expressed in cancer cells etc.

2.3.2 Other reprogramming factors

Following the first reports of derivation of iPSCs using the four factors, investigations into finding alternative factors that could replace any of the four factors or enhance the efficiency of reprogramming ensued. One of the first genes identified this way was the ophan nuclear receptor Esrrb [80]. The authors were able to use Esrrb as a replacement for Klf4 and generated iPSCs from mouse embryonic fibroblasts using Esrrb along with Oct4 and Sox2. To show that Esrrb can replace Klf factors in the genetic regulatory network of ESCs, they showed that Esrrb can sustain self-renewal following triple knockdown of Klf2, Klf4 and Klf5, which had been shown to act redundantly [75]. Heng et al. screened 19 nuclear receptors for their ability to enhance reprogramming efficiency and found that orphan receptors Nr5a2 (also called Lrh-1) and Nr1i2 can enhance the reprogramming efficiency when added to the OSKM cocktail [81]. Nr5a2, which had previously been shown to regulate Oct4 expression by binding to its proximal promoter and proximal enhancer regions at the epiblast stage of embryonic development [37], was even able to replace Oct4 from the traditional cocktail. Nr5a1 and Nr5a2 have recently also been identified, through a genomic screening approach, to reset EpiSCs to the ground or naïve state of pluripotency [36].

In another screen conducted for identifying enhancers for reprogramming, Zhao et al. found that addition of gene Utf1 and the siRNA for tumor suppressor p53 to the traditional four factor cocktail (OSKM) increased the reprogramming efficiency by almost a 100 fold [82]. While, UTF1 was already known to be a downstream target of the Oct4-Sox2 complex in ESCs [83], the role of p53 was a new finding although not a surprising one. The tumor suppressor pathway involving p53 (Trp53 in mouse and TP53 in human cells) and its downstream gene p21 can be activated in cells in response to increase in expression of oncogenic factors (like c-Myc and Klf4) and DNA damage. The role of p53 was further elucidated by a string of reports that showed that the p53-p21 pathway acts as a safeguard in iPSC reprogramming and prevents transformation of cells, similar to its role in tumorigenesis [84-88]. The elimination of this roadblock increases reprogramming efficiency, possibly through preventing transformation of suboptimal cells, but this should obviously be used sparingly to maintain quality of reprogrammed cells.

P53 and c-Myc also influence cellular reprogramming through micro RNAs. Micro RNA (miRNA) are small, 22 nucleotide, noncoding RNAs that cause translational repression by RNA silencing or degradation. Mouse embryonic fibroblast-specific miR-21 and miR-29a, which are suppressed by the reprogramming factor c-Myc, have a negative effect on reprogramming and this can be alleviated by suppressing their expression [89]. It was also shown that inhibition of these miRNAs led to a decrease in expression of p53. Another miRNA family that has a negative effect on reprogramming is miR-34, which was shown to be a downstream target of p53 (along with p21) in pluripotent cells. While p21 influences reprogramming mostly by affecting cell proliferation, miR-34 exerts its affect, at least in part, by repression of pluripotency genes Nanog, Sox2 and N-Myc [90]. Some micro RNA families have also been implicated positively in reprogramming. Members of the miR-290 cluster, miR-291-3p, miR-294 and miR-295, that have been shown to be regulated by c-Myc, increase the efficiency of 3 factor reprogramming [91]. While this report used the miRNAs along with the induction of Oct4, Sox2 and Klf4, Anokye-Danso et al. were able to reprogram mouse and human somatic cells using the miR302/367 cluster without any exogenous factors [92]. The cluster has been shown to be a direct target of Oct4 and Sox2 in human ESCs [93].

2.3.3 Small molecules in reprogramming

The use of iPS cells for regenerative therapy would ideally require reprogramming without extensive genetic manipulation to minimize future risks of tumor formation and uncontrolled cell growth. Due to this concern, there has been a significant interest in using synthetic chemistry to develop small molecules that can be used to control cell fate and replace traditional genetic reprogramming approaches. Small molecules also offer other advantages such as faster and reversible temporal control of expression. Even before the derivation of iPSCs, small molecules had been discovered that supported self-renewal of ESCs and helped in better understanding the signaling involved in their self-renewal and differentiation. As mentioned earlier, the discovery of 2i medium containing small molecules that served as antagonists of ERK signaling and GSK3β, allowed the field to move towards more defined culture conditions and also strengthened the concept of ground state of self-renewal in absence of differentiation signals [25]. Using a phenotypic screen, Chen et al. discovered a small molecule called Pluripotin (SC1) that was able to maintain self-renewal of mESCs in absence of feeders and cytokines [94]. The molecule was shown to inhibition of ERK1 and RasGAP signaling.

In cellular reprogramming, small molecules have been used to either enhance the reprogramming efficiency or to replace one of the reprogramming factors. An example of the former approach is Valproic acid (VPA), a histone deacetylase (HDAC) inhibitor, which increased the efficiency of traditional 4 factor reprogramming of MEFs by more than a 100 fold [95]. The efficiency was also significantly improved (about 50 fold) in 3 factor reprogramming, without c-Myc. The effect of such a chromatin modifying molecule could be large scale upregulation/downregulation of genes that creates a favorable epigenetic environment for reprogramming. Ichida et al. found a small molecule RepSox (E-616452) in a screen to look for small molecules that can replace Sox2 in the 4 factor reprogramming cocktail. This molecule replaces Sox2 by acting as an inhibitor for TGF-β signaling and activating Nanog.

Another small molecule that has been shown to be able to replace Sox2 in the reprogramming cocktail is BIX-01294, a G9a histone methyltransferase inhibitor [96, 97]. This small molecule was even shown to be able to replace Oct4, albeit at a lower frequency of iPSC generation, in reprogramming neural progenitor cells. Although this was an interesting finding, it is important to note here that NPCs express Sox2 endogenously, although at a level lower than what is seen in pluripotent ESCs or iPSCs.

2.3.4 Milestones in reprogramming: endogenous gene expression and epigenetics

Further investigations into the sequence of events during reprogramming showed that the process follows a rather defined path where some markers such as alkaline phosphatase (AP) and SSEA-1 start getting expressed early on (early events) , whereas transgene silencing and X chromosome reactivation are late events. Using a Doxycycline-inducible lentiviral system, Stadtfeld et al. showed that in mouse IPS reprogramming, the expression of transgenes is required till day 10 following which, the cells are in a self-sustaining, poised state for reprogramming [98]. Also, a much higher percentage of cells (3-4 %) are seen to become AP or SSEA-1 positive very early (day 3 to 4) on but only a much smaller fraction of these cells are able to complete the regimen and become completely reprogrammed [99] which also highlights the stochastic nature of the process (next section).

2.3.5 Mechanism of transgene induced reprogramming

Over the last few years, since the initial derivation, iPSCs have been derived reproducibly from multiple starting cell types. Even though the protocol has been improving and has become quite reproducible now, the efficiency of reprogramming is still quite low, around 0.01%-0.1%. Two models have been proposed for the process of reprogramming-elite deterministic versus stochastic model [100]. The term 'latency' is used to define the number of cellular events, such as cell divisions, that occur before a cell gets reprogrammed to iPSC state. The elite deterministic model suggests that only a small subset of cells in the starting population have the ability to get reprogrammed and these cells have fixed latency. This propensity to participating in the reprogramming by these 'elite' cells could be innate (predetermined elite) or imparted through the viral transfection (induced elite). The predetermined elite cells would be similar to more immature or 'stem-like' cells that have been shown to be present, albeit in a very small percentage, in adult tissues and organs. The elite model has not been widely accepted though due to several experimental observations that go against the 'elite' hypothesis. Improvements in reprogramming efficiency through the use of small molecules such as valproic acid [101] suggest an increase in the starting number of cells that undergo reprogramming over increasing the reprogramming rate of the preexisting 'elite' cells. More importantly, Hanna et al. successfully reprogrammed terminally differentiated B cells [102]. The authors used this system because the maturation of B cells can be tracked through the genetic rearrangements of light and heavy chain immunoglobulin loci. The 'induced elite' model proposes that the integration of viruses in the host genome primes some cells to reprogram through activation or inactivation of genomic loci. The evidence that strongly goes against this model is that no specific pattern of integration or specific integration sites have been determined [103]. Also, iPS cells have been derived using integration-free techniques too, albeit with lower efficiency [104].

Stochastic model has found the most approval among all these model of reprogramming. It proposes that all the cells in the starting population have the ability to get reprogrammed but a series of stochastic events have to follow to allow successful and complete reprogramming. The latency here is not fixed and varies from one cell to another. In an elegant study using B cells from Nanog-GFP reporter knock in mice with Doxycycline-inducible expression of four factors (OSKM), Hanna et al. demonstrated that reprogramming is, quite likely, a continuous stochastic process. The authors suggest that the transition from a somatic cell to a reprogrammed one are caused by 'drifts in cell state', which could occur due to the inherent stochastic nature of gene expression in cells [105].

2.3.5 How similar are ESCs and iPSCs?

While, ESCs and induced pluripotent stem cells have many similarities, as discussed above, it is important to understand if there are any differences between the two cell types and how these differences might affect their uses.

In most studies involving iPSCs, it is seen that the gene expression signature of these cells is highly similar to that of their embryonic counterparts, having become increasingly dissimilar to the expression signature of their parent somatic cells as the reprogramming progresses. But as the iPSC field grew, the focus on whatever small gene expression differences did exist between the two pluripotent cell types increased. Chin et al. carried out an extensive analysis to compare the two pluripotent cell types using microarray analyses, array CGH, miRNA profiling and histone modification profiling [106]. The authors compared gene expression of early and late hIPSCs from multiple experiments performed across multiple labs with hESCs and found that there is a distinct hIPSC signature which is not stochastic. An important observation made was that despite the disparate starting cell types in the experiments compared, there was a considerable extent of overlap in the genes that were differentially expressed between the iPSCs and ESCs, which seems to suggest that these differences are characteristic of the iPSC state in general and not due to random, stochastic events. Very Small but definite differences were also observed in the miRNA expression signature and the histone modification at promoter regions in the two cell types, leading the authors to believe that the iPSC state is a distinct state of pluripotency and is not identical to their embryonic counterparts. This is not completely surprising because it is understood that reprogramming is a continuing process and the iPSCs keep moving closer to the ESC state with increasing passages. Guenther et al. carried out a similar analysis comparing the gene expression between ESCs and iPSCs from early and late passages obtained from different experiments conducted by multiple labs [107]. However, the conclusion of this study was opposite of what Chin et al. concluded in ether earlier study in that the authors did not see any consistent differential gene expression between iPSCs from different experiment and

ESCs. It was proposed that this discrepancy between the results of the two studies was due to a different, and perhaps less stringent, statistical data analysis and interpretation. As an example, Chin et al. did not use multiple hypothesis testing and exclusively relied on fold changes for their analysis. The authors also did not look the direction of fold change to look at common differentially expressed genes between different samples. The conclusions about the H3K4me3 and H3K27me3 profiles in iPSCs versus ESCs though, matched very well between the two studies. . Guenther et al. also noticed that the gene expression data clustered in a lab-specific manner, which suggests that the differences between ESCs and iPSCs observed could be lab-specific which, in turn, implies that these differences are indeed stochastic. This was also confirmed by a later study which used available gene expression data from multiple studies, including data from Chin et al., to look at trends in differential gene expression [108]. In line with conclusions of Guenther et al., the authors found that the majority of differential expression between iPSCs and ESCs are lab-specific and could be owing to factors such as reprogramming and culture techniques, progenitor cell type and yet undiscovered microenvironment effects. This study again emphasized the stochastic nature of differential gene expression between ESCs and iPSCs. In other aspects of cell homeostasis, such as reactive oxygen species (ROS) defense mechanisms and mitochondrial regulation, the two cell types are also very similar [109]. Human ESCs prevent themselves from ROS damage through limited mitochondrial biogenesis and high expression of antioxidant enzymes [110, 111]. Human iPSCs are able to reduce their mitochondrial genome copy number to levels similar to hESCs and also upregulate expression of hESC-specific, key antioxidant genes such as

GSTA3, NRF1, UCO4 etc. To conclude, gene expression signature of iPSCs is influenced for a number of factors such as the passage number, parent somatic cell type, reprogramming technique etc. However, what has been established without doubt is that with increasing passage number, the two cell types become increasingly similar and the residual donor cell expression reduces.

There is also the question of whether reprogrammed iPSCs have any ‘epigenetic memory’. It has been demonstrated in numerous studies that iPSCs acquire the correct, gene inducing histone modifications at the promoter regions of key pluripotency genes while imposing the suppressive H3K27me3 modification on some differentiated cell type-specific genes. Polo et al. addressed this question in their elegant study by obtaining genetically matched iPSCs from different mouse fibroblasts, hematopoietic and myogenic cells and observing the transcriptional and epigenetic profile of these cells at early passage (4 to 6) [112]. It was found that the cells were indeed transcriptionally and epigenetically distinguishable at this stage, despite having been proved to be bona-fide iPSCs. While, the epigenetic differences have been shown to be more subtle than transcriptional differences in such experiments, the authors did see high levels of activating H3Ac and H3K4me3 modifications at the promoter regions of genes that were characteristic of parental cell type and expressed at high levels in them, such as, CXCR4 and ITGB1 in skeletal muscle precursor-iPSCs. In line with these epigenetic and transcriptional differences between different iPSC lines, a difference in the *in-vitro* differentiation potential was also seen where tail tip fibroblast-derived iPSCs produced significantly fewer erythrocytes and macrophages than the splenic B cell-derived iPSCs.

The authors also noted that these differences get minimized with eventual passaging and it could be attributed to passive, replication-dependent loss of somatic marks in such cells. Other studies investigating this issue have also arrived at similar conclusions. Kim et al. looked at the in-vitro blood-forming potential of iPSCs derived from non-hematopoietic cells such as fibroblasts and neural progenitors and observed poor differentiation capability owing to residual methylation at loci important for hematopoietic fate [113]. The authors also show that this epigenetic memory can be reset by multiple rounds of differentiation and reprogramming or by using chromatin modifying drugs. Similarly, it has been shown that human beta cell derived iPSCs maintain an open chromatin at important beta cell loci and thus, show enhanced differentiation capability into insulin producing cells [114].

2.4 Transcriptional Network in Pluripotent Stem Cells

The discovery of iPS cells lead to a surge of scientific interest in transcriptional regulation of pluripotency. Now that it was clear that acquisition of pluripotency was not a one way street, there was renewed curiosity about the molecular determinants of this phenomenon. But more importantly, this discovery of the reprogramming progress created novel opportunities to probe the ‘stemness’ of pluripotent cells. It was now possible to study the transition of cells as they rolled up the Waddington landscape, not just as they rolled down, as seen in development and cellular differentiation. Many elegant studies prior to iPSC derivation had led to identification of key transcription factors that play important roles in maintenance of ‘stemness’ and self-renewal. Oct4, Sox2, Nanog and Klf4 were already known to be crucial components of the transcriptional machinery [58-60, 62, 70, 71, 115]. In addition to these key genes, other factors such as Zic3, Utf1, Nr5a2, GCNF, Lefty1, had also been identified by various labs that seemed to be tied in some way to the core transcriptional machinery in maintenance of these cells [37, 70, 83, 116-119]. In addition, these studies had also shown that these genes are part of an extremely intricate transcriptional network where change in the expression of one often had a cascade effect on the expression of the other known players. Furthermore, derivation of pluripotent stem cells lead to the understanding that some of these genes are master regulators and had the ability to cause a global change in cellular transcription leading to induction of pluripotency in differentiated, somatic cells. Subsequently, there was an immediate effort to map the binding sites of these master regulators in the genome [64, 120-122]. These studies have corroborated many previously discovered interactions with physical binding data. In the meantime, efforts to identify new genes that play important roles in ESC and iPSC self-renewal continue successfully.

It is argued that despite all this, a complete understanding of the ‘core’ transcriptional network in pluripotent stem cells has not been achieved. While this argument is not invalid, it is important to remember that computational and experimental studies have shown that the transcriptional network in mammalian cells has a scale-free, hierarchical arrangement where a smaller number of highly interconnected, pivotal factors or genes act as master regulators or ‘hubs’. These hubs are connected to a much larger number of genes[123]. In light of this, it is important to focus on finding these master regulators of the pluripotent state and analyze their interactions with other master regulators of the stem cell fate.

2.5 Rodent Multipotent Adult Progenitor Cells (MAPCs)

Adult stem cells can be derived from the bone marrow of rat [2]. These cells, termed rat multipotent adult progenitor cells or rat MAPCs, express key pluripotency markers such as OCT4, SSEA-1 and REX1 and have been shown to differentiate in vitro into functional smooth muscle cells and hepatocyte-like cells, amongst others [1, 124, 125]. These cells have a unique gene expression signature in that while they express some key ESC-associated pluripotency markers such as OCT4, REX1 and SALL4, they do not express other important genes such as NANOG or SOX2 [126]. Also, these cells express many genes characteristic of nascent hypoblast such as GATA6, GATA4 and SOX17. Having been isolated from the bone marrow of an adult organism, this gene expression is quite peculiar and strikingly similar to precursors to the hypoblast (primitive endoderm) in vivo [4]. These cells also bear a striking resemblance to hypoblast precursor-like cells that have been derived in vitro from pre-implantation rat embryo. Similar, but not identical, cells have also been derived from the mouse embryo. These cells, termed XEN cells, express key hypoblast markers such as GATA4, SOX7, and SOX17 but, unlike XEN-P cells or MAPCs, do not express any pluripotency genes [127].

2.5.1 Isolation of rat MAPCs from bone marrow

The isolation of a homogenous population of MAPCs from the highly heterogeneous bone marrow is a long process spanning 16-18 weeks [2]. The cells can be isolated from the tibia and femurs of 4-6 week old rats or hind limbs of fetal/newborn rats. The cells are depleted of hematopoietic progenitors after 4 weeks in culture and plated at 5-10 cells per well density into single wells of a six well plate. Spindle-shaped cells exhibiting typical MAPC-like morphology gradually arise in low density cultures which grow quite rapidly and can be expanded and maintained in MAPC culture conditions (as discussed in Materials and Methods). The exact source of these cells in the adult bone marrow has not been confirmed. There has been no confirmed report of an OCT4 positive cell present in the adult bone marrow. Lengner et al. used a Cre-lox based genetic approach for tissue-specific activation of Oct4 and found no affect of this on the bone marrow [128]. It is possible that the rat MAPCs arise in culture due to induced reprogramming, especially given the long derivation time.

2.5.2 Hypoblast/Primitive Endoderm-Precursor cells *in-vivo*

As mentioned above, in the early mammalian development, the first differentiation event causes formation of the trophectoderm lineage and the ICM, which further differentiates to give rise to the primitive or extraembryonic endoderm and the epiblast. NANOG expressed in the epiblast cells and GATA6 expressed in the primitive endoderm cells, are markers of the two lineages arising out of the ICM and have been shown to play an important role in their specification [129]. Nanog null ICM does not produce epiblast but only gives rise to primitive endoderm cells and Nanog-deficient ESC differentiate to primitive endoderm too [130]. Similarly, forced expression of GATA6 causes ESC to differentiate to primitive endoderm [131]. GATA6 knockout doesn’t lead to primitive endoderm defects though. Instead, members of GRB2 signaling pathway, including FGF4 and FGFR2 do [132, 133]. The ICM prior to its differentiation into the two lineages, E3.5 in mouse gestation, was believed to be a homogenous population. This idea was supported by the identical morphology of every cell of the ICM at this point in development. Two elegant studies published in 2006 showed that this wasn't the case and these seemingly identical cells of the ICM are already biased towards the lineage that they would give rise-the epiblast (Epi) and primitive endoderm (PE) [4, 134]. Chazaud et al. performed an elegant lineage tracking experiment to show that at E3.5 to show that NANOG and GATA6 were expressed in a complementary 'salt and pepper' fashion which showed no consistent spatial pattern across multiple embryos. The authors also confirmed the role of GRB2 signaling in PE specification by showing that all cells in the GRB-2 null ICM take up epiblast identity. Kurimoto et al. employed a single cell cDNA-amplification procedure to look at global gene expression profile of individual cells of the ICM using microarrays and discovered two very distinct gene expression signatures. They identified the two populations as epiblast-precursor and primitive endoderm-precursor cells with the former expressing high levels of NANOG and low levels of GATA6 and the latter showing the opposite trend. Both these reports observed similar levels of OCT4 expression in the two precursor cell types. Guo et al. also conducted single cell analysis and identified a strong inverse correlation in the expression of FGF4 and FGFR2 in the cells of ICM as the earliest indicator of EPI/PE differentiation [68] further supporting the idea that FGF4 produced by the Epi-precursor cells signals through the FGFR2 receptor on PE-precursor cells and activates PE- markers through GRB2 signaling.

All these studies point towards the existence of a primitive endoderm-precursor cell types in the developing embryo which expresses OCT4 but not Nanog and primitive endoderm markers such As GAT6, GATA4 and SOX17. These precursor cells share a great deal of similarity in their gene expression with rat MAPCs. The primitive endoderm does not contribute to the embryo proper in development and supports the embryo by forming extraembryonic tissue such as yolk sac. In morula aggregation srudies, bone marrow-derived rat MAPCs also contribute exclusively to the parietal yolk sac and not the embryo proper (LoNigro et al., Unpublished work).

2.5.3 Primitive endoderm precursor cells *in-vitro*

Cell lines have been derived *in-vitro* which have properties of the three lineages in early mammalian development. While embryonic and epiblast stem cells are representative of the early and late epiblast in-vivo, respectively, trophoblast stem cells are representative of the extra embryonic trophectoderm lineage and differentiate to trophoblast subtypes *in-vitro* [135]. The trophoblast stem cells show dependence on FGF4 for self-renewal like their *in-vivo* counterpart. Primitive endoderm or hypoblast is the other extra embryonic lineage which, in the mouse embryo, is formed at E4.5, right before implantation. The role of this lineage is to provide support and signaling cues to the developing embryo proper. By E5.0, the primitive endoderm differentiates into two subpopulations: visceral and parietal endoderm. Prior to derivation of any bona fide primitive endoderm stem cell line, embryonal carcinoma cell line, P19, was used as model to study differentiation to primitive endoderm, upon induction of differentiation using retinoic acid [136]. Cells that are representative of the primitive endoderm or more specifically, the parietal endoderm have been derived from both mouse and rat blastocysts [137, 138] but without extensive characterization of gene expression or *in-vivo* contribution to development. Interestingly though, both papers reported the presence of two distinct, inter-convertible morphologies in their cultures. Kunath et al. reported the derivation on a primitive endoderm line from E.5 mouse blastocyst, which were termed XEN (extra embryonic endoderm) lines, [127]. The cell lines expressed markers of both parietal endoderm, such as PDGFRα and SPARC, and visceral endoderm, such as FOXA2 and HNF4α. Like the previous reports, XEN cells also synthesized basement membrane components such as laminin and heparin sulfate proteoglycan. No expression of epiblast genes such as OCT4, NANOG or SOX2 was reported in these cells. The authors noted that their cells were very similar, if not identical, to the cells described in the earlier reports by Fowler et al. and Notarianni et al., and exhibited epithelial-mesenchymal transitions in morphology. Both Notarianni et al. and Kunath et al. observed similar morphology and subcellular organization using scanning electron microscopy, such as the presence of microvillus and extended cellular processes or pseudopodia. Chimeras generated with XEN cells showed their extensive contribution to visceral primitive endoderm but not to the embryo proper or the trophectoderm. The cells, however, showed a strong bias to from parietal endoderm and contribute to parietal yolk sacs. The authors ruled out a role of FGF4 in sustenance and derivation of XEN cells but they did see active LIF-STAT3 signaling in culture, suggesting that might play a role. Interestingly, the XEN cells also bear some resemblance to the primitive endoderm-like cells derived by GATA6/GATA4 over expression in mouse ESCs [131]. Expression of these GATA factors is able to differentiate mouse ES cells to the primitive endoderm even in the presence of LIF signaling.

While there had been some evidence of the presence of a primitive endoderm precursor cell type expressing GATA6, as well as, OCT4 *in-vivo* (Section 2.5.2), the primitive endoderm lines described above seemed clearly more differentiated with no OCT4 expression. In 2009, Debeb et al. reported the first isolation of a cell type that showed a nascent hypoblast/primitive endoderm phenotype [3]. These cells, termed XEN-P (extra embryonic endoderm-precursor), were isolated from the E4.5 rat blastocysts on a layer of mouse fibroblasts in medium containing leukemia inhibitory factor or LIF. In addition to OCT4, these cells also expressed the pluripotency marker REX1 (ZFP42) but no Nanog or SOX2. Also, in addition to parietal endoderm markers shown by XEN cells, the XEN-P cells also expressed markers of visceral endoderm such as DAB2 and EOMES.

An interesting observation was that the rat XEN-P cells expressed SSEA-1 in a small population of the cells, like the previously derived rat parietal endoderm line by Notarianni et al. but this expression was absent in the mouse XEN cells suggesting that there might be inherent differences between the rat and mouse extra embryonic cells. Like the previous reports of extra embryonic cells, the XEN-P culture also consisted of dual morphology: a round, refractile versus flat, epithelial morphology. Debeb and colleagues showed that the round, refractile cells are positive for OCT4, alkaline phosphatase (AP) and SSEA-1 and suggest that these are the true precursor cells in their culture. On the other hand, the epithelial cells were SSEA-1$^-$, AP$^-$, SSEA-3$^+$ and express low OCT4. The authors suggest that latter population emerges as a consequence of spontaneous differentiation of the former. The in-vivo contribution of XEN-P cells was also found to be biased towards the parietal endoderm (84%) although their contribution to visceral endoderm (12%) was significantly higher than what was seen for the mouse XEN cells. This again points to the precursor-like nature of the XEN-P cells in comparison to the XEN cells.

Another report looking at the developmental potential of rat XEN-P cells showed that while the *in-vivo* contribution of mouse XEN cells is restricted to parietal yolk sac, XEN-P cells can contribute to the parietal and visceral yolk sac, extra embryonic ectoderm and, surprisingly, also to the trophectoderm lineage [139]. The contribution to trophectoderm was more predominant upon injection in mouse versus rat embryos and was also significantly higher when cells were added at the morula stage instead of the blastocyst stage.

2.6 High throughput gene expression profiling

Microarray technology has revolutionized the field of functional genomics since its discovery less than two decades ago. The technology allowed simultaneous monitoring of expression levels of a large number of genes in a biological sample [140, 141]. It has served as an invaluable tool for discovering functions of genes and their involvement in different biological processes and pathways. Another big advantage is that current platforms require a very small amount of sample and allows identification of novel genes and pathways involved in biological processes. The DNA microarray technology is based on complimentary of A-T and G-C bases in DNA sequences. RNA from samples of interest is extracted and copied to get cRNA or cDNA. This sample is then hybridized onto a chip containing complimentary oligonucleotide probes. The cRNA or cDNA is usually labeled with a fluorescent probe which is excited by a laser following hybridization. The abundance of every fluorescent spot is captured by a camera during scanning and the information from this digital image is used as an indicator of abundance of a particular gene/probe.

2.6.1 Microarray design and processing

A DNA microarray essentially consists of a small, solid substrate, usually a glass slide, onto which ‘probes’ that are complimentary to cDNA are spotted. These probes can be synthesized oligonucleotides, cDNA samples or PCR fragments. The probes can be synthesized directly onto the microarray chip or can be synthesized prior to deposition followed by spotting onto the chip. Microarray designs can also be classified as two channels versus one channel arrays although the former is not widely used anymore. In two channel arrays, cDNA from two samples is labeled with different fluorescent probes, such as Cy3 and Cy5. The relative strength of Cy3 and Cy5 fluorescent signal at a spot is then used as an indicator of the relative abundance of mRNA of the gene/probe associated with that spot.

As the cost of manufacturing DNA microarrays has gone down with the advancement of technology, multiple manufacturers have come up with different chip designs for profiling gene expression. Typically, these designs differ in characteristics such as the total number of probes on the chip, number of probes associated to a single gene, length of probes (25mer, 60mer), *in situ* synthesis versus spotting, etc. It is common to have multiple probes for every gene such that the different probes target different sequences across the length of the gene. The term 'probe set' is used to refer to all the probes corresponding to a particular gene on a chip. These probes can be concentrated at the 3'end of the mRNA (Affymetrix® 3' expression arrays), distributed across the length of the whole transcript (Affymetrix® Whole Transcript arrays) or specifically target exon sequences (Affymetrix® Exon arrays). While, the probes in Affymetrix® arrays are usually 25mers, other manufacturers such as Agilent Technologies® and NimbleGen® have used 60mer probes in their array design. Agilent Technologies® offers the option of dual mode operation with their arrays which allows them to be used for single channel, as well as, two channel analyses. As an example, Figure 3 shows the schematic of an Affymetrix GeneChip® Rat Gene 1.0 ST array which is a whole transcript array.

The GeneChip® whole transcript array has over 700,000 25mer probes approximating about 26 probes per gene. It covers an estimated 27,342 genes and requires a small amount of initial RNA sample, 50-100ng.

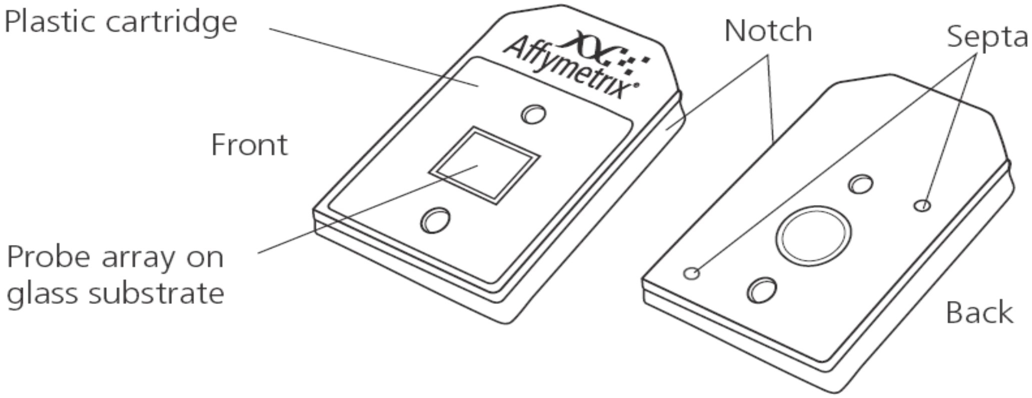


Figure 3. Schematic showing Affymetrix GeneChip® Rat Gene 1.0 ST array (www.affymetrix.com)

2.6.2 Microarray data analysis

An image file is generated when the hybridized microarray chip is processed by a scanner. This image file is then processed further to generate an intensity file that consists of intensity values corresponding to every probe on the chip. Before this raw intensity file can be used to arrive at biologically relevant conclusions, the high throughput data needs to be processed, corrected and summarized. A number of free and licensed computational tools have been developed to this end but the underlying statistical concepts are the same. The first step typically involves detection of background noise which needs to be subtracted from the actual signal. Almost all microarray platforms are designed to include certain detection or control probes to allow detection of this background signal. Following this, the next step of normalization is critical for obtaining data that is free of any bias or errors. Microarray experiments, by their very nature, are prone to a number of systematic or random errors. Systematic errors could result from variations in *in-vitro* transcription or DNA labeling, among others. Random errors could easily arise from fluctuation and limitations in scanning devices. Fortunately, statistical tools can be used to detect and remove these discrepancies from the data before doing any differential expression analysis. A simple linear normalization adjusts for the fact that for equal amount of starting RNA, the overall intensity from every chip in an experiment should be the same. This is a mere scale-up or scale-down approach though and doesn't correct for non-linear variations in the raw signal. Such variations can quite often be corrected by using quantile normalization that works on the assumption that between any two biological samples, the number of probes undergoing significant differential expression is a much smaller number than the total number of probes on the chip. This entails that the intensity distribution plots for different chips should overlap.

As mentioned earlier, in almost all microarray platforms, every gene or probe set is represented by multiple probes that could hybridize to different parts of the target gene; the next step in microarray data processing is that of summarization. Here, signal from multiple probes representing the same gene/probe set is combined to arrive at a single value that can then be used for differential expression analysis.

Following these three main processing steps, the microarray data is now ready for estimating fold changes to determine which genes are different or same between the biological samples of interest. While fold change is a relevant and informative number, more rigorous statistical testing should also be done to determine false positives and statistical significance of the fold change. Parametric or non-parametric tests are usually performed to determine a test-statistic. For example, a t-statistic can be calculated by performing a Student's T-test. The significance of this t-statistic is then evaluated by calculating a p-value based on Student's T-distribution. If the p-value is less than a pre-specified value, such as the commonly used value of 0.05, the null hypothesis is rejected and the gene is determined to be statistically different between the samples. Non-parametric tests, such as Wilcoxon ran sum test, are used when the microarray data does not follow a normal distribution [142].

Since microarray datasets are big, consisting of tens of thousands of probes, the concept of multiple hypothesis testing becomes important. When the null hypothesis, which assumes that a gene is not differentially expressed between two samples, is tested multiple times, type I and type II errors can occur just by chance. A type I error is when a gene is falsely identified as differentially expressed and type II error occurs when a differentially expressed gene fails to be identified as such. Such errors associated to multiple hypothesis testing can be dealt with using statistical criteria such as False discovery rate or FDR [143]. Significance analysis of microarrays (SAM), a popular computational tool for differential expression analysis, False discovery rate determined by SAM gives an estimate of the number of differentially expressed genes, identified as such through a p-value cutoff, by chance [144]. Thus, adjusting the false discovery rate to desired cutoff limits the number of falsely identified genes and permits higher statistical confidence in the fold changes identified [145].

2.6.3 Functional analyses

The goal of a microarray study is usually not just the identification of a list of differentially expressed genes, but to identify a biological context or some sort of functional enrichment. A variety of tools have been developed to derive physiological relevance from microarray data analysis. These tools can be aimed at identifying enrichment of known functions and pathways in data, as well as, discovering new networks and novel regulations.

Licensed software such as Ingenuity Pathway Analysis (IPA®) or Metcore™ are popular tools that can map microarray data onto known maps of pathways and interactions to identify functional enrichment. Both these software rely upon their databases which contain large amount of interaction data curated from public databases such as GO, literature etc. For a list of differentially expressed genes, IPA can overlay the data onto known canonical pathways to detect enrichment and can also generate de novo networks that are based on non-directional interactions in the ingenuity database. A statistical p-value is also calculated to estimate the probability of generation of a network by chance.

Gene set enrichment analysis or GSEA is another popular and free tool that allows detection of functional enrichment in pre-defined gene sets [146]. These gene sets can be a combination of known pathways and user defined sets. The highlight of GSEA is that it determines functional enrichment on the basis of a small but consistent upregulation or downregulation across a majority of genes involved in a particular gene set, as opposed to focusing only on the top differentially expressed genes in the entire dataset. The idea being that there is statistical significance associated to a function or pathway when a sufficiently large number of genes in that function or pathway change simultaneously even if the change in individual expression levels is not very large. The program calculates an enrichment score for every gene set and determines the significance level of that enrichment score. It also accounts for multiple hypothesis testing by normalizing the enrichment score.

2.6.4 Pattern detection and clustering

While determining functional enrichment that largely relies on known interactions is extremely useful tool, microarray data can also be used for discovering new interactions, clusters and networks. Statistical tools for data clustering and pattern detection can help identify novel trends in the data.

Clustering techniques, such as hierarchical or K-means clustering, allow grouping of samples and genes, on the basis of similarity of expression data.

Matrix factorization techniques such as Principal component analysis (PCA) and Non-negative matrix factorization (NMF) are popular tools for dimensionality reduction in microarray data. PCA reduces the complexity of data by transforming it to a new set of variables, called the principal components, which together summarize the variation in the data [147]. The first few principal components capture most of the variation in the data while the last ones mostly capture noise. While PCA is a good tool for visualization post dimensionality reduction, it does not always allow for interpretation of the principal components and clusters [148]. NMF, on the other hand, works better for non negative data, such as microarray data, because the 'metagenes' that are defined by the algorithm are interpretable and often biologically meaningful [149]. The seminal study by Lee and Seung [150] showed that NMF was able to learn the parts of an object (the image of a face) quite well while PCA is only able to learn 'holistic' faces as eigenvalue.

Both PCA and NMF factorize the intital data matrix A, into matrices B and C such that

$$A_{jk} \approx \left(BC\right)_{jk} = \sum_{a=1}^{r} B_{ja} C_{ak}$$

Here, r refers to the rank of reduction. In PCA, entries in B and C can be of arbitrary signs, thus producing eigen values that lack intuitive meaning. In NMF, on the other hand, negative entries are not allowed and this allows identification of 'parts' that are all additive in nature with respect to the initial matrix. This makes NMF an effective tool for class discovery and for determining new patterns in the data. Brunet et al. showed how NMF can be used to predict biologically relevant classes in the data and has the capability to pick out nesting in the data structure [151]. The authors used NMF to recover patterns in cancer-relate microarray datasets and implemented the concepts of consensus clustering to obtain meaningful partitioning of their data and showed that NMF performed better than self-organizing maps (SOMs) for pattern detection and class discovery.

2.6.5 Reverse engineering of cellular networks from gene expression data using ARACNe

Reverse engineering refers to the identification of underlying cellular network by using experimental data such as high throughput gene expression data. The tools for functional analysis discussed in the previous section only allow the user to see if known networks and functions are enriched in an experimental dataset. IPA and other such tools, with their extensive database of interactions, can help discover some new interactions in the data; completely new interactions can't be predicted. Reverse engineering algorithms are directed towards the exact same issue of predicting novel networks using data at hand. One limitation of this tool though is that a large amount of experimental data is needed to do this. Algorithms for clustering are able to pick out genes which are similarly co-regulated but such techniques have not been very successful for large, genome-scale reverse engineering because these algorithms assume a hierarchical nature of gene networks which isn't necessarily true.

Basso et al. developed an algorithm called ARACNe (algorithm for reconstruction of accurate cellular networks) for reverse engineering and showed its success in building hierarchical, scale-free networks in gene expression data from human B cells [123]. They validated their approach by using c-Myc as an example of a regulatory hub and found that ARACNe could successfully discover interactions that have been pre-validated in experiments. ARACNe works though two important concepts of mutual information (MI) and Data processing inequality (DPI) the details of which are beyond the scope of this thesis. Mutual information refers to the extent of relatedness of genes determined using statistically significant co-regulation. The concept of Data processing inequality is then invoked to remove indirect interactions that could be happening through intermediates [152]. While ARACNe is one of the first reverse engineering tools to have been used successfully for eukaryotic network prediction, its application is limited by the requirement of a large amount of data. In their study with B cells, Basso et al. used more than 300 microarray samples for reverse engineering.

CHAPTER 3
Materials and Methods

3.1 Rat MAPC culture

3.1.1 Rat MAPC medium

Basal medium for culturing rat Multipotent adult progenitor cells or MAPCs consists of Dulbecco's Modified Eagle medium low (1g/L) glucose (60%v/v) and MCDB-201 solution (40%v/v). The MCDB-201 solution is prepared by dissolving 17.7 g of MCDB powder in 1000 ml of MilliQ distilled water. The pH of the solution is adjusted to 7.2 before using. The basal medium is further supplemented with 1x Insulin-Transferrin-Selenium (ITS), 1× Linoleic acid–Bovine serum albumin (LA-BSA), Penicillin (100 IU/ml) and Streptomycin (100 μg/ml), 10^{-4} M l-Ascorbic acid (256 mg of l-Ascorbic acid to 100 ml PBS), 2% qualified FBS, Recombinant human platelet-derived growth factor (10ng/ml), Mouse Epidermal growth factor (10ng/ml), Dexamethasone (0.05 μM), Mouse leukemia inhibitory factor (1000U/ml) and β–Mercaptoethanol (55 μM).

3.1.2 Cell Maintenance

Rat MAPCs are cultured on cell culture dishes coated with 100ng/ml mouse fibronectin. The dishes are pre-coated for 1 h at 37 °C. The fibronectin is removed prior to the plating of cells. For routine maintenance, cells are cultured at a density of 300 cells/cm^2. For passaging, cells are detached from the plates by using 0.05% Trypsin-EDTA (Invitrogen).

3.1.3 Infecting rat MAPCs with virus

Rat MAPCs were seeded at 4000-5000 cells/cm^2 in a 24-well tissue culture plate (BD Falcon™). 12-24 hr later, medium was changed to 300 µl fresh MAPC medium containing 5 µg/ml Protamine Sulfate (Sigma-Aldrich). 5-10 µl concentrated virus is added to this and the plate is moved back to the incubator. 12-18 hr later, the medium is removed, cells are washed gently with PBS (to avoid detachment) and fresh MAPC medium (without Protamine Sulfate) is added to allow infected cells to recover and grow. If the lentivirus plasmid expressed a constitutive selection gene, the selection was usually started 24 hr past this medium change to allow enough time for infected cells to start expressing lentiviral transcripts. Alternately, the cells can be passaged (24 hr post medium change) into multiple wells of a 6 well tissue culture dish before starting selection. This is particularly useful for antibiotics where selection is slow, such as Zeocin. In this case, the cells should be replated at a density that allows cells to undergo selection for 10-14 days without becoming confluent at any point during selection.

3.2 Mouse embryonic fibroblast culture

Mouse embryonic fibroblasts were obtained from E13-14 CF-1 mice (Charles River Laboratories, Wilmington, MA). The culture medium consists of high glucose DMEM (Gibco) with 10% (v/v) fetal bovine serum (Hyclone) and 1 mM L-glutamine (Gibco). The cells were mitotically inactivated using gamma irradiation (3000 rads).

3.3 Rat embryonic stem cell culture

3.3.1 Maintenance of rat embryonic stem cells

The rat ESCs are cultured in the rat ESC medium on a layer of CF-1 mouse fibroblasts. Routine passaging is done every 2-3 days by trysinizing cells and splitting in 1:4 ratios. Trypsin is neutralized using serum-based mouse embryonic fibroblast medium since the rat ESC medium in serum free. The freezing medium for rat ESCs is a 90:10 (v/v) mix of mouse embryonic fibroblast medium and DMSO (Sigma).

3.3.2 Rat embryonic stem cell medium

Rat ESC basal medium is a 50/50 (v/v) mix of DMEM/F-12-N2 medium and Neurobasal/B27 medium with 0.1 mM β-Mercaptoethanol (Gibco). DMEM/F-12-N2 medium consists of 25 μg/ml Insulin (Sigma), 100 μg/ml Apo-transferrin (Sigma), 6 ng/ml Progesterone (Sigma), 16 μg/ml Putrescine (Sigma), 30 nM Sodium Selenite (Sigma) and 50 μg/ml Bovine serum albumin (Gibco) in DMEM/F-12 (Gibco). Neurobasal/B27 medium is a 50:1 (v/v) mix of Neurobasal™ medium (Gibco) and B27 supplement (Gibco) with 1-2 mM L-Glutamine (Gibco). The medium with inhibitors consisted of the basal medium with 3 mM CHIR99021 (Stemgent) and 0.5 Mm PD0325901 (Stemgent).

3.4 Lentiviral Packaging

293T cells were plated on a 15cm. tissue culture dish pre-coated with Poly-D-Lysine (P6407, Sigma-Aldrich). The concentration of PDL is 50 μg/ml. and the coated plates are incubated at 37 °C for an hour prior to cell passaging. Medium for culturing 293T cells consists of DMEM 90%(v/v), FBS 10% (v/v), L-Glutamine (2mM). Cells are plated 12-24 hr prior to the start of packaging and should not be more than 70% confluent at the start of packaging protocol. Second generation lentivirus was packaged using one of the two packaging systems: pCMV-dR8.2 dvpr and pCMV-VSVG (Weinberg lab) or psPAX2 and pMD2.G (Trono lab). The protocols for both are as follows:

Using pCMV-dR8.2 dvpr and pCMV-VSVG (Weinberg lab)

In a 15ml. conical, add 10 µg of pCMV-dR8.2 dvpr and pCMV-VSVG each and 40 µg of lentivirus plasmid to 1.5 ml. OptI-MEM™ reduced serum medium (Gibco). Shake and incubate at room temperature for 5 min. In another 15 ml. conical, add 36 µl Lipofectamine™ 2000 (Gibco) to 1.5 ml. OptI-MEM™ reduced serum medium, shake and incubate for 5 min at room temperature. Combine the contents of the two conical, shake and incubate at room temperature for 20 min. Add this to 3 ml. of 293T medium (from above) and add the entire contents onto the 293T tissue culture plate. The medium is changed to 9ml of 5 % serum medium (without antibiotics) 6-8 hr later. The virus can be collected at interval of 24 hr past this point. The virus can be concentrated using Amicon Ultra-15 centrifugal filters (Millipore) and stored at -80 °C in smaller aliquots.

Using psPAX2 and pMD2.G (Trono lab)

The protocol is identical as above except for the initial amounts of packaging and lentivirus plasmid. Here, 5 µg of pMD2.G, 15 µg of psPAX2 and 30 µg of the plasmid are added to 1.5 ml of Opti-MEM ™ medium.

3.5 Retroviral packaging

Retroviral packaging was done by using the Platinum-E (Plat-E) packaging line based on the 293T cell line [153]. The medium for culturing PLAT-E cells consisted of Dulbecco's Modified Eagle medium High glucose (90% v/v), Fetal bovine serum (10% v/v), L-Glutamine (2nM), Gentamicin (20 μg/ml), Blasticidin (10 μg/ml) and Puromycin (1 μg/ml). The cells are plated in a 15 cm tissue culture dish at less than 50 percent confluence. When the cells reach 60-70 percent confluence, plasmid DNA transfection is performed using Lipofectamine 2000. 70 μg plasmid DNA and 90 μl Lipofectamine 2000 are added to 7.5 ml of opti-MEM medium(Gibco®) and the mixture is incubated at room temperature for 20 min. 8 ml of fresh PLAT-E medium is added to this mix and the entire contents are transferred to the PLAT-E tissue culture dish. 6-8 h post transfection, the culture was switched to 18 ml of medium with High glucose DMEM containing low serum (5% v/v) and no antibiotics. First viral harvest is performed at 24 h post the medium switch and a second one is performed another 24 h later. The virus is concentrated by using Amicon Ultra-15 centrifugal filters (Millipore) and spinning the supernatant at 4000 rpm at 9 °C for 25 min. The concentrated virus is stored in smaller aliquots at -80 °C.

3.6 Promoter methylation

3.6.1 Primers for detecting Nanog promoter methylation

MethPrimer (www.urogene.org/methprimer) was used to design methylation primers. The left and right primers span a 213 bps which is 189 bps upstream of the transcription start site.

Left primer: TTTTTTTGTATTATAATGTTTTTGGTGG

Right primer: ATCAACCTATCTAAAAACCAACAACTC

10 μM was used as the primer stock concentration.

3.6.2 Bisulfite conversion of genomic DNA

Conversion was performed on 200-500 ng of genomic DNA by using EZ DNA Methylation-Gold™ kit (Zymo Research), as per the manufacturer's protocol. The converted DNA was eluted in 10 ul. Elution buffer and stored at -20 °C. Sequential PCR was performed by using 1 ul of this converted DNA as the starting amount for amplifying Nanog promoter sequences. PCR was performed in a 50 ul. total volume consisting of ZymoTaq™ DNA polymerase (25 ul), left primer (1 ul) and right primer (1 ul). The first PCR was performed using 1 ul. converted DNA as template while the second one was performed using 5 ul. of the PCR product from the first one as template. The PCR program is as follows: 95 °C for 10 min (1X); 95 °C for 30 s, 50 °C for 40 s, 72 °C for 1 min (35X); 72 °C for 7 min (1X). The second PCR product was gel purified using Wizard® SV gel purification and PCR clean up kit (Promega).

3.6.3 Cloning and sequencing of converted DNA

The gel-purified, PCR product was ligated into pCR®4-TOPO® (Invitrogen) and transformed into One Shot® TOP 10 Chemically competent E. Coli (Invitrogen) as per manufacturer's protocol. The transformation mixture was plated onto Kanamycin resistant plates. Resistant colonies were cultured overnight in LB medium and the plasmid was isolated using Wizard® Plus SV Minipreps DNA Purification Systems

(Promega). The isolated plasmids were sequenced using M13 reverse primer provided by the manufacturer.

3.7 TRIPZ vector construction and packaging

The GATA6 shRNA was moved from the GIPZ vector system to the TRIPZ system. The ~350 bps shRNA containing sequence was excised from the GIPZ vector (V2LHS_42850, Open Biosystems) by using XhoI and MluI, and gel isolated. The TRIPZ vector backbone is similarly digested and ligated with the gata6 shRNA containing sequence. The ligation was confirmed by using the sequencing primer: 5’-GGAAAGAATCAAGGAGG-3’.

The vectors were packaged using second generation plasmids psPAX2 and pMD2.5 (Trono lab)

GATA6 shRNA sequence: GTGTTAAATCATTTGCATA (mature sense), TATGCAAATGATTTAACAC (mature antisense) (V2LHS_42850, Open Biosystems)

SOCS3 shRNA sequence: CACCTGGACTCCTATGAGA (mature sense),

TCTCATAGGAGTCCAGGTG (mature antisense), (V2THS_46878, Open Biosystems)

3.8 RNA isolation and quantitative PCR

RNA extraction was performed by using RNeasy micro kit (Qiagen) as per the manufacturer’s protocol. cDNA synthesis was done by using Superscript ^{III} First strand synthesis (Invitrogen). The first step of cDNA synthesis was performed in a 10 ul. volume with the following components: 2 μg RNA, 1 μl Random hexamers (50 ng/ml), 1 μl DNTP mix (10 mM) and water. The sample is incubated at 65 °C for 5 min followed by holding at 4 °C. The second mixture is formed by adding the following to the first step mix: 2 ml 10× RT buffer, 4 ml MgCl2 (25 mM), 2 ml DTT (0.1 M), 1 ml RNAseOUT (40 IU/ml), and 1 μl Superscript III RT (200 IU/μl). cDNA synthesis reaction is performed on a thermocycler as follows: 10 min at 25°C, 50 min at 50°C, 5 min at 80°C, and return to 4°C for thermal hold. 1 μl of RNAse H was added to the sample and incubated for 20 min at 37°C followed by cooling down to 4°C.

Real time, quantitative polymerase chain reaction was performed using SYBR® Green PCR Master Mix (Applied Biosystems). PCRs are performed in a total volume of 12 μl per well per sample by adding 5.5 μl of cDNA-water mix (2 ul cDNA, 3.5 ul water) to 6.5 ul of SYBR green-primer mix (0.25 ul forward primer, 0.25 ul reverse primer and 6 ul SYBR green) . The PCR program used is as follows: 50 °C for 2 min (1X); 95 °C for 10 min (1X); 95 °C for 00:15 min, 59 °C for 1 min (40X); 95.0 C for 00:15 min (1X); 59.0 C for 00:20 min; 95.0 C for 00:15 min. The RT-qPCR reaction was run on a Realplex mastercycler (Eppendorf) using the above program. Transcript abundance relative to GAPDH was expressed as $\log_2$ (Transcript expression relative to GAPDH) and calculated as ΔCt (dCt) which is Ct (gene of interest)-Ct (GAPDH). Transcript abundance with respect to another reference sample was calculated as ΔΔCt (ddCt) which is ΔCt (reference)- ΔCt (sample). The fold change with respect to reference is calculated as 2^ ΔΔCt.

3.9 Flow cytometry

The RFP/DsRed expressing cells were washed with PBS and trypsinized using 0.05% Trypsin-EDTA. The trypsin is neutralized using PBS containing 3% FBS. The cells are centrifuged and resuspended in PBS containing serum and transferred to FACS tubes for analysis in FACSAria (BD Biosciences).

3.10 cDNA amplification

cDNA amplification was done using the MessageBOOSTER™ cDNA synthesis from cell lysates kit (Epicentre Biotechnologies). Cells were sorted using FACSAria (BD Biosciences). 50-100 cells were added to the 3 µl RNA extraction solution followed by cell lysis by freezing with liquid nitrogen and thawing at room temperature. The first round of cDNA synthesis and *in vitro* transcription were performed as per manufacturer's protocol. Following the transcription, the RNA cleanup was performed using RNeasy MinElute Cleanup Kit (Qiagen). To bring the volume down to the specified amount of 3-8 µl, a SpeedVac concentrator (Thermo Fisher) was used. This was followed by round two of cDNA synthesis. The cDNA dilution depends on the abundance level of the transcript. For rat Nanog, the sample was diluted 1:20 before performing QRT-PCR.

3.11 Genomic DNA isolation and PCR

The PCR was run using a programmable thermo cycler using the following program: 95°C for 2 min; 25-40 cycles at 95°C for 30 s, 55-60°C (annealing temperature) for 45 s, 72°C for 45 s; 72°C for 7 min.

3.12 Alkaline Phosphatase staining

Culture medium was aspirated and cells were washed once with Phosphate buffered saline (PBS). Cells are then fixed by adding 10% Formalin solution (Sigma Aldrich) and incubating at room temperature for 30 min. The fixing solution is then aspirated and PBS is added to the fixed culture. A solution of Napthol, Fast red violet dye and 0.2% Tween-20 solution (in water) is prepared by combining them in the ratio 1:2:3, respectively. PBS is aspirated from the fixed culture and the above solution is added, followed by incubation at room temperature in dark for 25 min. The cells are then rinsed with 0.05 % Tween-20 solution (in PBS). The stained cells can be stored in PBS at 4 °C.

3.13 Immunostaining for Nanog and SSEA-1

The cells are fixed by using 10% Formalin solution (Sigma Aldrich) and incubating at room temperature for 30 min. The fixing solution is aspirated and Phosphate buffered saline containing 0.2% Triton-X is added followed by incubation for 15 min. Non-specific binding is blocked by adding PBS containing 0.1% Tween and 1% Bovine serum albumin and incubating for one hour. The primary antibody for Nanog (Abcam ab80892 Rb polyclonal) was used at 1:250 dilution. The incubation with primary antibody was performed overnight at 4 °C. The primary antibody for SSEA-1 (MAB2155, R&D Systems) was used at a 1:100 dilution. The secondary antibody was used at a 1:1000 dilution and incubation was performed for 1-2 hours.

3.14 Rat microarray hybridization and processing

The rat iPSC, ESC and MAPC samples were hybridized to GeneChip Rat Gene 1.0 ST Array (Affymetrix). Total RNA was isolated from the samples and quality was confirmed using Agilent Bioanalyzer. Poly A RNA controls were added to the total RNA prior to 1st cycle, 1st strand synthesis. 1st cycle cDNA synthesis, in vitro transcription and 2nd cycle cDNA synthesis were performed using Ambion WT expression kit (Ambion). cDNA fragmentation and terminal labeling was performed using WT terminal labeling kit (Affymetrix). Hybridization of sample and scanning of thr array was performed by the Biomedical Genomics Center (BMGC, University of Minnesota). The raw data was normalized using the recommended setting for whole transcript gene arrays by Expression Console software (Affymetrix).

Gene	**Species**	**Forward Sequence**	**Reverse Sequence**
Gata6	Rat	GTCTGGATGGAGCCACAGTT	ATCATCACCACCCGACCTAC
Nanog	Rat	AGCGCCGGTGGAGTATCCCA	ATGGAGCGGAGCAGCGTTCC
Gata4	Rat	CTGTGCCAACTGCCAGACTA	AGATTCTTGGGCTTCCGTTT
Sox2	Rat	CCCGCAGCAAAATGACAGCTGCG	AACTCCCTGCGAAGCGCCTAAC
Pou5f1	Rat	TCAGACTTCGCCTTCTCACCCC	AGGGGCCGCAGCTTACACATG
Pdgfra	Rat	CTTGCAATCCGTCAGTAGCA	ACGTCCTCGGCTAGGGTTAT
Sox17	Rat	GCAAGATGCTAGGCAAATCC	GTACTTGTAGTTGGGATGGTC
Gapdh	Rat	TaqMan® Rodent GAPDH Control	(Applied Biosystems, Cat# 4308313)
Pou5f1	Human	CTTCGCAAGCCCTCATTTC	CCTTGGAAGCTTAGCCAGGT
Sox2	Human	AACCCCAAGATGCACAACTC	CGGGGCCGGTATTTATAATC
Nanog	Human	TTTGGAAGCTGCTGGGGAAG	GATGGGAGGAGGGGAGAGGA

Table 1. List of quantitative real time PCR primers

CHAPTER 4
Gene Expression Signature of Pluripotent Embryonic stem cells

4.1 Introduction

Embryonic and induced pluripotent stem cells have two defining characteristics that separate them from other stem and somatic cell types: indefinite self-renewal and pluripotency. The gene expression signature that imparts these features to a cell begets immense interest from the scientific community. Before the discovery of induced pluripotent stem cells, many studies had been conducted to unravel the gene network responsible for the maintenance of human and mouse ESCs. After their discovery, it became clear that these features don’t have to be inherent in a cell; they can be acquired. And iPSCs introduced new scientific ways to test and ask the questions of self-renewal and pluripotency. While ESCs represented the ball that is rolled down the Waddington’s developmental landscape to lose its defining features, iPSCs represented the opportunity to roll up a ball and observe it arriving at the top (Figure 4) [56]. The genetic and epigenetic features that define a pluripotent stem cell can be acquired by jumpstarting a series of cellular event and allowing conducive culture conditions for the stabilization of eventual, desired cell types. It is important to understand that while Figure 4 shows singular paths to differentiation and reprogramming, in theory, there are infinite paths to both of these events. It is known that multiple differentiation protocols that promote differentiation of ESCs to the same final cell type, follow different trajectories on the genetic landscape. These trajectories depend on the signaling cues that are presented to the cells along these trajectories, as well as, inherent, stochastic genetic events. Similarly, reprogramming of different somatic cell types carves out disparate trajectories on the landscape.

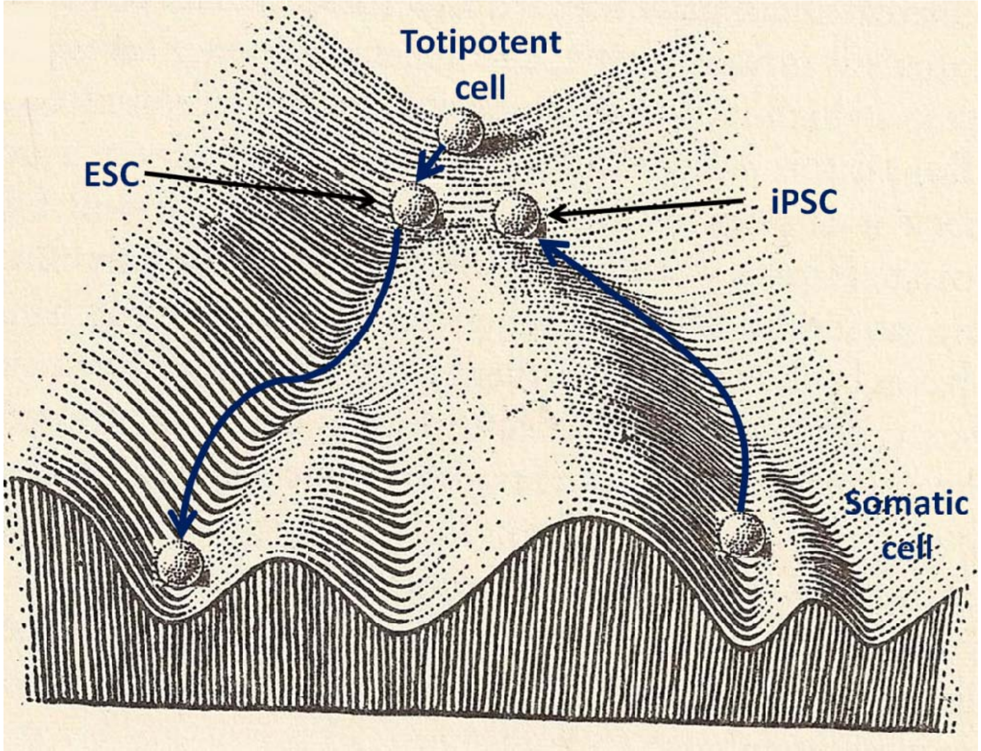


Figure 4. The differentiation of an ESC and the reprogramming of a somatic cell to an induced pluripotent stem cell (iPSC) represented on the original Waddington's developmental landscape [56]

Many studies have been conducted to understand the gene expression changes that take place during both differentiation and reprogramming. While these studies have been very useful and enhancing our understanding of what makes a pluripotent stem cell, the inherent drawback is the small and limited number of trajectories captured in these studies. Thus, to arrive at a core gene expression signature that can help define the state of ‘stemness’ in pluripotent stem cells, a meta-analysis is important. Combining data from multiple studies that capture different trajectories along the genetic landscape, both up and down the slope, as shown in Figure 4 can help arrive at a core set of genes that are truly representative of the pluripotent state.

4.2 Meta-analysis data compilation

The microarray data used in this study was compiled from the public repository for high throughput genomic data, Gene Expression Omnibus (GEO). As discussed in Chapter 2, mouse and human pluripotent stem cells differ in their morphology, culture conditions, as well as, gene expression. While the mouse cells seem to represent the ICM stage of embryonic development, the human counterparts are more similar to the epiblast stage. Thus, in this study, data relevant to both species was collected, processed and analyzed separately. But there obviously are important similarities too between the pluripotent stem cells of the two species. Chapter 5 discusses the common gene sub- networks that are active in the pluripotent cells of the two species.

249 microarray samples were collected from 20 different studies involving mouse embryonic and induced pluripotent stem cells, as shown in Figure 5. All studies used Affymetrix GeneChip Mouse Genome 430 2.0 Arrays for analysis. The raw data was normalized using Affymetrix microarray suite version 5.0 (MAS 5.0) algorithm and the average chip intensity was scaled to 500. The probe sets with detection p-value greater than 0.4 in all the 249 samples were termed as absent. The absent probe sets and those with an intensity value less than 50 (averaged across all the 249 samples) were excluded from further analysis. The redundancy in probe sets was removed by using the highest

intensity probe set for further analysis. This filtering brought down the number of probe sets for further analysis to 18845.

	GEO Accession Number	Cell Types		
		ESCs	iPSCs	Other cell types
1	GSE3653 [154]	2	0	14
2	GSE4189 [59]	5	0	9
3	GSE13235	3	0	6
4	GSE10210 [155]	0	0	16
5	GSE10806 [156]	3	6	2
6	GSE9563 [157]	18	0	18
7	GSE11044 [158]	0	0	0
8	GSE10574 [159]	2	0	0
9	GSE10477 [159]	1	0	2
10	GSE10476 [159]	2	0	2
11	GSE10534 [159]	1	0	2
12	GSE7528 [160]	18	0	0
13	GSE8503 [161]	6	0	0
14	GSE9760 [162]	0	0	12
15	GSE6933 [126]	3	0	12
16	GSE5671 [163]	0	0	18
17	GSE8128 [164]	4	0	5
18	GSE7688 [165]	1	30	4
19	GSE10871 [166]	0	0	2
20	GSE4307 [4]	0	0	20

Figure 5. Studies used in meta-analyses involving mouse embryonic and induced pluripotent stem cells

132 samples were obtained from 13 GEO studies involving human pluripotent stem cells, as shown in Figure 6. All studies were conducted using Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays. Raw data files were processed similar to the mouse samples and resulted in 20949 unique probe sets for further analysis.

		Cell Types		
	GEO Accession Number	**ESCs**	**iPSCs**	**Other cell types**
1	GSE8884 [167]	2	0	2
2	GSE9865 [168]	2	9	2
3	GSE7234	3	0	1
4	GSE9086 [167]	0	0	8
5	GSE6561 [169]	4	0	0
6	GSE9832 [170]	1	8	10
7	GSE9709 [171]	0	8	2
8	GSE9940	3	0	15
9	GSE7896 [172]	8	0	0
10	GSE12390 [173]	3	15	3
11	GSE12034 [174]	0	0	6
12	GSE10809 [175]	4	0	4
13	GSE11350 [176]	3	6	3

Figure 6. Studies used in the meta-analyses for discovering pluripotency-associated gene signature in human ESCs. ESCs: Embryonic stem cells , iPSCs: Induced pluripotent stem cells

4.3 Non-negative Matrix Factorization (NMF) for data clustering

As shown in figures 5 and 6 above, the data collected contained many different types of biological samples. In additions to pluripotent stem cells, multiple cell types that represented different stages of differentiation and reprogramming were present. Thus, to obtain any relevant biological information from the data, it was important to discover natural patterns and classes in the data. Non-negative matrix factorization has been shown to be able to pick out novel trends in gene expression data [151, 177]. For a specified value of the rank r, the algorithm gives the closest approximate factorization of the matrix W (n × m) into matrices H (n × r) and K (r× m). With respect to gene expression data, W is the matrix containing non-negative, logged intensities of n genes in m samples. For the human and mouse datasets, the algorithm is applied for different values of k starting from k equals to 2. The cophonetic coefficient, developed by Brunet et al., is a measure of the stability of factorization and this was used to differentiate between strong versus weak partitioning. Analogous to the facial features identified by Lee and Seung in their study [150], NMF identified 'metagenes' and metasamples' in expression data. These concepts can be understood as follows. The starting matrix W (n x m) consists of expression values of n genes in m samples. For rank r equals to 2, the factorized matrices are H (n x 2) and K (2 x m) such that $W_{nm} \approx H_{n2}K_{2m}$. The two columns in H define two metagenes where $\mathbf{h_{i1}}$ defines the coefficient of gene i in metagene 1 and $\mathbf{h_{i2}}$ defined the coefficient the gene in metagene 2. Every column in K represents the metagene expression pattern in that sample. So, $\mathbf{k_{1j}}$ represents the expression pattern of metagene 1 in sample j. An alternate view of this factorization and one that is more useful for the purpose of this study in the concept of metasamples. Every row in K can be defined as a metasample. So, for r equals to 2, the data has been factorized into two metasamples. Now, $\mathbf{k_{1i}}$ represents the coefficient of sample i in metsample 1 and $\mathbf{k_{2i}}$

represents the coefficient of sample i in metasample 2. In this case, each row of H defines the expression pattern of that gene in the two metasamples.

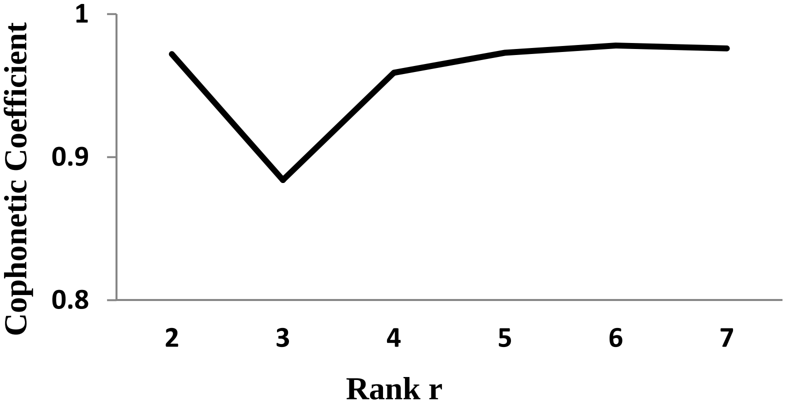


Figure 7. Variation of cophonetic coefficient with rank r in mouse dataset

Using this concept, NMF was performed on both the human and mouse datasets for multiple values of the rank r, starting at 2 and going on till the cophonetic coefficient starts to decline in value, as described in Brunet et al.

Figure 7 shows the change in cophonetic coefficient with increasing values of the rank r in the mouse dataset. The value is high for k equals to 2 but goes down for k equals 3. It reaches its maximum at k equals 6 and steadily goes down from this value (data not shown). Figure 8 shows the ordered consensus matrix which gives a visual indication of the effectiveness of the factorization and shows the poor clustering at k=3. One of the highlights of NMF is that with increasing k, it can pick out finer sub-partitioning in the data and this is what is observed in the mouse dataset. For k=2, the entire data is separated into two classes where pluripotent and slightly differentiated cells were thrown in one group and the other somatic and differentiated cells were in the other group. With increasing k, NMF is able to further segregate the data into sub-classes, except for k=3. At the highest value of cophonetic coefficient at k=6, the entire dataset was factorized into biologically relevant classes as shown in Figure 8. NMF was able to separate out the completely undifferentiated pluripotent stem cells from the partially and terminally differentiated cells.

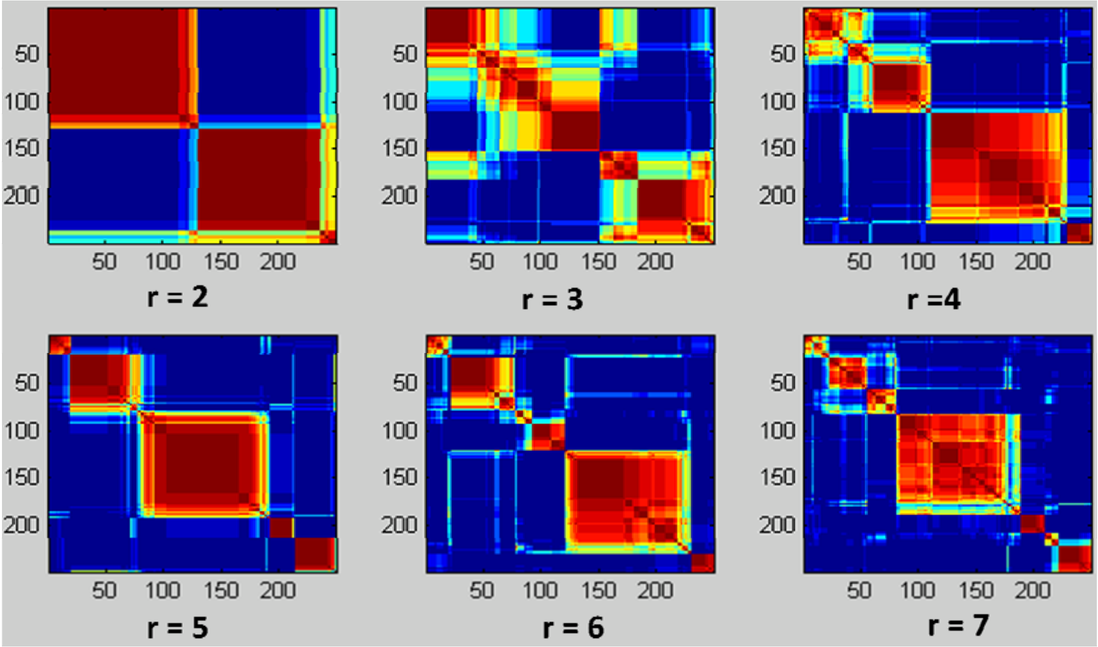


Figure 8. Reordered consensus matrices averaging 100 connectivity matrices computed for rank r from 2 to 7

Similar analysis for human ESCs produced similar trend for cophonetic coefficient, as shown in Figure 10. After k=2, the second highest value of cophonetic coefficient occurred at k=5. Beyond this, the value dropped steadily. A visual representation of this is shown in Figure 11 though the ordered consensus matrix.

Cluster	Number of samples	Class of samples
1	28	Partial iPSCs, Mouse ESCs expressing 4 factors (Oct4, Sox2, Klf4 and c-myc)
2	15	Bone marrow, bone marrow-derived adult stem cells, Mesenchymal stem cells (MSCs), ESC derived cardiomyocytes
3	56	Neurogenin 3 inducible differentiation on day 3, Neural stem cells (NSCs), Day 3.5 endothelial differentiation
4	22	Adult brain, liver, kidney and heart; Hematopoietic stem cell (HSCs); Day 7 and 9 Embryoid bodies
5	108	Undifferentiated ESCs and iPSCs
6	20	ICM

Figure 9. Broad classification of mouse dataset into six biologically relevant clusters by NMF

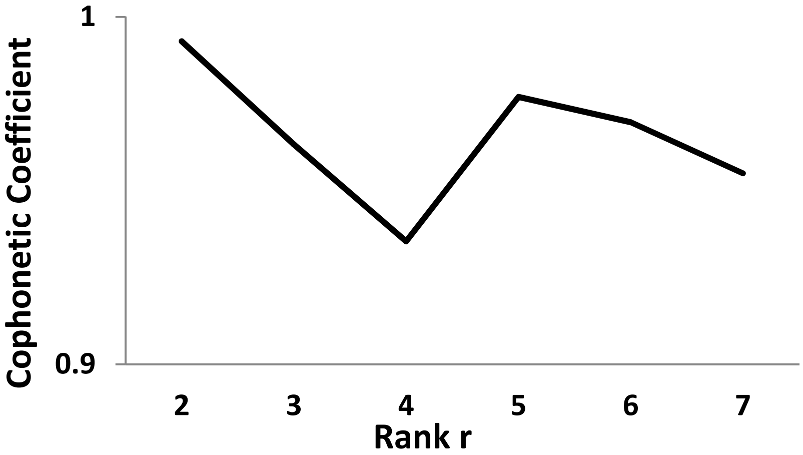


Figure 10. Variation of cophonetic coefficient with rank r in human dataset.

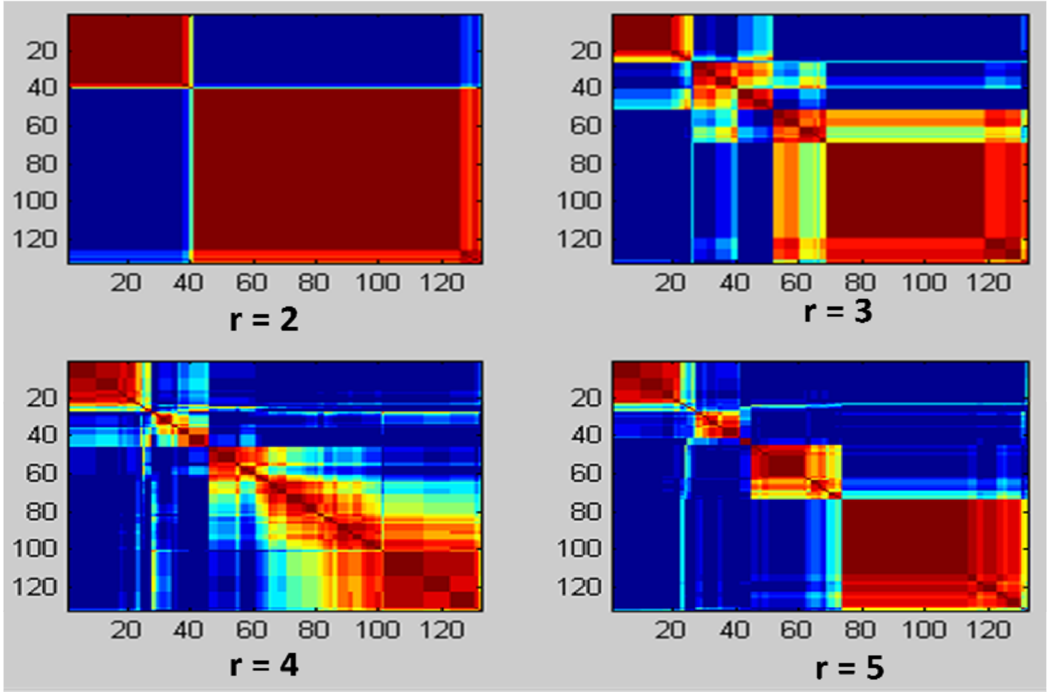


Figure 11. Reordered consensus matrices averaging 100 connectivity matrices computed for rank 2 to 5 in human dataset

Cluster	Number of samples	Class of samples
1	3	Huamn oocytes
2	14	Day 3.5 Embryoid bodies; Blast cells differentiated from ESCs
3	29	Day 6, 10 and 17 Embryoid bodies
4	59	ESCs and iPSCs
5	27	Partially reprogrammed cells

Figure 12. Broad classification of human dataset into five biologically relevant classes by NMF

The broad classification of the human dataset into biologically relevant classes is shown in Figure 12.

4.4 NMF detects Metagenes and Metasamples in expression datasets

As discussed in section 4.3, NMF does not just cluster samples into intuitive biological groups; it also detects metagenes relevant to those metasamples. While a metasample is a group of biologically similar (in terms of gene expression) samples predicted by NMF, every metagene is a group of genes that are responsible for identification of individual metasamples. So, for a metasample that represents pluripotent stem cells, NMF also identified the genes that distinguish that metasample from all the other samples in the analysis. Such a metagene represents the core set of genes that are responsible for the self-renewal and pluripotency of embryonic and induced pluripotent stem cells.

For example, in the mouse dataset, cluster 5 represents pluripotent embryonic and induced pluripotent stem cells. For 6 overall clusters, the gene expression matrix is factorized into two matrices: H (n x 6) and K (6 x 249), where n is the total number of genes. Every row of H represents a metasample with entry $\mathbf{h_{ij}}$ indicating the representation of sample j in metasample i. This can be explained by plotting two rows/metasamples of the H matrix obtained by applying NMF to the mouse dataset.

As shown in Figure 13, the metasample represented by row 1, referred as k1, is representative of cluster 5 and has highest values corresponding to this metasample representing ESCs and iPSCs. Similarly, metasample k5 is representative of cluster 1 representing partially reprogrammed iPSCs. The top 20 genes associated with the metagenes relevant to mouse and human pluripotent stem cells are listed in Table 1 and Table 2 below.

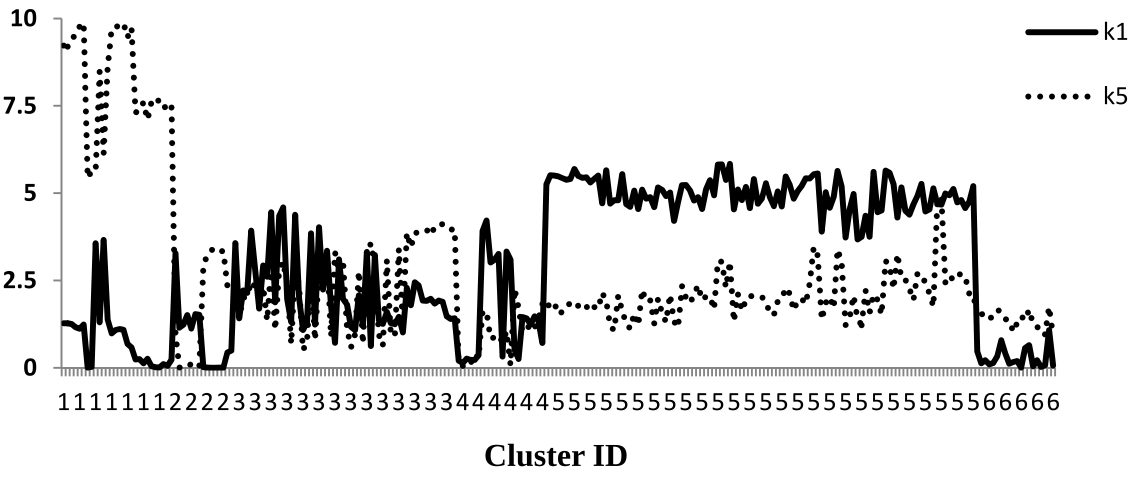


Figure 13. Two rows of matrix K representing metasamples in mouse dataset. k1 represents metasample corresponding to pluripotent stem cells while k5 corresponds to partially reprogrammed induced pluripotent stem cells

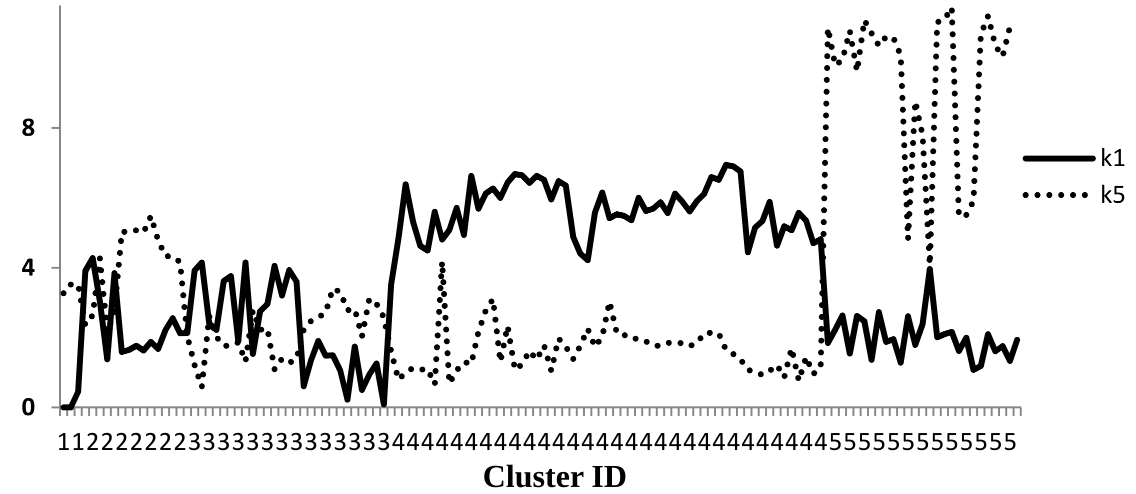
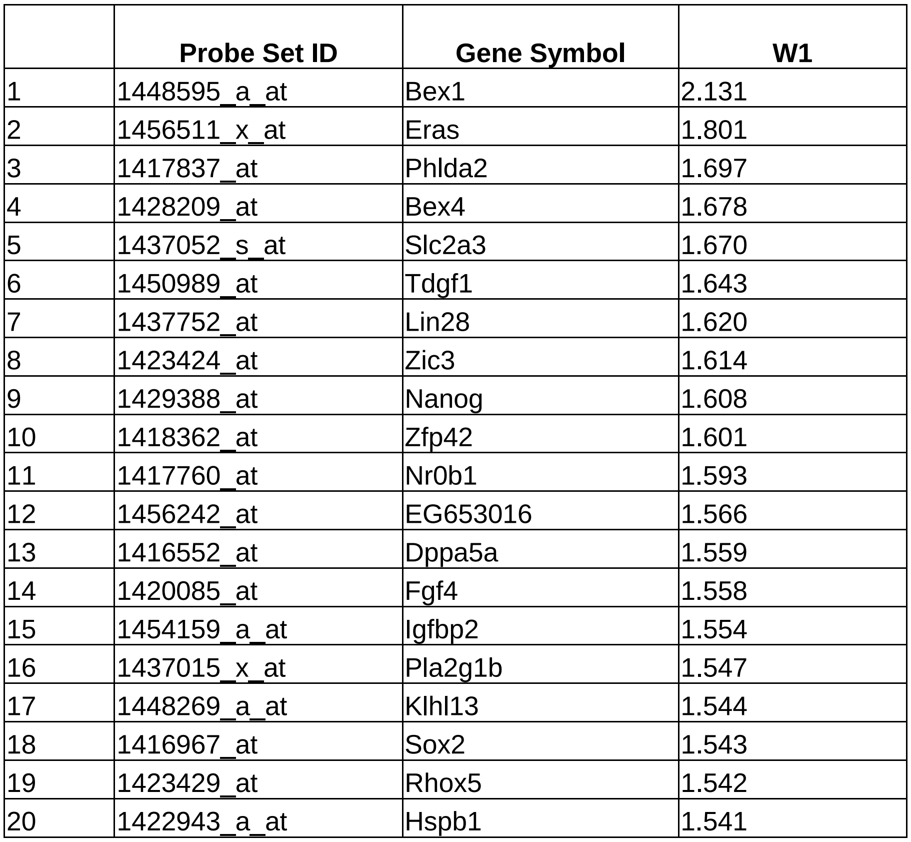


Figure 14. Two rows of matrix K representing metasamples in human dataset, k1 represents metasamples corresponding to pluripotent stem cells while k5 corresponds to partially reprogrammed iPSCs,



	Probe Set ID	Gene Symbol	W1
1	1448595_a_at	Bex1	2.131
2	1456511_x_at	Eras	1.801
3	1417837_at	Phlda2	1.697
4	1428209_at	Bex4	1.678
5	1437052_s_at	Slc2a3	1.670
6	1450989_at	Tdgf1	1.643
7	1437752_at	Lin28	1.620
8	1423424_at	Zic3	1.614
9	1429388_at	Nanog	1.608
10	1418362_at	Zfp42	1.601
11	1417760_at	Nr0b1	1.593
12	1456242_at	EG653016	1.566
13	1416552_at	Dppa5a	1.559
14	1420085_at	Fgf4	1.558
15	1454159_a_at	Igfbp2	1.554
16	1437015_x_at	Pla2g1b	1.547
17	1448269_a_at	Klhl13	1.544
18	1416967_at	Sox2	1.543
19	1423429_at	Rhox5	1.542
20	1422943_a_at	Hspb1	1.541

Table 2. Top 20 genes in the metagene associated with mouse pluripotent stem cells

It is immediately noticeable that important pluripotency associated genes TDGF1, NANOG and DDPA4/DPPA5 are common between the two species. As discussed before, human and mouse ESCs are different in their morphology, as well as, gene expression. Thus, it is not surprising that the exact same genes are not present in the two tables. Comparing the list of top 100 genes in both the species brings out other known factors such as MYCN, ZFP42 (REX1) and LEFTY1.

	PROBE SET ID	GENE SYMBOL	W1
1	231381_at	HESRG	1.727
2	220184_at	NANOG	1.719
3	214240_at	GAL	1.642
4	206286_s_at	TDGF1	1.573
5	214022_s_at	IFITM1	1.548
6	216379_x_at	CD24	1.548
7	214023_x_at	TUBB2B	1.541
8	204285_s_at	PMAIP1	1.523
9	204141_at	TUBB2A	1.502
10	207714_s_at	SERPINH1	1.495
11	222392_x_at	PERP	1.495
12	206268_at	LEFTY1	1.486
13	210905_x_at	POU5F1P4	1.479
14	222154_s_at	LOC26010	1.456
15	202790_at	CLDN7	1.456
16	219955_at	L1TD1	1.454
17	217728_at	S100A6	1.454
18	232985_s_at	DPPA4	1.447
19	231896_s_at	DENR	1.440
20	204409_s_at	EIF1AY	1.431

Table 3. Top 20 genes in the metagenes associated with human pluripotent stem cells

4.5 Predicting novel ‘stemness’ markers from metagene profiles

The identification of known pluripotency associated factors, such as NANOG, TDGF1, and ZFP42, as being common in the gene expression signature that differentiates pluripotent cells from differentiating/differentiated cells is a validation for our approach. But this approach is also an important discovery tool that allows identification of new candidates important for the pluripotency gene expression in both human and mouse embryonic and induced pluripotent stem cells. The gene APOE is such as example. Apolipoprotein E, or APOE, belongs to a group of apolipoproteins that regulate lipoprotein transport through interaction with cell surface receptors [178]. In addition, APOE has been shown to play a role in inhibition growth of many cell types and preventing tumor cell growth and metastasis [179]. In embryonic development, it has been shown to be highly expressed during organogenesis from day 9.5 post coitum in murine development [180] and has been detected in the blastula stage of embryonic development in non-mammalian vertebrates [181]. It has been shown to be upregulated by bone morphogenetic protein- 2 (Bmp-2) in murine mesenchymal progenitor cell line causing differentiation [180]. Though its deficiency has not been shown to be embryonic lethal in mouse development [182], its precise role in human and mouse ESC regulation or maintenance has never been probed.

4.6 Predicting signaling hubs and cellular networks using ARACNe

The metasample pattern identified by NMF, corresponding to the pluripotent stem cells cluster, as discussed in section 4.4, while consistently high in that cluster, shows a variable profile in other clusters (Figure 13). It is due to the fact that the values in row k1 (metasample row corresponding to pluripotent mouse stem cells) represent the expression of metagene representing pluripotent stem cells in specific samples. For example, k_{1j} represents the expression of metagene1 in sample j. The highest values in k1 correspond to the cluster of pluripotent stem cells. If the values are plotted in decreasing order, it can be seen that they are representative of the extent of differentiation or reprogramming of samples. Figure 15 demonstrated this trend by taking example of two studies. Three samples each from study GSE10477 and GSE3653 are shown. GSE10477 involves analysis of conditional knockout of Oct4 in mouse ESCs and the plot below shows that unmanipulated day 0 ESCs have the highest value of k_{1i}, followed by day 1 knockout cells and then day 2 knockout cells. Similarly, for another study GSE3653 which involves differentiation of mouse ESCs, undifferentiated ESCs have high k_{1i} of 5. The value for day 3 sample differentiated by forming embryoid bodies (EBs) is about 3.9. Finally, the day 10 embryoid sample has presumably undergone extensive differentiation and falls to a k_{1i} value of 1.4.

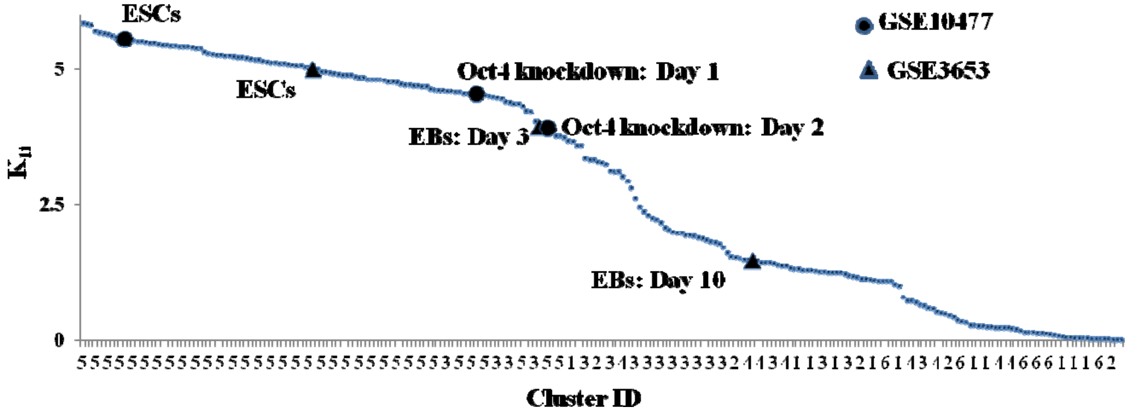


Figure 15. Trend in values of metasample represented by first row of matrix K versus cluster ID in mouse pluripotent stem cells

Similarly, for human dataset, the coefficients in metasample representative of the pluripotent stem cells exhibits similar trend. Figure 16 shows displays the trend in three samples from study GSE9865 involving reprogramming of neo-natal human dermal fibroblasts by induction of four factors, Oct4, Sox2, Klf4 and c-Myc. NMF separates out the samples into three separate clusters. The fully reprogrammed iPSCs have a high value of the coefficient in metasample representative of the pluripotent stem cell cluster. This value goes down successively from these iPSCs to partially reprogrammed iPSCs and then to unmodified fibroblasts.

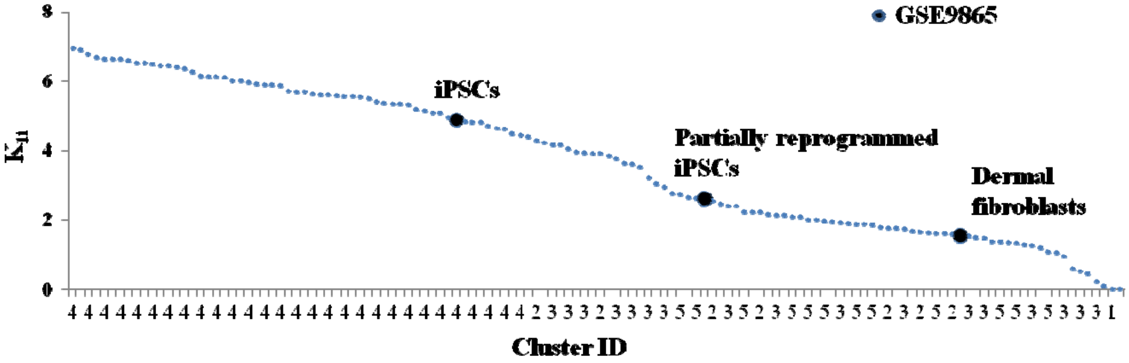


Figure 16. Trend in values of metasample represented by first rwo of matrix K versus cluster number in human pluripotent stem cells

This success in ranking samples according to the metagene pattern representative of pluripotent stem cells allows us to separate data into classes based on their extent of differentiation or reprogramming. For example, in the mouse dataset, the value of k_{1i} can be used to classify cells as fully reprogrammed and undifferentiated (k_{1i}>4.4) cells; early differentiation and late reprogramming samples (4.4> k_{1i} >1.5) and the final class of terminally differentiated and somatic cells (k_{1i} <1.5). This allows a numerical estimate of the extent of differentiation or reprogramming and enables us to conduct differential expression analysis between cells at different stages of these processes.

ARACNe is a tool that can be used to reverse engineer cellular networks using gene expression data [123]. The algorithm identifies co-regulation through gene expression data, removes indirect interactions and builds hierarchical, scale-free networks that contain few, highly interconnected genes called hubs. This is, indeed, representative of biological signaling networks where a few signaling centres are intricately connected to a large number of genes. In contrast, most genes, have rather small number of interactions. ARACNe was used to determine major cellular 'hubs' involved in pluripotent human and mouse stem cells. The genes signature associated with the pluripotency signature identified by NMF (described above) was used as the input to the algorithm.

Figure 17 shows the top 15 network hubs identified in the mouse stem cells and the number of direct interactions of each gene. A surprise candidate was the transcription factor Tcf15 which was identified as the biggest signaling hub. Tcf15, also known as bHLH-EC2, has been shown to be expressed in mouse ESCs and has been suggested to be involved in maintaining ESCs in a non-committed state in absence of any differentiation cues [183]. The predicted first neighbors of Tcf15 include genes Foxd3, Etv4, Fgf17, Zic3, Dmrt1, Eras, Zfp42, Ash2l, Dppa2, Rex2, Tcea3, Nr0b1, and Nodal. If the depth of the network is increased to include, in addition to direct interactions, genes connected to Tcf15 through one intermediary, the number of interactions increases to 518. Prominent genes connected to Tcf15 through a single intermediary are Dppa3, Lefty2, Mycn, Klf4, Jarid2, Klf5, Klf2, Nanog, Dppa4, Gdf3, Rest, Utf1, Fgf4, Dnmt3i, Tcea1, Bmp4, Tdgf1, Klf9, Sox2, Pou5f1, Trp53. Important pluripotency factor, Nanog, and Tcf15 are extensively connected to each other through 16 intermediaries including Zfp42, Dmrt1, Foxd3, Eras and Syce1. Sox2 and Tcf15 share 7 intermediaries between themselves including Foxd3, Zic3, Ash2l and Tdh. Similarly, Oct4 and Tcf15 share 4 intermediaries, namely, Zic3, Tdh, Nr0b1 and Nr3c1.

Nodal is the next biggest signaling hub which also interacts directly with Tcf15. In addition, the two genes share about 40 first neighbors including genes such as Foxd3, Dmrt1 and Eras. Esrrb, an important transcription factor in pluripotent stem cells, was the third largest signaling hub. It was predicted to interact directly with important genes such as Rex1, Klf4, Jarid2, Mybl2, Dppa5, Klf2, Klf5, Dppa4, Nanog, Gdf3, Utf1 and Zfp42.

A similar analysis with the human dataset predicts network hubs as shown in figure 18. Rex1 (Zfp42), a well-known 'stemness' associated gene was the biggest signaling hub. It has direct interactions with Jarid2, Dnmt3b, Pou5f1, Fgf4, Cyp26a1, Dppa4, Mybl2 and Tdgf1. Again, if the network is expanded to include genes connected to Rex1 through an intermediate gene, the number of interactions increases dramatically to 794 and includes almost all pluripotency associated genes such as Nanog, Sox2 , Bmp2, Tp53, Mycn , Utf1, Etv4, Ash2l, Zic3, Nodal and Terf1. Tdgf1 is the next biggest transcription factor hub and direct interactions were predicted with genes like Nanog, Gdf3, Dnmt3b, Pou5f1, Dppa4, Fgfr4, Fgf4, Lefty2, Dppa2, Lefty1,

Seven genes, namely, Pou5f1, Nanog, Sox2, Tdgf1, Mycn, Nodal and Gdf3 are important common signaling hubs in the two species and suggest that these factors are important in maintaining stemness in pluripotent stem cells of both species.

	Gene	Direct Interactions
1	Tcf15	152
2	Nodal	134
3	Esrrb	112
4	Fgf4	76
5	Sohlh2	74
6	Pl12g1b	72
7	Rest	69
8	Nanog	66
9	Utf1	64
10	Dmrt1	56
11	Gdf3	51
12	Tcea3	51
13	Nr0b1	49
14	Klf2	45
15	Tcfcp2l1	45

Figure 17. Top signaling hubs and their number of direct interactions in mouse embryonic and induced pluripotent stem cells, as predicted by ARACNe

	Gene	Direct Interactions
1	Zscan10	157
2	Tdgf1	137
3	Epha1	112
4	Sox2	99
5	Zic3	99
6	Sirt1	99
7	Lck	94
8	Pou5f1	94
9	Foxh1	92
10	Nanog	90
11	Gabrb3	83
12	Mycn	80
13	Foxa3	79
14	Gal	75
15	Tead4	73

Figure 18. top signaling hubs and their number of direct interactions in human embryonic and induced pluripotent stem cells, as predicted by ARACNe

CHAPTER 5
Conserved Sub-networks in Human and Mouse Pluripotent Stem Cells

5.1 Introduction

Human and mouse pluripotent stem cells share many similarities between them. They both, presumably, represent very early stages of development where cells are pluripotent and can differentiate to any of the three germ layers. As a result, they can be differentiated in vitro to cell types of all the three lineages. *In vivo*, they form benign teratomas which again consist of multiple tissue types that are representative of endoderm, mesoderm and ectoderm. While their morphologies are not identical, both cell types grow in culture as compact colonies with high nucleus to cytoplasm ratio. They share expression of important markers of 'stemness' such as Nanog, Oct4, Sox2, amongst others. But despite these similarities, these cells are not identical. For some time, it was believed that the observable differences between the mouse and human pluripotent stem cells are attributable to the difference of species. This idea came under questioning when another type of pluripotent stem cell type was discovered from the mouse embryo: epiblast stem cells. These cells showed remarkable similarity with human ESCs in terms or morphology and gene expression. This led to the understanding that human and mouse ESCs could be representative of slightly different stages during embryonic development. But even with this new understanding about possible developmental asynchrony between human and mouse ESCs, there is considerable interest in the differential gene regulation between these cells. As mentioned earlier, these cells do share expression of many important stemness-associated markers. What is not completely understood right now though are which components of cellular networks and signaling are common between the two cell types and which are unique to each of them?

The assimilation and analysis of large amount of gene expression from human and mouse embryonic and induced pluripotent stem cells allows us to compare gene expression and figure out which genes are express highly in both and which ones are not (as discussed in Chapter 4). But, differential expression alone does not offer much insight into functional context. Tools such as protein-protein interaction maps can be used by overlaying gene expression data on them to determine underlying function [184]. Functional interaction networks that combine data from diverse sources such as protein-protein physical interaction data [185, 186], disease/phenotype data [187], phylogenetic profiles [188, 189] etc. incorporate a broad range of interactions and functional relationships [190-192]. Because of their extensive coverage, more so than physical protein interactions maps, these functional maps offer a unique possibility of capturing novel interactions and functional contexts.

5.2 Overlaying differential expression data onto functional networks for subnetwork prediction

We developed an algorithm to find conserved active subnetworks across species by using fold change values from differential expression analysis. The approach overlays gene activity scores on the respective functional linkage or interaction networks to discover dense subnetworks with a large number of differentially active genes with similar expression patterns in both species (Figure 17). Starting from a seed gene, subnetworks are grown in both species by adding nearby genes from the interaction network. Only those neighbors are added that maximize what is defined as an ‘average activity score’ of the subnetwork, while also maintain a minimum desired clustering coefficient of the genes in the subnetwork. Average activity score or the’ network score threshold’ insures that there is a high degree of differential expression in the subnetwork. Activity scores are normalized fold change values from the differential expression analysis. While this criterion guarantees that only those subnetworks are identified that exhibit a high degree of differential expression, the other criterion of clustering coefficient makes sure that dense and highly interconnected subnetworks are identified. The average weighted clustering coefficient of the subnetwork is the ratio of existing connections between the neighbors to the total pairs of neighbors possible. Subnetwork growth is stopped when the average activity score reaches a minimum threshold. This process is then repeated with each differentially active gene in either species serving as the seed. The result is a set of highly clustered subnetworks with a high density of matched differential expression in both species.

To estimate the significance of the obtained subnetworks, randomization experiments were carried out. For both species, the differential expression values were shuffled independently relative to the gene names to remove any connection between them.

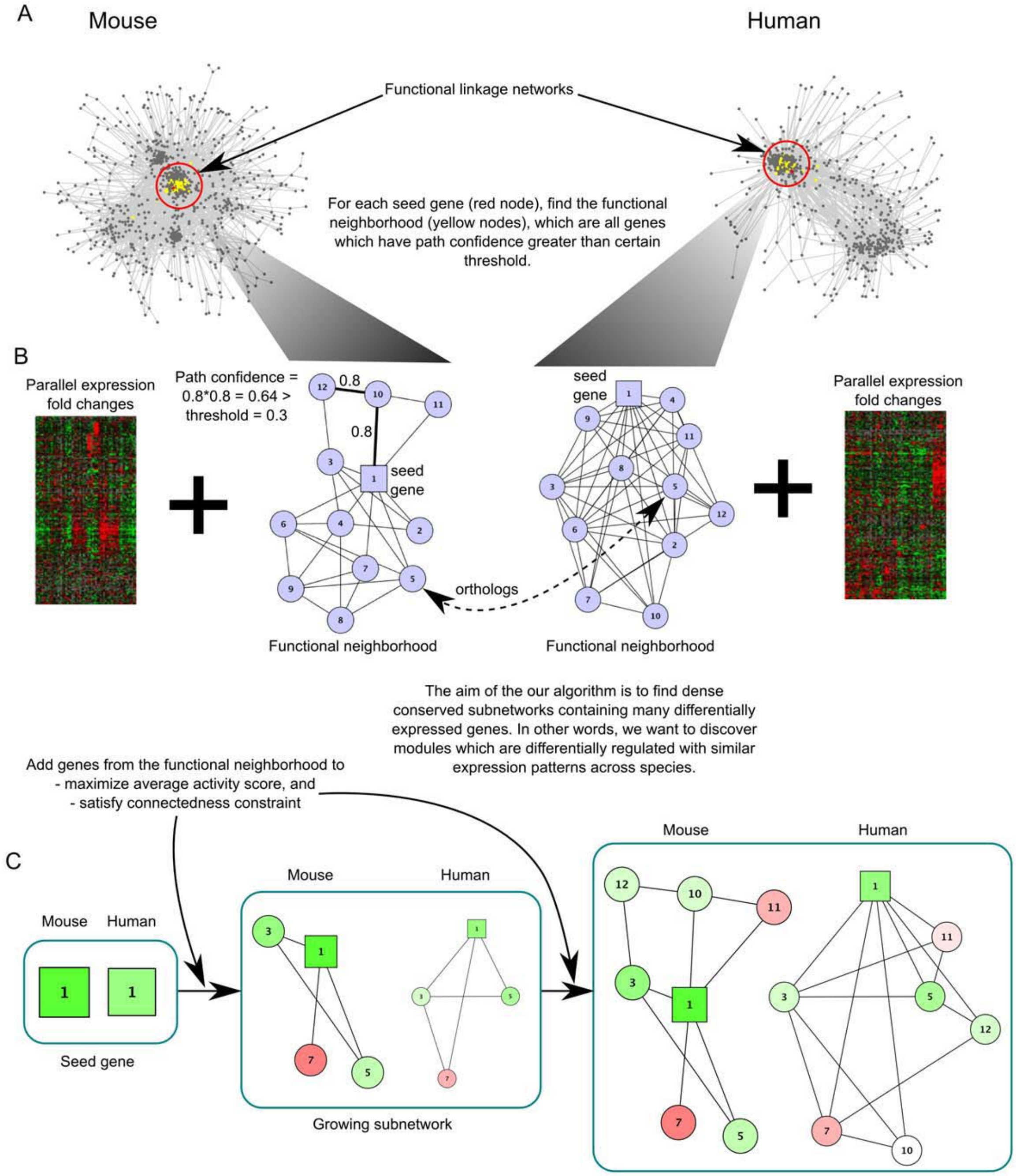

A
Mouse
Human
Functional linkage networks
For each seed gene (red node), find the functional neighborhood (yellow nodes), which are all genes which have path confidence greater than certain threshold.
B
Parallel expression fold changes
Path confidence = 0.8*0.8 = 0.64 > threshold = 0.3
0.8
0.8
seed gene
seed gene
orthologs
Functional neighborhood
Functional neighborhood
Parallel expression fold changes
The aim of the our algorithm is to find dense conserved subnetworks containing many differentially expressed genes. In other words, we want to discover modules which are differentially regulated with similar expression patterns across species.
Add genes from the functional neighborhood to
- maximize average activity score, and
- satisfy connectedness constraint
C
Mouse
Human
Seed gene
Mouse
Human
Growing subnetwork
Mouse
Human

Figure 19. Methodology for discovering conserved active subnetworks across human and mouse pluripotent stem cells. (A) Flowchart describing growth of a subnetwork from a candidate seed gene (red) in the functional network. (B) Genes whose path confidences from the seed gene are above a certain threshold are considered to be functionally related to it. (C) The candidate subnetwork is grown by adding genes iteratively from the functional neighborhood so as to maximize the average expression activity score of the genes in the subnetwork. (Green: Upregulated in stem cells; Red: Downregulated in stem cells) At all iteration steps, the connectivity constraint must be satisfied. The nodes in the subnetwork represent genes and the edge-weights are derived from the functional networks of either species.

The idea behind performing these randomizations is to detect false positives that do not have much biological significance. Fold change values were only shuffled among genes present in the functional linkage network, while the functional linkage network was kept the same. The network discovery algorithm was then run on the shuffled expression data to discover any conserved subnetworks. This entire process was repeated several times to establish a mean and standard deviation for the number of conserved subnetworks identified by chance, which was used to assign confidence values for the real subnetworks. The results of the randomization added confidence to the results because much fewer conserved subnetworks were identified post randomization. Also, the average size of a subnetwork obtained post randomization is 5.7 genes, much lower than the average size of a subnetwork obtained prior to randomization which is about 22 genes.

5.3 Conserved subnetworks in human and mouse pluripotent stem cells

Using the cross-species network discovery algorithm, we are able to find subnetworks reflecting conserved functional modules between mouse and human pluripotent stem cells. We found many of these subnetworks to be monochromatically active in stem cells or differentiated cells. This was not a prerequisite for network discovery, but reflects that the majority of genes supporting a local process are regulated in the same direction. Monochromatic subnetworks up-regulated in stem cells were our primary focus because these reflected potential candidate processes that are necessary for maintaining a pluripotent, self-renewing stem cell state. One of the most significant conserved subnetworks of this type captures the core pluripotency circuit in ESCs (Figure 20(A)).

This network recovers associations between important transcription factors such as POU5F1, NANOG, SOX2 and FGF4, all of which have been shown to form an important transcriptional circuit in embryonic stem (ES) cells, consisting of feed-forward and autoregulatory loops [57]. Chromatin immunoprecipitation experiments have shown that these three proteins exhibit a significant overlap in their binding sites in the genome [57, 59]. The subnetwork links FGF4 to the core signaling circuitry formed by POU5F1, SOX2, and NANOG. FGF4 has been shown to be expressed in the peri-implantation mouse embryo [193] and the SOX2/POU5F1 complex has been shown to activate transcription of FGF4 by binding to an enhancer element [69]. The role of this module has also been studied quite extensively in early embryonic development. FGF4 null mutants in mouse are embryonic lethal due to defective primitive endoderm [133]. The cells of the mouse ICM show a reciprocal expression pattern of FGF4 (ligand) and FGFR2 (receptor). It has been shown that the FGF4 secreted by the epiblast precursor cells is crucial to the differentiation and maintenance of cells of the trophectoderm and extraembryonic endoderm lineages [4, 68].

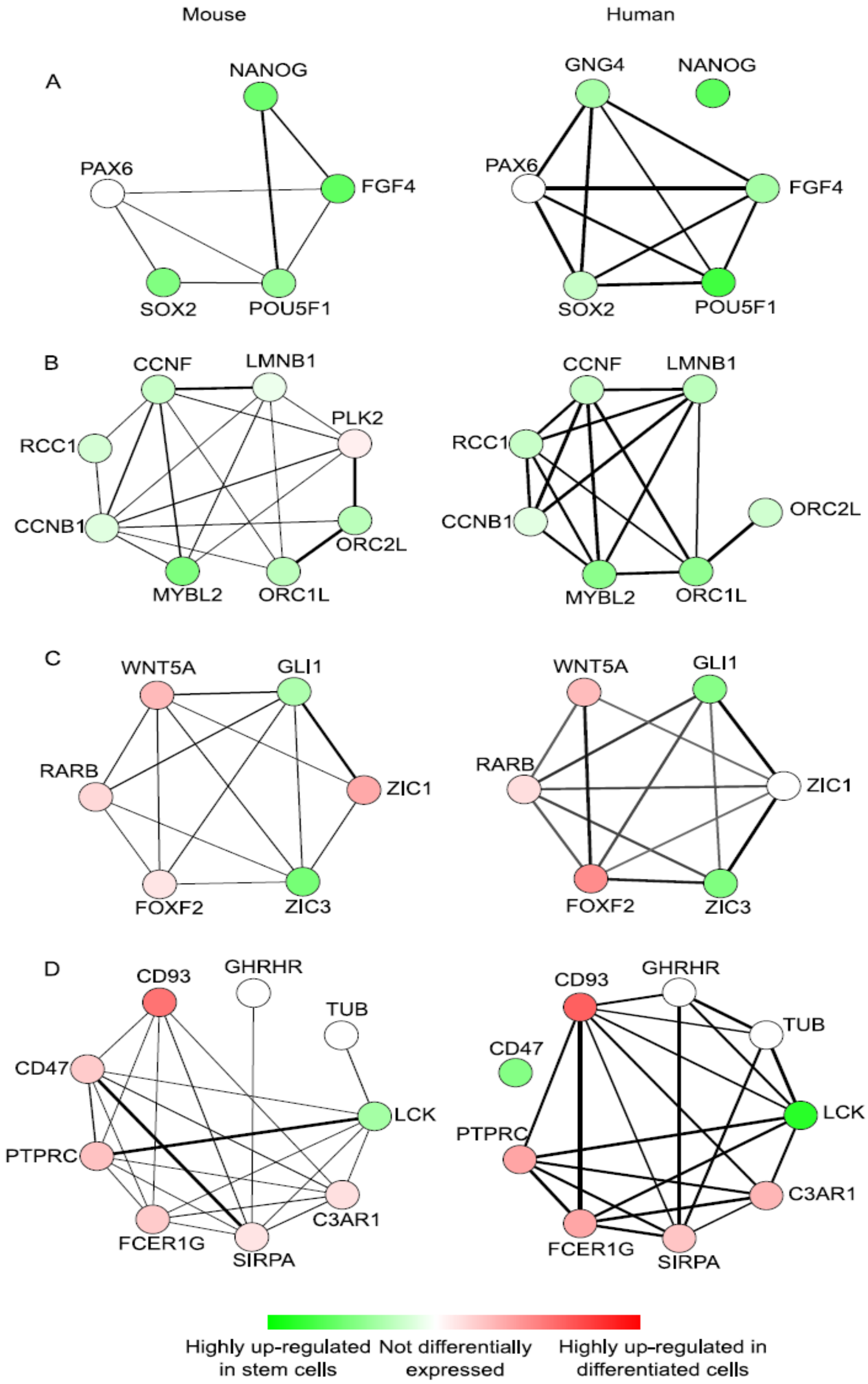


Figure 20. Examples of conserved subnetworks. The edge thickeness represents the confidence of the interaction in the functional network. The nodes represent genes in the subnetwork which are colored according to differential expression. Red: upregulated in differentiated cells; Green: Upregulated in stem cells; White: not differentially expressed. The color intensity indicates the magnitude of normalized differential expression or fold change.

Human ESCs show a striking resemblance to mouse epiblast derived stem cells in terms of morphology and maintenance culture conditions, amongst other characteristics [30, 31]. Thus, this network highlights a core, conserved module active in the pluripotent cells of both the species, irrespective of the downstream effects on cell signaling and morphology. FGF4 stimulation of ERK1/2 signaling in mouse ES cells has been shown to facilitate lineage commitment [194]. In human ES cells, FGF signaling promotes self-renewal by directly affecting the expression of NANOG [195, 196] as well as suppressing expression of genes responsible for reversion to an ICM-like state [32]. Another highly significant subnetwork discovered by our approach pertains to the control of cell cycle progression in ES cells. Both human and mouse ES cells have a very short G1 phase which can be attributed to the constitutively active CDK2/6 [197, 198]. CCNB1 and MYBL2 are two important cell cycle regulators that are expressed at high levels in undifferentiated.

ES cells and their expression decreases rapidly upon induction of differentiation [199]. This happens even before loss of the important regulator proteins such as POU5F1 or NANOG can be detected. The conserved subnetwork highlights the role of these two Genes in the maintenance of cell cycle progression in ES cells. Knockdown of MYBL2 has been shown to induce polyploidy/aneuploidy in ES cells and CCNB1 is a known target of MYBL2 [200]. B-MYB is also crucial for ICM development in mice embryos [201]. The role of CCNF in ESCs has not been explored but yeast two hybrid assays have shown that the NLS domain of CCNF can regulate nuclear localization of CCNB1 [202].

Many conserved subnetworks also included genes that are upregulated during the initiation of differentiation. This supports the idea that the maintenance of ES cell phenotype requires the suppression of differentiation-associated gene expression as well. One interesting example of this phenomenon was highlighted in a third subnetwork discovered by our approach, which was centered on the protein ZIC3. ZIC3 has been shown to be required for maintaining pluripotency of mouse ESCs by suppressing endoderm specification [117] while GLI1 has an important effect on ESC proliferation [203]. These two proteins are known to work in coordination for transcriptional activation or repression [116]. Both of these genes code for DNA binding zinc finger proteins and they share and recognize highly conserved zinc finger domains. The down-regulated genes in the subnetwork, namely, WNT5A, FOXF2 and RARB, play important roles in the differentiation of ESCs [204-206]. It is interesting to observe that these genes have GLI binding sites in their promoter region or cis-regulatory domains, which suggests that GLI1 and ZIC3 could potentially regulate their expression in ES cells [207, 208]. Also, GLI proteins participate in regulation of Hedgehog signaling, of which RARB and FOXF2 are members, and GLI is also known to regulate the members of WNT family [209]. These functional interactions and coordinated expression strongly suggest ZIC3 and GLI1 might be responsible for suppressing the expression of genes such as FOXF2, WNT5A and RARB. This network in particular provides an illustrative example of how subnetwork discovery can provide novel testable experimental hypotheses. This hypothesis could be explored experimentally through RNAi knockdown of ZIC3 and GLI1 in ESCs to check for resultant changes in expression of the other genes in the network. Lim et al. [117] conducted RNAi knockdown of ZIC3 in human and mouse ESCs and saw enhanced expression of endodermal transcripts like SOX17 and PDGFRA. Further experiments could also be used to check for direct binding of ZIC3 and GLI1 to the promoter regions of the differentiation associated genes. The subnetwork also highlights the striking observation that the gene ZIC1, despite sharing 69% homology with ZIC3, does not show the same trend in expression in either mouse or human pluripotent stem cells. While ZIC2 and ZIC3 have been suggested to have partially overlapping or redundant roles in suppressing endoderm in ESCs, the role of ZIC1 in this context has been not been explored much. Further overexpression studies of this gene could be used to elucidate its exact role in this network.

Another interesting subnetwork found by our approach was centered on the seed gene SIRPA. The only gene in the whole subnetwork that is found to be up-regulated in mouse and human pluripotent stem cells is LCK (Figure 20(D)). LCK is one of the eight SRC family kinase genes, which are known to play crucial roles in regulating signals from a variety of cell receptors, affecting a variety of cellular processes such as differentiation, growth and cell shape [210]. Members of this family, namely Hck and Lck, have been implicated in the maintenance of self-renewal of murine ESCs [196]. Cyes, along with Hck, have been shown to be regulated by LIF in mouse ESCs and the expression of their active mutants allows the maintenance of these cells at lower concentrations of LIF [211]. Other studies have also reported the evolutionarily conserved transcriptional co-expression of LCK in human and mouse ESCs based on transcriptomic studies [212]. LCK has also been shown to induce STAT3 phosphorylation and this is believed to cause transformation of cells having constitutive LCK activity [195]. All of the other genes in the sub-network are down-regulated in ES cells, which may be due to the fact that the expression of SFKs is generally associated to lineage-restricted patterns in the adult, such as, the expression of LCK in T lymphocytes. While the hypotheses suggested by the discovered subnetworks ultimately require experimental follow-up, these examples illustrate that the networks capture many of the well-characterized processes supporting stem cell pluripotency as well as implicating some novel players. In general, the process of active subnetwork discovery can play an important role in interpreting differential expression or other genome-wide data. Active subnetworks, and in particular those that are conserved across species, provide evidence that a whole process or pathway is up/down-regulated, which is more definitive than the type of information provided by a differential expression list, for example. A single highly differentially expressed gene is less compelling than an entire functional module with evidence of differential expression. Furthermore, because the underlying functional linkage networks are based on large collections of genomic data, our approach can potentially identify functional modules that are not yet characterized, but that play a critical role under the conditions being studied.

5.4 Identification of species-specific subnetworks

We modified the cross-species network discovery algorithm to discover subnetworks that are markedly different in the expression patterns between the two species. These subnetworks represent tightly interconnected groups of genes or proteins that are active only in one of the species or where the expression changes are in opposite directions, highlighting places where pluripotent stem cell signaling differs between human and mouse. Through randomization experiments similar to the conserved subnetwork identification approach we found that we were able to find such non-conserved network signatures approximately twice as frequently as on randomized expression profiles. We note that this is a substantially lower signal-to-noise ratio than for the conserved subnetwork discovery approach, for which we achieved approximately 20-fold improvement over random, suggesting that statistically significant species specific active subnetworks are harder to discover. This is not surprising given that the relatively frequent appearance of random subnetworks in a single species, which cannot be easily classified as statistical artifacts or biologically relevant changes across species. The species-specific network discovery problem is not able to take advantage of the noise filtering property of the conserved network search described above. Nevertheless, we find interesting subnetworks which highlight differences between gene expressions in mouse and human stem cells. For example, one species-specific subnetwork (Figure 21) recapitulates the well-known difference in BMP signaling between human and mouse ESCs. Mouse ESCs require BMP2/BMP4 to induce the expression of Inhibitor of differentiation (Id) genes via Smad pathway for self-renewal [22]. Thus, exogenous addition of LIF and BMP4/2 is required to maintain mouse ES cells in culture without differentiation.

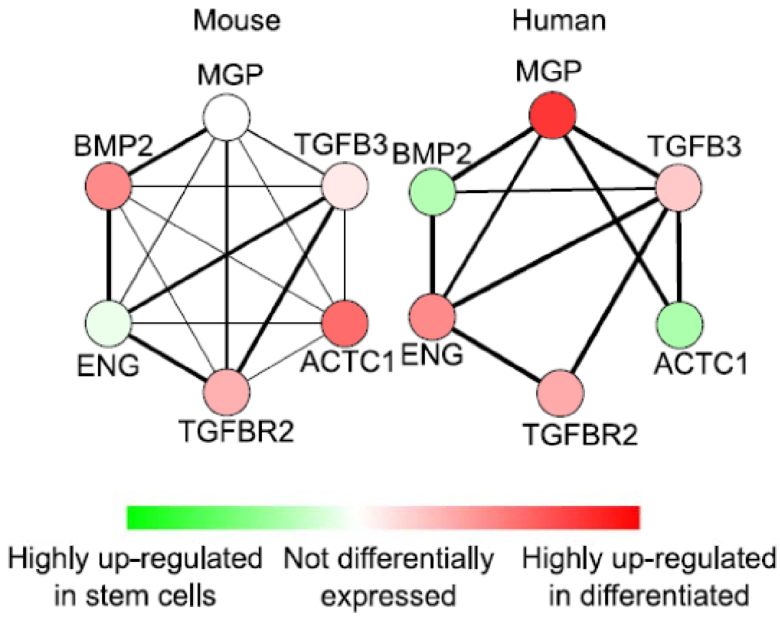

Mouse
Human
MGP
BMP2
TGFB3
ENG
ACTC1
TGFBR2
MGP
BMP2
TGFB3
ENG
ACTC1
TGFBR2
Highly up-regulated in stem cells
Not differentially expressed
Highly up-regulated in differentiated

Figure 21. An example of species-specific subnetwork that shows differential BMP signaling between human and mouse pluripotent stem cells. The thickness of the edges represents the strength of the interaction in the functional network. The color of the nodes indicate differential expression as indicated above. The intensity of the color indicates the extent of normalized fold change of the node/gene.

On the other hand, human ESCs cultured in unconditioned medium exhibit high levels of BMP signaling which causes the cells to differentiate. Mouse EpiSCs, like human ESCs, differentiate to trophoectoderm upon BMP4 induction [30]. This needs to be suppressed through an antagonist such as noggin to maintain these cells in an undifferentiated, self-renewing state [196]. The other genes in the subnetwork that show opposite trends in differential expression between human and mouse ES cells are MGP, ACTC1 and ENG. Endoglin (Eng) is an accessory receptor for several TGF-b growth factors, including BMP2, and has been shown to be crucial for embryonic hematopoiesis [213]. Matrix GLA protein (MGP) is a small matrix protein that has been shown to have a direct interaction with BMP2 and has been shown to modulate BMP signaling [214]. The potentially disparate role of these genes in mouse and human ES cells can be explored further.

5.5 Discussion

With the advent of microarray technology, it has become quite easy to study genome wide changes in transcription. High throughput gene expression data allows us to not only look at gene change across the whole genome, but also lends itself to developing hypotheses, discovering trends, etc. though the use of statistical analyses. With the ever increasing scientific interest in stem cells, there is a large amount of gene expression data that, when processed carefully to remove biases and noise, can be used to ask biologically meaningful questions.

Gene expression data involving human and mouse pluripotent cells was obtained from public repositories. This included data from a broad range of experiments involving both embryonic and induced pluripotent stem cells. The data was normalized to adjust for any systematic biases and filtered to remove absent probes. NMF, a dimensionality reduction tool, was used to detect biological classes in the data. NMF was able to not only pick out a distinct class of pluripotent stem cells from the large amount of samples, but it also predicted a gene expression profile characteristic of these cells, called the ‘metagene’. The metagene is a ranked order of genes that are responsible for separating the pluripotent cells from all other classes of samples. In future studies, this ‘metagene’ can be used to define a ‘gene space’ which is able to define ‘boundaries’ of the pluripotent landscape.

The similarities and differences between human and mouse pluripotent cells and how they relate to specific stages in embryonic development is an extremely interesting area of research. With the recent isolation of epiblast stem cells, the field has developed a much better understanding of the differences in pluripotent cells of the two species. But as mentioned earlier, in addition to experimental analysis, gene expression data can also be used to understand these differences better. By overlaying our gene expression data onto functional interaction maps of the human and mouse genome, we have discovered been able to detect small pockets in the transcriptional circuit that show conserved or species specific expression. Such analyses are important not just for developing insights into pluripotent stem cells but also into early embryonic development.

CHAPTER 6
Bone marrow-derived rat stem cells have Primitive Endoderm-Precursor Phenotype

6.1 Introduction

Bone marrow-derived rat adult stem cells or MAPCs have isolated after depleting the hematopoietic cells from the bone marrow. A rapidly proliferating population of small cells is obtained under MAPC-culture conditions in 8-18 weeks [2]. These cells are considered multipotent for their ability to differentiate in-vitro into multiple cell types of the endoderm and mesoderm lineage.

Upon the derivation of extra embryonic endoderm precursor or XEN-P cells, it was observed that MAPCs bear a striking resemblance to these cells, not only in morphology, but expression of important markers.

As an example, both cell types uniquely express the pluripotency markers Oct4 and Rex1, while simultaneously expressing markers of primitive endoderm such as Gata6, Pdgfra and Sox17 [3]. But there are some very obvious differences between the two cell types such as the tissue or stage of development that the cells are isolated from. While, XEN-P cells are isolated from E4.4 rat embryo, the MAPCs are obtained from the bone marrow of an adult rat. In an effort to understand how such very similar cells could be derived from such disparate sources, Verfaillie and colleagues (Katholieke Universiteit, Belgium) have worked towards deriving MAPCs from the embryo. Given the long period of derivation required for the isolation of MAPCs from bone marrow, Verfaillie et al. have proposed that MAPCs could be arise as a consequence of culture-induced reprogramming (LoNigro et al., Unpublished work). The authors have been able to derive MAPC/XEN-P-like cells from the E4.5 rat embryo under MAPC culture conditions. Unlike the long time period for MAPC derivation, the authors have observed that the embryo-derived, MAPC/XEN-P-like cells arise about a week and continue to proliferate. Since this period is much too small to induce any extensive reprogramming event in the cells of the embryo, it can be believed that these cells arise from the primitive endoderm-precursor population existing in the ICM, more so because there is evidence now of the existence of such a population in ICM [4, 134].

6.2 Microarray profiling of MAPC gene expression

A comparison of the whole genome gene expression in pluripotent ESCs (rESCs), bone marrow derived rat adult stem cells (rMAPCs) and the embryo-derived rat stem cells (rMAX8) shows that the embryo derived MAX8 cells show a striking similarity to the bone marrow-derived MAPCs. The ESCs are quite different to both the cell types, as shown in Figure 22.

Both MAPCs and MAX8 express high levels of markers of ICM, such as, KLF4, KLF5, TDGF1, REX1 (ZFP42), SSEA-1, TBX3, MYCN and POU5F1 (OCT4) but they do not express SOX2, KLF2, ZIC3, ESRRB etc (Figure 23). Both these cells do not express markers of late epiblast such as FGF5 and BRACHURY (T), just like rat ESCs. On the other hand, epiblast stem cells (EpiSCs) express these genes. From this analysis too, it was confirmed that NANOG expression showed the biggest fold change (40 fold higher in ESCs) between the pluripotent ESCs and MAPCs/MAX8 cells. On the other hand, while SOX2 expression is also very low in these cells, it is not completely absent like NANOG.

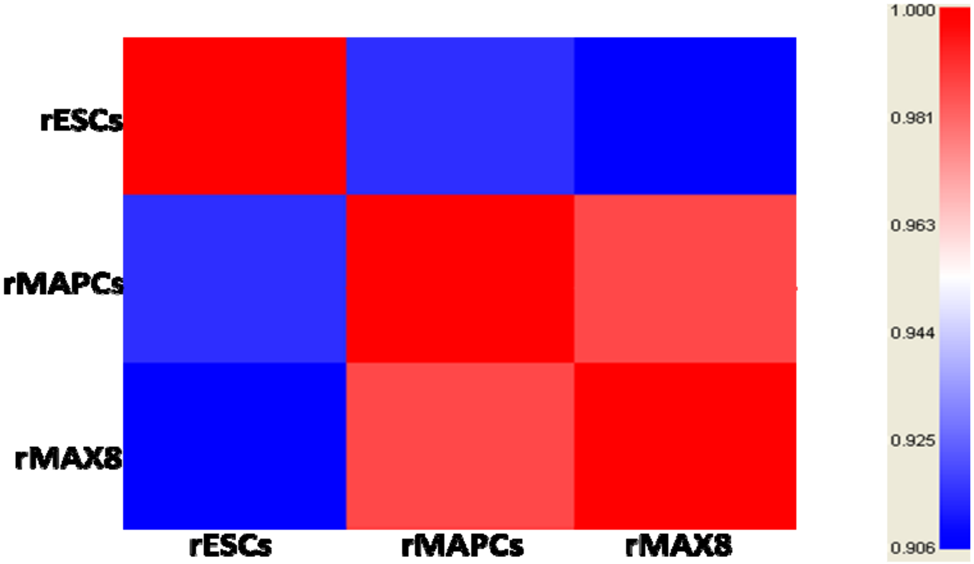


Figure 22. Pearson correlation between gene expression in rat ESCs (rESCs), bone marrow-derived rat adult stem cells (rMAPCs) and embryo-derived rat stem cells (rMAX8)

The MAPCs and MAX8 express high levels of markers of primitive endoderm such as GATA6, GATA4 and SOX7 (Figure 24).Most of these markers are either not expressed by rESCs or expressed at basal levels. MAPC and embryo-derived cell types also express high levels of parietal endoderm markers like TPA (PLAT) but not of some visceral endoderm markers like TTR, BMP2, and HNF4α. This is in line with the fact that XEN-P cells which are similar, if not identical to these cells, also lean towards parietal endoderm in their in-vivo contribution to development [3].

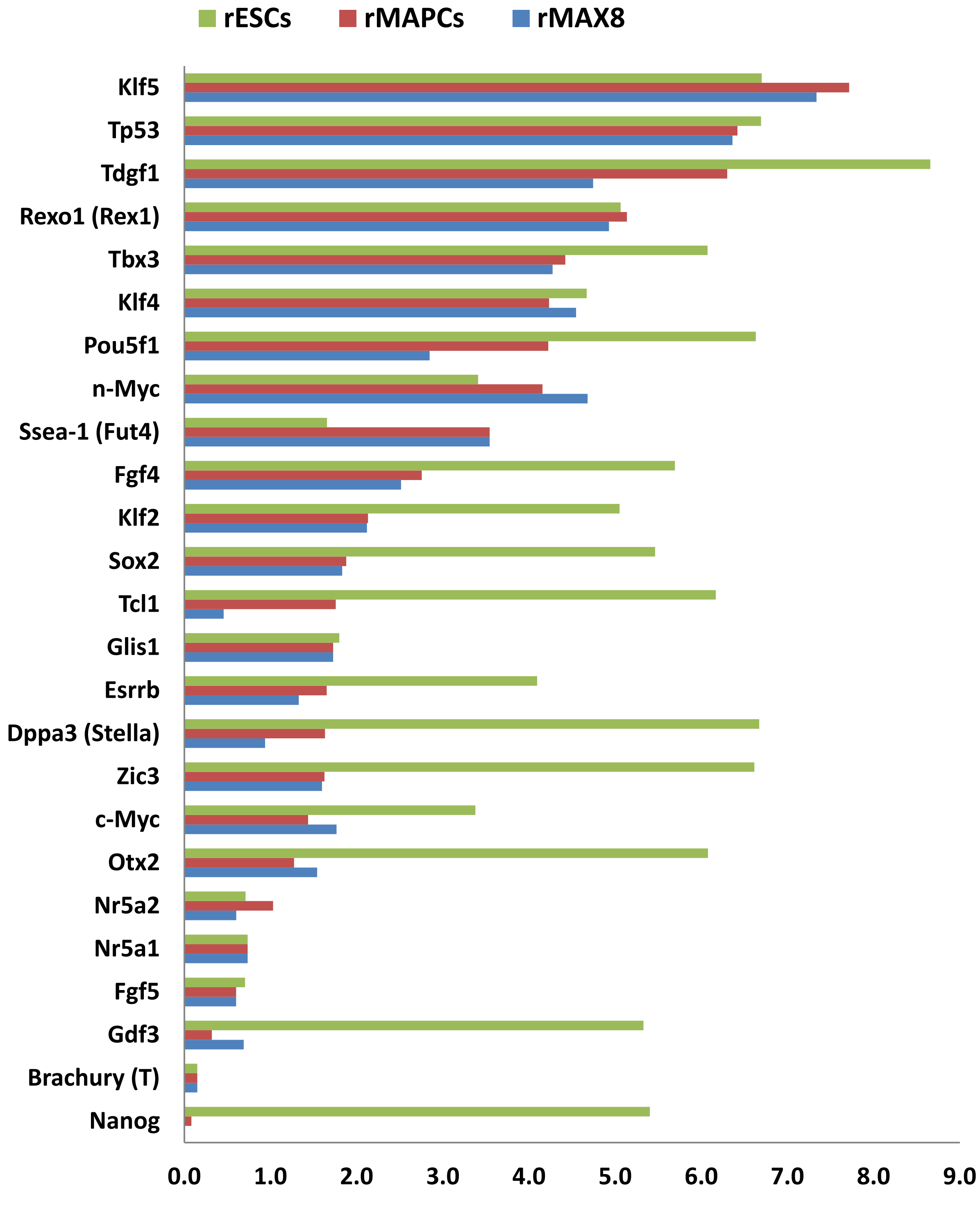


Figure 23. Log intensity of expression of ICM and epiblast specific markers in rat ESCs (rESCs), bone marrow-derived rat adult stem cells (rMAPCs), embryo-derived rat stem cells (rMAX8)

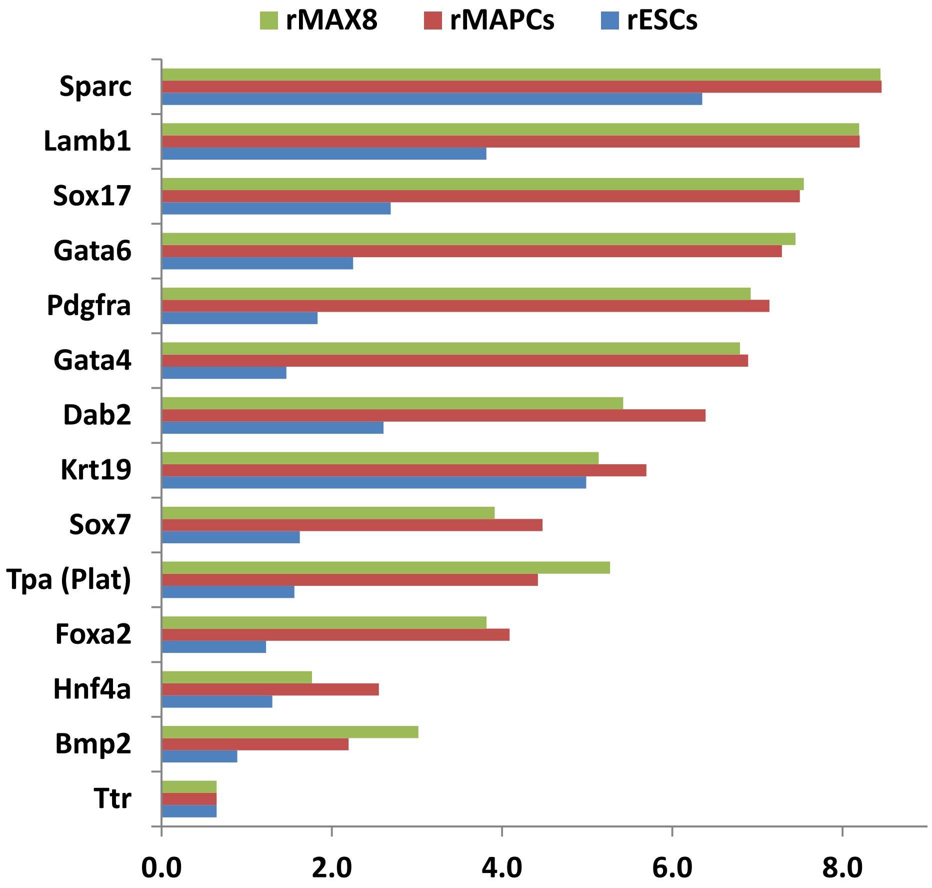


Figure 24. Log intensity of expression of markers of primitive, parietal and visceral endoderm in rat ESCs (rESCs), bone marrow-derived rat adult stem cells (rMAPCs) and embryo-derived rat stem cells (rMAX8)

The microarray analysis shows how bone marrow derived MAPCs and embryo-derived MAX8 cells are strikingly similar, if not identical, to each other and also the embryo-derived XEN-P cells. Also, the fact that these two similar cell types express certain markers of ICM but not the epiblast suggest that they likely represent the primitive endoderm population present within the ICM.

6.2.1 Isolation of MAPC-like cells from un-induced differentiation of rat embryonic stem cells

Given that ESCs are considered the in-vitro counterpart of the ICM of the developing embryo, the rat ESCs should be able to give rise to primitive endoderm precursor of MAPC/MAX-like cells *in-vitro*. Mouse ESCs have been shown to differentiate to primitive endoderm by over expression of the master regulator, GATA6, or by knockdown of Nanog. But as such cells can be readily derived from the ICM of the embryo in MAPC culture conditions; they could also be derived from the ESCs without the need for genetic manipulation. To probe this, we plated rat ESCs in MAPC culture conditions, on fibronectin in medium containing LIF and PDGF. We observed an onset of spontaneous differentiation during the first week of culture where round, refractile MAPC-like cells could be easily observed in a heterogeneous population of cells (Figure 25). Low density maintenance of these cells at 200-300 cell/cm^2 (similar to MAPC culture) quickly gave rise to a more homogenous culture on cells that are morphologically identical to MAPCs and MAX8.

Figure 26 below shows the change in gene expression of the ESCs from day 0 to day 10 in MAPC culture conditions. The cells quickly lose complete expression of Sox2 and Nanog but not Oct4, which is also expressed by MAPCs. The cells also gain high expression of primitive endoderm marker genes GATA6, GATA4, SOX17 and PDGFRα.

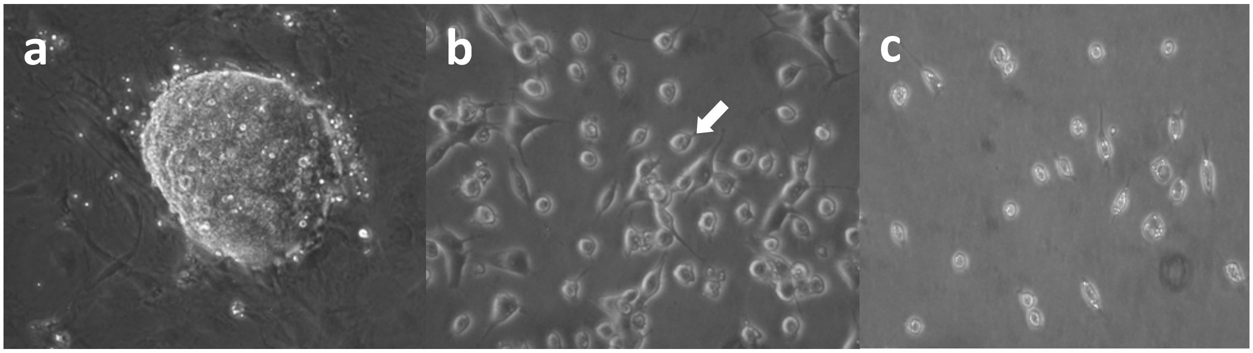


Figure 25. Culture induced differenetiation of rat ESCs. (a) Day 0 image of undifferentiated rESC colony, (b) Day 5 culture show heterogenous mix of cells with a significant number of MAPC-like cells, shown using the arrow, (c) Day 12 culture is almost uniformly MAPC-like cells

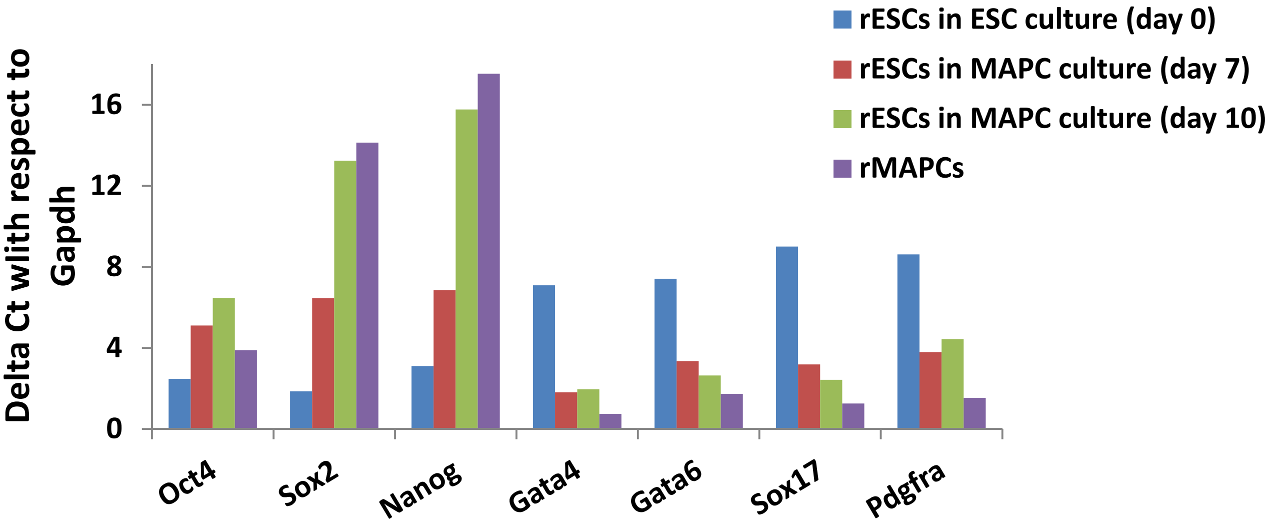


Figure 26. Gene expression in rat ESCs under MAPC conditions at day 7 and day 10

Though a complete analysis of the state of the final cell type derived by putting these ESCs in MAPC conditions is currently underway, the morphological resemblance of these cells to MAPCs is striking and the gene expression profile also points towards the similarity between the two cell types.

6.3 Discussion

Combining all this analysis, suggests that the rat bone marrow-derived MAPCs are very similar, if not identical to the primitive endoderm-precursor cells obtained from the rat embryo and the rat ESCs. While the derivation of such precursor cells from the ICM or the ESCs is direct and occurs by providing conducive culture conditions, in the form of MAPC culture system, for self-renewal of these cells *in-vitro,* bone marrow-derived MAPCs are possible arising due to extended culture-induced reprogramming.

Chapter 7
Transcriptional network in rat primitive endoderm-precursor cells

7.1 Introduction

In chapter 4, meta-analyses of gene expression data from pluripotent stem cells was performed to gain insights into the core transcriptional network for maintaining pluripotency. A 'metagene' signature was identified that clearly differentiates pluripotent embryonic and induced pluripotent stem cells from other partially or completely differentiated, or partially reprogrammed cell types. Important information about cellular signaling 'hubs' was discovered using reverse engineering tool 'ARACNe'.

As discussed in the previous chapter, MAPCs are strikingly similar to the primitive endoderm precursor cells recently observed in the ICM of mouse blastocyst [4]. These cells have high expression of the pluripotency marker Oct4 although the Nanog expression had decreased to very low levels and the cells had turned on the expression of key primitive endoderm marker Gata6. The epiblast precursor population, on the other hand, maintained high Nanog (while maintain their Oct4 too) and showed little to no expression of Gata6. The existence of a cell co-expressing Gata6 and Oct4 has been corroborated by other studies too [34, 134]. This gene expression shift that happens in early embryonic development is important to our understanding of how pluripotency is established and maintained in pluripotent cells. Due to this developmental proximity of MAPCs and MAPC-like other primitive endoderm cells to the pluripotent stem cells of the embryo, it is important to understand the core transcriptional network in MAPCs. This network is likely to be very closely associated to the gene network in pluripotent cells. Thus, MAPCs can be used as a model system to analyze the transcriptional dynamics in pluripotent stem cells of early embryo. Also, how the growth factor requirements of MAPCs are related to the maintenance of this unique gene expression profile is an important question to answer.

As discussed in the last chapter, the culture conditions used for deriving adult stem cells from bone marrow of adult rats can be used for successfully deriving primitive endoderm-precursor cells from the rat embryo, as well as, rat ESCs. The culture conditions, as discussed in Chapter 3, consist of low density culture of cells (200-300 cells/cm^2) in basal medium supplemented with growth factors LIF (leukemia inhibitory factor), PDGF (Platelet derived growth factor) and EGF (epidermal growth factor). While EGF is believed to help in supporting the growth of the cells, the first two growth factors, LIF and PDGF, are believed to be critical for the maintenance of the primitive endoderm phenotype. We wanted to study the effect of both these cytokines on the primitive endoderm precursor phenotype of bone marrow-derived MAPCs.

7.2 Role of LIF in maintenance of MAPCs

LIF has been shown to sustain self-renewal of mouse embryonic and induced pluripotent stem cells by signaling through JAK/STAT3 [215]. This signaling has also been shown to be pivotal in the maintenance of the ‘naive’ state of pluripotency [216]. To observe the effect of LIF withdrawal on MAPCs, we cultured the cells in MAPC medium without LIF at normal growth density of 200-300 cells/cm^2 and analyzed the cell growth and gene expression. Almost an immediate effect was observed on the growth rate of the cells with the doubling time increasing from about 12 h to 18 h in two passages (Figure 27). Increased cell death could be seen in culture and no cells could be obtained by passage 4 for reliable cell count. Analysis of gene expression following LIF withdrawal showed a steep decrease in the expression of OCT4 even as early as day 2. Surprisingly, while expression of GATA4 and SOX17 decreased too, the expression of GATA6 either stayed the same or showed a very slight increase (Figure 28). While the exact interpretation of this change in not understood right now and required further analysis, one possible way to interpret this is to assume some sort of differentiation towards parietal endoderm. Both XEN and XEN-P cells have a predisposition to differentiate towards the parietal endoderm. Also, it has been shown in embryonic development that continued expression of GATA6 is seen in parietal endoderm cells while it can no longer be observed in visceral endoderm cells that continue to express GATA4 [217, 218]. These results highlight the importance of LIF in marinating a precursor phenotype of these cells where pluripotency-associated genes continue to express while the cells are also primed to differentiate to primitive endoderm. In such a scenario, withdrawal of LIF could lead to collapse of the self-renewing state and cause cells to differentiate to the extra embryonic lineage.

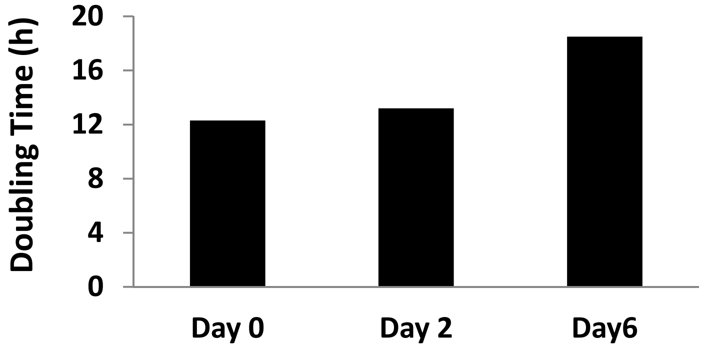


Figure 27. Increase in doubling time of MAPCs upon LIF withdrawal

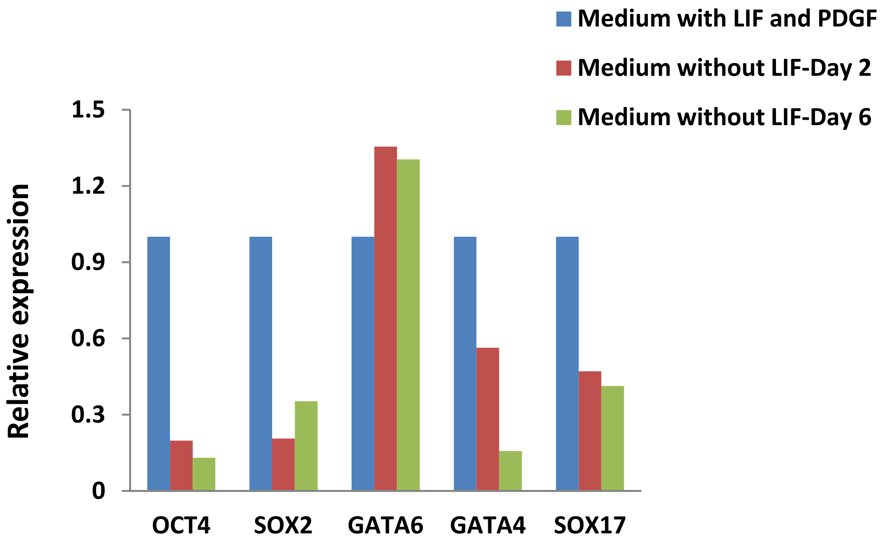


Figure 28. Effect of LIF withdrawal on gene expression days 2 and 6 post withdrawal

7.2.1 LIF signaling in primitive endoderm precursor-like cells

As discussed above, LIF signaling is crucial for maintenance of MAPCs. This cytokine has been shown to be equally important for the self-renewal of mouse ESCs [219] and investigations into its precise role has revealed complex signaling machinery in place.

The cytokine signals by binding to the heterodimeric receptor complex consisting of the LIF receptor and gp130 receptor. Upon its binding, the tyrosine kinase, Janus kinase (JAK), phosphorylates tyrosine resides on both LIFR and gp130 receptor, which leads to the recruitment of signal transducers and activators of transcription (STAT1 and STAT3) to their SH2 domains. These proteins are then phosphorylated by JAK-mediated phosphorylation causing formation of homodimers and heterodimers which translocates to the nucleus and act as transcription factors [220]. The stimulation by LIF also signals through other pathways through the SHP2 effector protein which binds to intracellular domains of both gp130 and LIFR [221]. Phosphorylated SHP2 can bind to the adaptor protein Grb2 and activate Ras/ERK signaling or bind to Gab (Grb2-associated binder) to activate the PI3 kinase pathway. Out of these 3 signaling cascades that LIF binding can initiate, JAK/STAT3 and PI3 kinase have been linked to self-renewal of ESCs while the Ras/ERK signaling has been shown to suppress Nanog and induce differentiation into extra embryonic endoderm [222-224]. The self-renewal in murine ESCs is due to a balance between the differentiation-inducing signals from Erk/MAPK pathway versus the self-renewal signaling from JAK-STAT3.

7.2.1.1 STAT signaling in naïve state of pluripotency

While the importance of JAK-STAT3 signaling in maintaining self-renewal of mESCs, following activation through LIF, is conclusive, the exact mechanism by which this happens is not completely understood yet. Some self-renewal-associated genes have been shown to have a link to the phosphorylated STATs. Niwa et al. identified KLF4 as a downstream target of JAK-STAT3 signaling in mouse ESCs. They showed that this cascade predominantly activates SOX2 although this is not completely exclusive [224]. The authors also found TBX3 as a target gene of the PI3 kinase signaling also activated through LIF, which is an activator for downstream NANOG. KLF4 is an important component of the gene network in pluripotent stem cells but it is not the sole target of JAK-STAT3 signaling in these cells. Other, yet undiscovered, targets of this signaling have been shown to be limiting for complete reprogramming and establishment of the ground/naïve state of pluripotency in epiblast stem cells [216].

7.2.1.2 Role of SOCS3 in JAK/STAT3 signaling

Suppressors of cytokine signaling or SOCS proteins are negative regulators of LIF and Interleukin-6 signaling [225-227]. These proteins, SOCS-1 and SOCS-3 in particular, have a central SH2 domain that can interact with the tyrosine kinase domain of JAK proteins and prevent it from activation STAT proteins. SOCS3 is activated in mouse ESCs in response to LIF signaling and at its physiological level in these cells; it is required for self-renewal [228]. SOCS3-null mouse ESCs show enhanced STAT3 phosphorylation and Erk/MAPK signal but reduced capacity for self-renewal and a propensity to differentiate to primitive endoderm. This was due to the enhanced and constitutive SHP-2 activation in absence of any competition from SOCS-3 leading to an increase in the Erk/MAPK differentiation-inducing signal. This phenotype could be rescued using inhibitors of Ras-MAPK signaling. Similar to OCT4, SOCS3 levels in murine ESCs are also required to be within a window of expression and elevated levels of SOC3 have been shown to be detrimental to self-renewal of cells. Over expression of SOCS3 reduces STAT3-dependent transcription and cell growth [229]. This is due to the association of increased SOCS3 to JAK1 and JAK2, which leads to reduced phosphorylation of STAT3 by these factors, as well as, proteasomal degradation of these proteins [230, 231]

In combination with the results of growth factor signaling, microarray analysis can also lend useful insights into the genes active in maintaining the MAPC or XEN-P phenotype. Figure 29. shows a comparison of expression of targets of LIF signaling between primitive endoderm-precursor cells and ESCs. Almost all known transcription factor targets of LIF signaling, as shown below, are expressed at higher levels (fold change greater than 2) in rat ESCs, except for SOCS3. SOCS3 is expressed at almost two fold higher expression levels in MAPCs, as well as, embryo-derived MAX8 cells. This is an important difference between the pluripotent ESCs and the primitive endoderm-precursor cells and could explain the disparate outcomes of LIF signaling in the two cell types. One of the downstream targets of LIF signaling, SOX2, is strongly differentially expressed between the two cell types, showing high expression in ESCs but very low expression in MAPCs and MAX8 cells. An elevated level of SOCS3 could account for this difference in expression of this very important pluripotency-associated gene. Similarly, enhanced SOCS3 levels could also lead to reduced PI3 kinase signaling since SOCS3 binds to the phosphorylated gp130 receptor through its SH2 domain, thereby preventing the binding of SHP2 protein to the receptor. While this could contribute to reduced activation of Nanog in the primitive endoderm precursor-like cells, this is not the likely to be the dominant suppressor of Nanog expression in these cells. As mentioned earlier, Nanog expression is strikingly absent in these cells which express high levels of OCT4 and some basal levels of SOX2, possibly suggesting very strict control of its expression occurring through multiple mechanisms. Two such important mechanisms will be discussed next in this document.

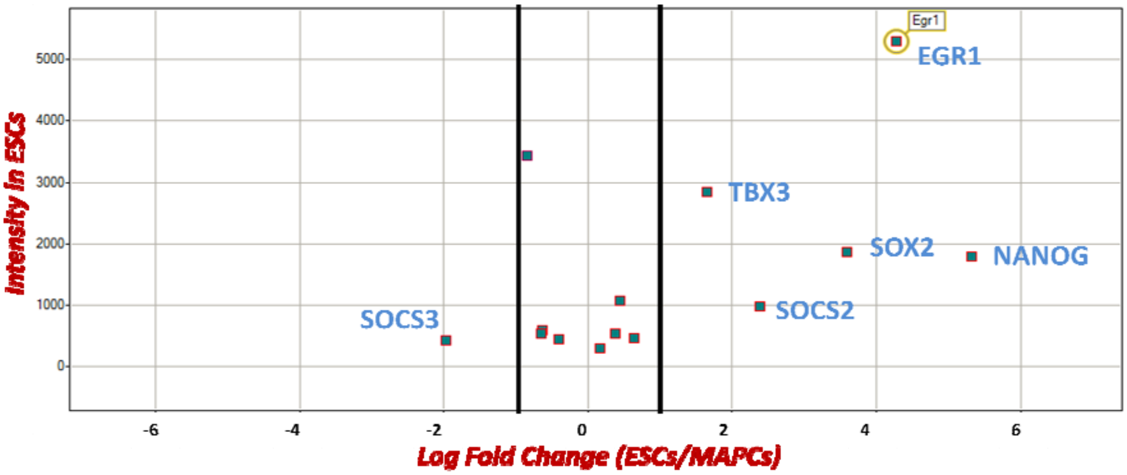


Figure 29. Fold change of targets of LIF signaling between rat ESCs (ESCs) and bone marrow derived rat primitive endoderm-precursor cells (MAPCs)

7.2.1.3 Effect of SOCS3 on expression of Nanog

As discussed above, reduced levels of PI3 kinase signaling could influence NANOG levels in rat MAPCs and MAX8 cells but there is yet another aspect of LIF independent JAK signaling that could add to the suppression of NANOG by elevated SOCS3 levels. JAK2 can phosphorylate tyrosine 41 in histone H3 (H3Y41) and this prevents HP1α from binding to H3 and thereby assembling at the promoter regions of genes [232]. In mouse ESCs, JAK2 was shown to directly affect the expression of NANOG through this mechanism [233]. Overexpression of SOC3 has been shown to bind to JAK2 and target it for degradation and this could hamper NANOG activation through this mechanism in MAPCs and MAX8 cells.

7.3 Effect of PDGF withdrawal

PDGFRα is an important marker of primitive endoderm in the developing blastocyst [234]. It has been shown to be activated by GATA6 in early blastocyst and its expression is crucial for rhe derivation of XEN cells in vitro [235]. Primitive endoderm precursor cells express high levels of the receptor on their surface and the PDGF in medium can signal through this receptor to turn on downstream targets such as MEK1/2 and PKC. To understand the role of this receptor signaling, we cultured MAPCs in medium without PDGF but containing LIF. We observed reduction in growth of the cells around passage 3 when the doubling time increased to about 15 h from the initial 12 h (Figure 30). An important difference here, in comparison to LIF withdrawal, was that growth did not progressively reduce. The cells seemed to have reached a state of lower growth rate and continued to maintain morphology and proliferate at that rate. Even extended passaging did not show a further reduction in growth. This seemed to suggest that after the initial slowing down, the cells reached another stable state and the cultures did not terminate like what was seen upon LIF withdrawal.

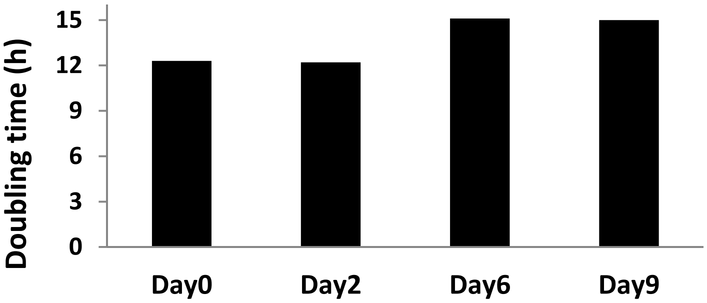


Figure 30. Increase in doubling time of MAPCs in medium without PDGF

Analysis of gene expression change upon PDGF withdrawal shows an initial decrease in expression of OCT4 and SOX2, as well as, GATA4 (Figure 31). Surprisingly, the GATA6 expression increased quite significantly from day 2 to day 6.On day 9, however, the OCT4, GATA6 and GATA4 levels returned to their day 0 levels and only SOX2 levels remained much lower. This is in coherence with what is seen from the doubling time plot. The cells seem to have possibly reached a new self-renewing state by day 9 where the OCT4 and GATA6 levels are similar to day 0 levels. It can be concluded that PDGF has an effect on cell proliferation but is not critical for the maintenance of the primitive endoderm-precursor phenotype. This is also similar to what is observed during development where PDGFRα helps in proliferation of the primitive endoderm but is not necessary for establishing the primitive endoderm [235].

In terms of maintenance of MAPC or MAX8 phenotype, these results suggest that LIF provides the necessary signal required for self-renewal of these cells, in absence of which, the cells can not be maintained in given culture conditions.

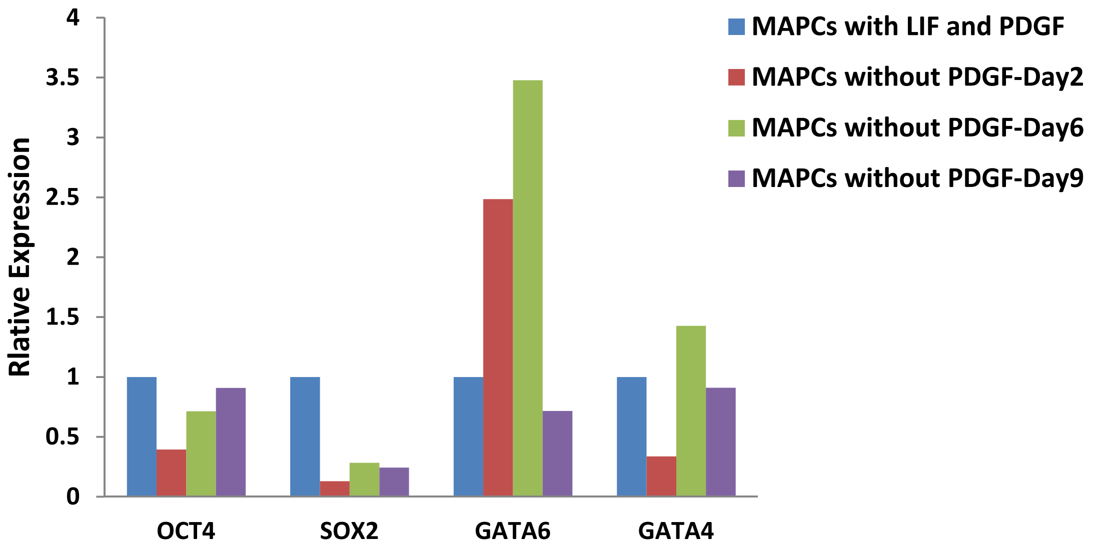


Figure 31. Effect of PDGF withdrawl on expression of Oct4, Sox2, Gata6 and Gata4 in MAPCs

7.4 Transcriptional Network in MAPCs

Based on the results of our microarray analysis, growth factor withdrawal and the existing understanding of gene regulation in early embryo and primitive endoderm; a simplified depiction of the genetic network active in primitive endoderm precursor-like cells is shown in Figure 32. NANOG and GATA6 are at the heart of this network and these transcription factors presumably control the fate transition between the pluripotent ICM and primitive endoderm *in-vivo*, and the pluripotent ESCs and primitive endoderm precursor-like cells, such as MAPCs and XEN-P, *in-vitro*. The expression of NANOG seems to be controlled by multiple signaling factors, most notably, the suppression by GATA6. Over-expression of GATA6 in mouse ESCs has been shown to suppress NANOG and lead to differentiation of cells to the primitive endoderm [131]. Through our microarray analysis we also discovered another potential suppressor of NANOG in MAPCs and MAX8, the nuclear receptor NR6A1 (GCNF). GCNF is a transcriptional repressor that suppresses gene expression by binding to response elements in the promoter region of genes and plays a critical role in retinoic acid-induced differentiation of mouse ESCs by down regulating OCT4 and NANOG [119, 236]. GNCF is more than two fold upregulated in rat MAPCs and MAX8 with respect to the ESCs (Figure 33).

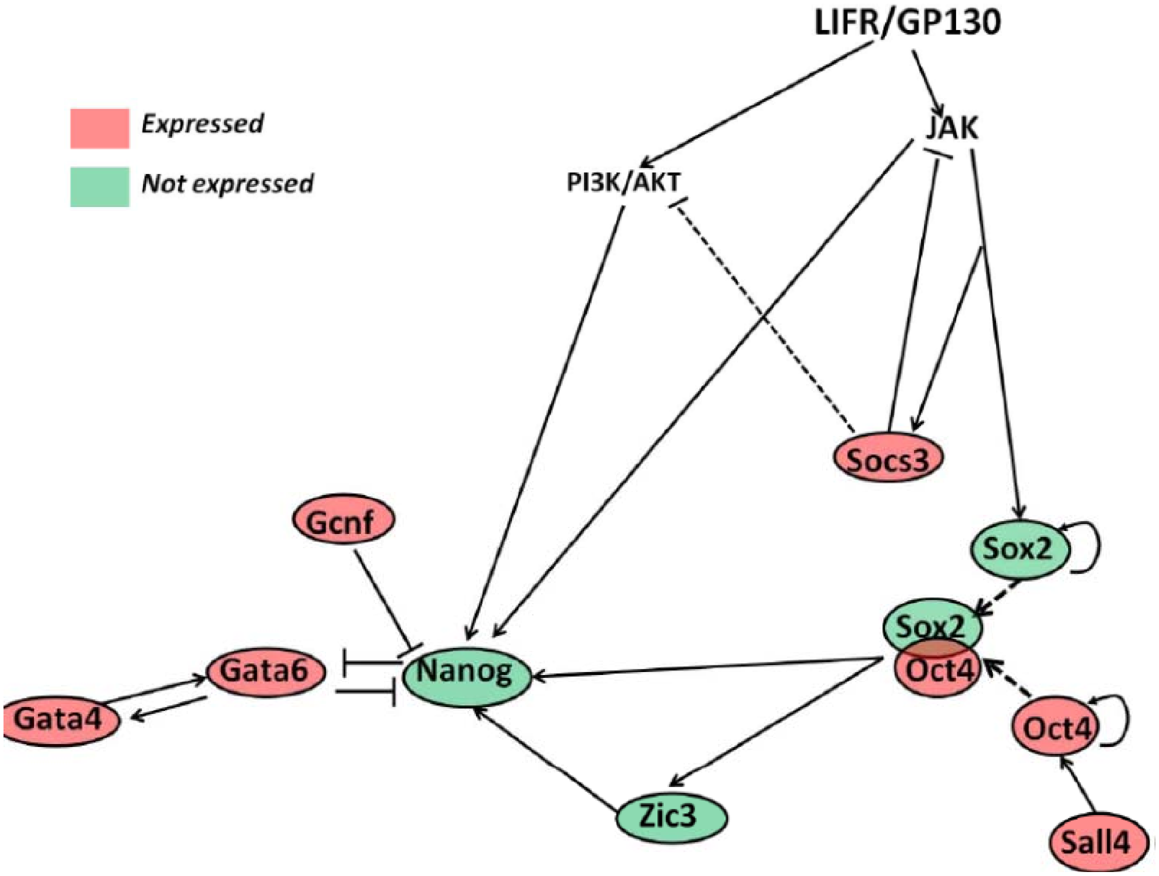


Figure 32.A simplified depiction of the transcriptional network active in rat primitive endoderm precursor-like cells (MAPCs and MAX8)

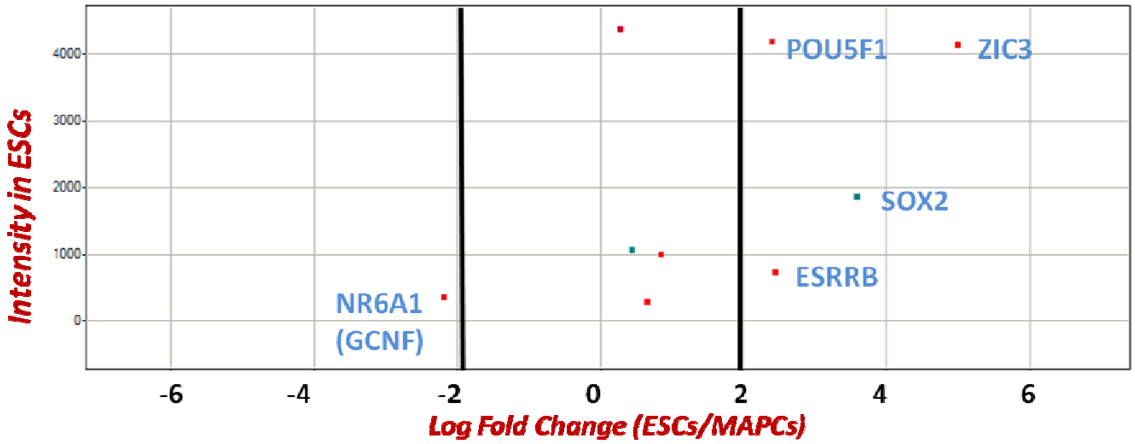
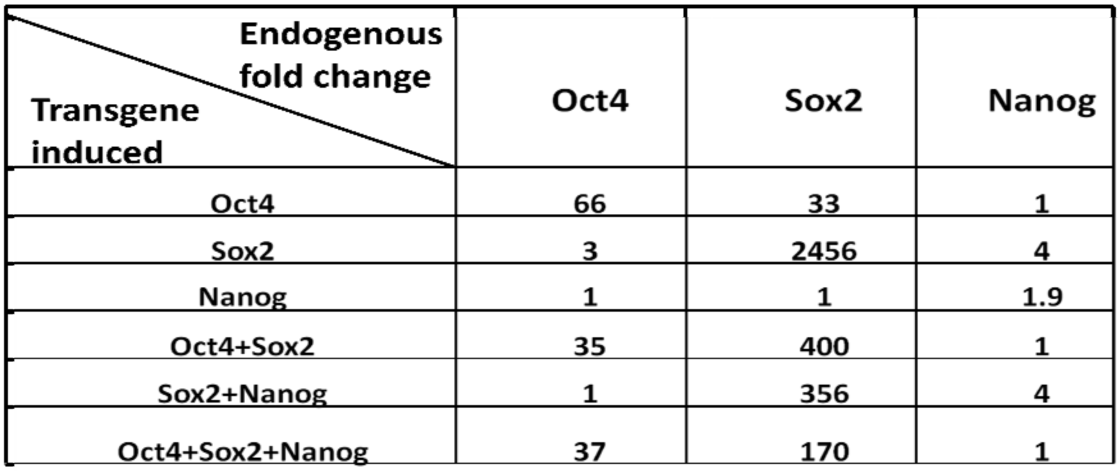


Figure 33. Expression of activators and suppressors of Nanog in rat ESCs (ESCs) and rat bone marrow-derived primitive endoderm precursor-like cells (MAPCs)

7.5 Inducing exogenous expression of pluripotency genes in MAPCs

While these primitive endoderm precursor cells express potential suppressors of NANOG, such as GATA6 and GNCF, it can be argued that failure of these cells to express genes that can activate NANOG, such as SOX2 and ZIC3, could lead to lack of expression of this ‘gatekeeper’ of pluripotency. To look at this, we forced expression of pluripotency factors such as SOX2 and even exogenous NANOG to determine if endogenous expression can be induced. Rat MAPCs were infected with lentiviruses expressing human OCT4, SOX2 and NANOG in different combinations, a stable pool of cells expressing the exogenous gene(s) was selected using puromycin and the endogenous expression was analyzed on day 5 post infection. Figure 34 shows the change in expression of endogenous OCT4, SOX2 and NANOG due to expression of exogenous factors. Consistent with the transcriptional network shown in figure 33, when MAPCs are induced with high levels of exogenous OCT4, both endogenous OCT4 and

SOX2 expression levels increase. It is known that OCT4 binds to its own promoter and self-activates. Thus, an increase in OCT4 protein would cause endogenous gene to be expressed at higher levels. This, in turn, would cause some increase in the endogenous SOX2 because OCT4 also binds to SOX2 promoter region.



Endogenous fold change Transgene induced	Oct4	Sox2	Nanog
Oct4	66	33	1
Sox2	3	2456	4
Nanog	1	1	1.9
Oct4+Sox2	35	400	1
Sox2+Nanog	1	356	4
Oct4+Sox2+Nanog	37	170	1

Figure 34. Change in expression of endogenous genes upon exogenous induction of OCT4, SOX2 and NANOG in rat MAPCs, day 5 post infection with lentiviruses

Similarly, it was observed that inducing exogenous SOX2 causes a very steep increase in expression of endogenous gene. This can be explained by the formation of OCT4-SOX2 dimer that binds to the promoter region of SOX2 and activates its transcription. Since MAPCs already express OCT4, the induction of exogenous SOX2 would cause rapid formation of the dimer that not only activates the endogenous SOX2, but also increases the endogenous OCT4 levels due to binding of the dimer to OCT4 promoter region. In all the combinations shown above, no significant increase in endogenous NANOG was observed. NANOG does not activate itself and thus, exogenous NANOG fails to induce endogenous gene expression. The most striking observation was that even with very high endogenous levels on OCT4 and SOX2 (comparable to expression in rat ESCs), no induction of endogenous NANOG could be seen. NANOG promoter has been shown to have binding sites for the OCT4-SOX2 dimer and the lack of NANOG induction above suggests that its expression in MAPCs could be controlled by other factors too, such as primitive endoderm genes like GATA6/4. In this case, the presence of activating OCT4-SOX2 dimer would not be sufficient to activate endogenous gene expression.

To look at slightly longer term effects of exogenous factor induction, the experiment was repeated and endogenous gene expression was analyzed at day 13 post induction. Surprisingly, an increase in expression of endogenous Nanog was observed in MAPCs induced with exogenous OCT4 or exogenous OCT4 and SOX2 (Figure 35). This delayed increase in Nanog expression could indicate possible alleviation of suppression by another factor or promoter demethylation and activation. In any case, it is clear that Nanog expression is carefully and strictly regulated in MAPCs by both activating and inhibiting factors.

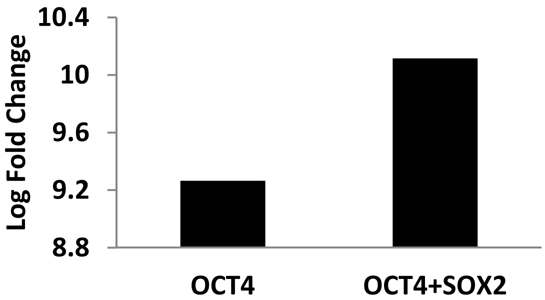


Figure 35. Increase in expression of endogenous Nanog in MAPCs upon sustained expression of transgenes on day 13 post induction

The effect of induction of exogenous pluripotency factors Oct4 and Sox2 can also be seen on other genes not expressed by MAPCs, such as Zic3 (Figure 36). Zic3 has been shown to be a regulator of Nanog expression in pluripotent stem cells and its expression shows a remarkable increase on day 13 post induction.

Taken together, these results suggest that an increase in expression of Oct4, combined with a sustained induction of Sox2, can induce expression of pluripotency genes not expressed by MAPCs, such as Nanog and Zic3. This increase in pluripotency factor expression however is not accompanied by suppression of primitive endoderm as no decrease in expression of Gata6, Gata4 and Sox17 could be observed on day 13 in these cells.

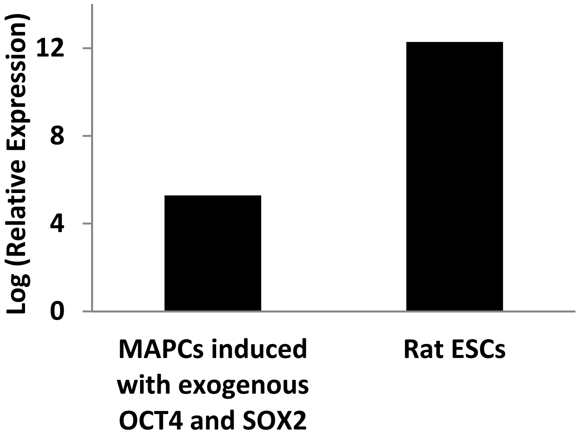


Figure 36. Increase in Zic3 expression in MAPCs induced with exogenous Oct4 and Sox2, relative to uninduced cells

7.6 Nanog promoter methylation in MAPCs

As mentioned above, one possible cause of delayed increase in Nanog expression in MAPCs could be the gradual alleviation of suppressive epigenetic effects. Epigenetic processes such as modulation of chromatin structure and DNA methylation play a pivotal role in gene expression. DNA methylation, catalyzed by enzymes called DNA methyltransferases or DNMTs, is the chemical addition of methyl groups on cytosines in CpG dinucleotides. Genes having CpGs in their promoter regions have been shown to show an inverse correlation between gene expression and dinucleotide methylation. This effect has been shown to be important in development, as well as, cancer progression [237, 238]

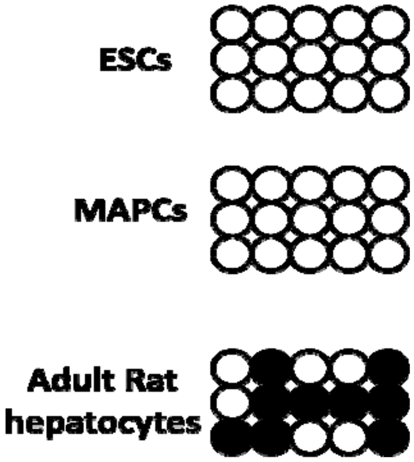


Figure 37. DNA methylation analysis of Nanog promoter region, 213 bps upstream of 5' UTR, in rat ESCs, bone marrow-derived MAPCs and adult rat hepatocytes

In ESCs, the promoter regions of important pluripotency-associated genes, such as, Oct4 and Nanog, are predominantly demethylated. Upon spontaneous or induced differentiation, the promoters of these genes undergo methylation, thereby correlating the methylation status with the expression [239, 240]. The promoters of these genes have also been shown to undergo demethylation upon reprogramming to induced pluripotent stem cell state. Thus, DNA methylation is an active and important process in controlling gene expression in pluripotent stem cells.

To determine if the Nanog expression in MAPCs is also influenced DNA methylation, bisulfate sequencing was used to determine the methylation status of CG dinucleotides in the promoter region upstream of the 5' UTR. As shown in Figure 37, the analyzed promoter region was found to be completely unmethylated in these bone marrow-derived adult stem cells. Rat ESCs were used as a positive control and showed an identical methylation pattern. Similarly, adult rat hepatocytes not expressing Nanog, showed extensive methylation in this promoter region.

This shows that the Nanog expression in these adult, primitive endoderm precursor-like cells in not actively suppressed by epigenetic DNA methylation. The gene expression is likely to be under the control of positive regulatory factors such as Oct4, Sox2, Zic3, as well as, negative regulatory factors such as Gata6 and Gata4. At the same time, other epigenetic and genetic factors such as histone modifications and miRNA control could also play a role in modulating expression of Nanog and other pluripotency genes in these cells.

7.7 Effect of GATA6 knockdown on primitive endoderm

Gata6 is an important marker of primitive endoderm and is one of the earliest lineage markers during segregation of ICM into primitive endoderm and epiblast [134]. Nanog, a marker for pluripotency, continues to be expressed in the epiblast and becomes mutually exclusive in expression with Gata6. There has been increasing evidence suggesting that these two genes mutually suppress each other, either directly or through intermediate regulatory factors. In mouse ESCs, sustained expression of Nanog prevents them from differentiating to primitive endoderm while forced expression of Gata6 causes Nanog suppression and differentiation to primitive endoderm [115, 131]. MAPCs are bone marrow-derived stem cells that have a primitive endoderm precursor-like gene expression signature and express very high levels of primitive endoderm markers like Gata6 and Gata4. These cells express Oct4 at reasonably high levels and Sox2 at basal but Nanog expression is completely absent in these cells. Also, as discussed earlier, while Sox2 expression can be quickly induced through elevated Oct4 levels, Nanog levels only increase modestly upon sustained expression of Oct4 and Sox2. To determine is Gata6 is a putative suppressor of Nanog in these cells, doxycycline inducible suppression of Gata6 was performed in MAPCs. The pTRIPZ vector is an inducible lentiviral vector system that allows doxycycline-dependent expression of shRNA of interest (Figure 38). The tet response element (TRE) also drives the expression of turbo RFP, along with the shRNA, giving a visual indication of the extent of expression. The lentiviral plasmid was packaged using the second generation packaging plasmid system.

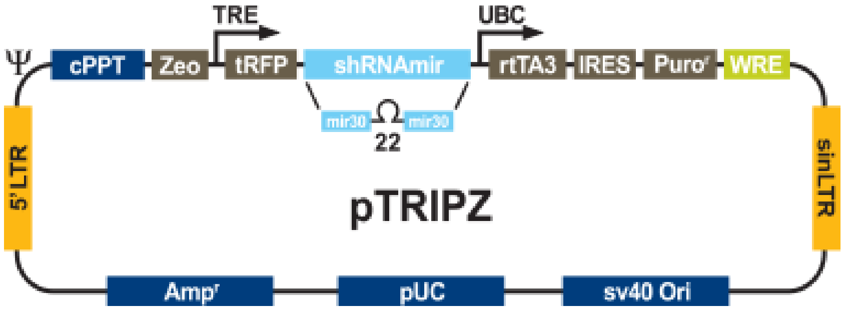


Figure 38. Schematic of the pTRIPZ lentiviral inducible shRNAmir vector (www.openbiosystem.com)

A stable pool of cells was selected where the lentivirus had integrated and the cells were induced using 1000ng/ml. Doxycycline (Sigma-Aldrich). Following induction, RFP expression could be seen early as 24 hours. An increase in RFP was observed over the next 2 days after which it stabilized.

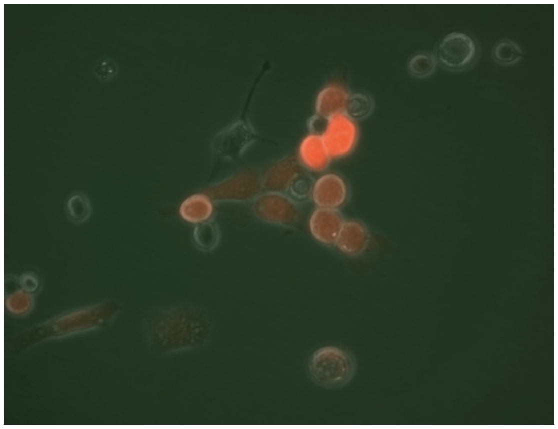


Figure 39. RFP expression in doxycycline induced MAPCs expressing shRNA against GATA6 on day 6 post induction

A decrease in Gata6 expression could be seen as early as day 2 using quantitative real time PCR. The suppression effect of shRNA increased until day 5 beyond which it stabilized. This decrease in Gata6 expression matched with a temporal increase in Nanog expression (Figure 40) too. The highest expression of Nanog was achieved on day 6 at which point, more than a fourfold increase could be seen in the expression.

As shown in figure 39, different cells in the selected stable pool, express varied levels of RFP. The extent of induction and expression of RFP/shRNA would depend upon the site of integration in the genome and the number of copies of the lentiviral plasmid that integrated. So, cells expressing highest level of GFP expression were isolated from the bulk culture and their Nanog expression was analyzed. This was done to confirm the increase in Nanog as a result of Gata6 suppression by showing that cells exhibiting highest levels RFP also show higher Nanog induction that what can be observed in the bulk culture. As shown in figure 41, the RFP expression correlated well with the extent of knockdown observed with the cells expressing high levels of RFP showing higher knockdown. Nanog expression was then compared between the highest and lowest Gata6 expressing samples, namely the RFP negative and the top 2% RFP positive cells. A higher than 16 fold increase in Nanog expression was seen in the latter.

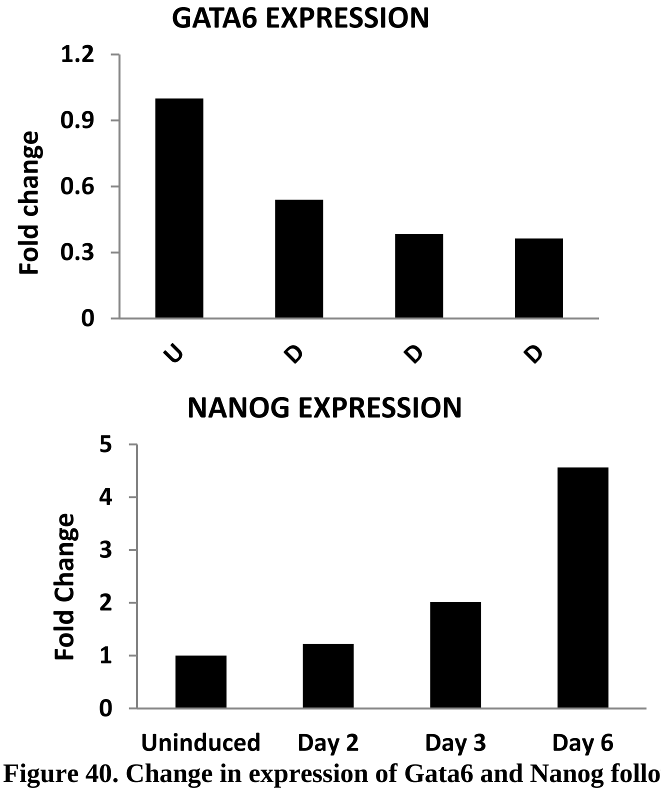


Figure 40. Change in expression of Gata6 and Nanog following induction of shRNA targeting Gata6

These results show that Nanog expression in MAPCs is suppressed, directly or indirectly, by Gata6 and the extent of suppression depends on the level of Gata6 expression.

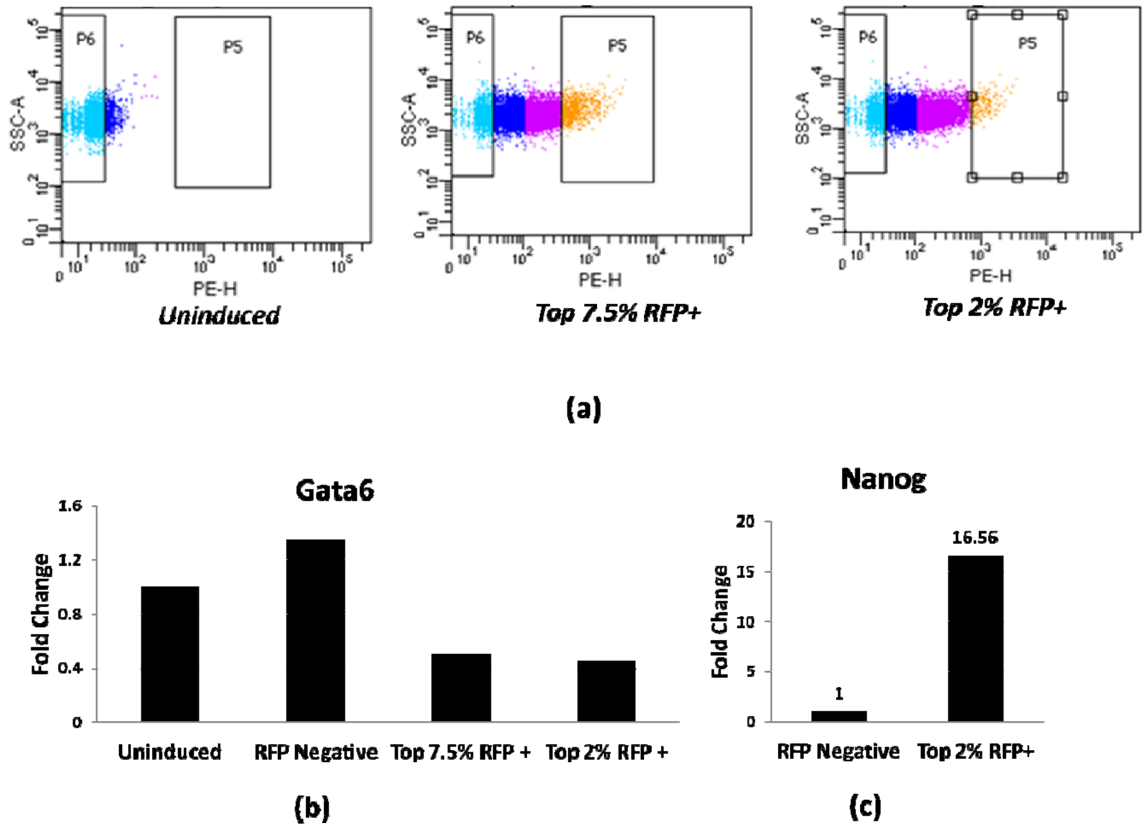


Figure 41 (a) RFP expression distribution in doxycycline induced MAPCs. Four populations were collected using FACS-uninduced cells, RFP negative induced cells and, top 7.5% and 2% RFP positive cells. (b) Gata6 expression in uninduced, RFP negative-induced cells and top 7.5% and 2% RFP positive cels. (c) Nanog expression in top 2% RFP positive cells in comparison to the RFP negative populations shows a more than 16 fold induction.

7.8 Discussion

To better understand the unique properties of primitive endoderm precursor-like cells such as MAPCs and MAX8, a model for the transcriptional network active in these cells was constructed and the elements of this network were explored through multiple approaches. The culture medium for initial derivation and maintenance of MAPCs consists of two important growth factors, Leukemia Inhibitory factor (LIF) and Platelet-derived Growth factor (PDGF). The roles of these growth factors were explored by culturing cells in medium devoid of one growth at a time and observing the changes in growth and gene expression. Similar to its role in maintenance of mouse embryonic stem cells, LIF was determined to be critical for the growth, as well as, Oct4 expression. In the absence of PDGF though, the cell growth slowed down but the cells could still be maintained for many passages and no significant effect on expression of Oct4 or Sox2 could be determined. These results, in combination with analysis of gene expression data from microarrays and, published literature, were used to construct a gene regulatory network active in MAPCs. Through induction of gene expression using exogenous factors, the inter-relationships between Oct4, Sox2 and Nanog were explored. It was observed that endogenous Sox2 could be easily induced through increase in endogenous Oct4, as well as, the expression of exogenous Sox2. Nanog, on the other hand, could only be induced upon sustained expression of exogenous Oct4 and Sox2. Even with this, the induction was not as high as what was observed for endogenous Sox2. To further explore this aspect, the relationship between Nanog and its potential suppressor Gata6 was explored. It was determined that suppression of endogenous Gata6 using shRNAs could indeed induce expression of endogenous Nanog.

While some important aspects of the proposed gene network in MAPCs (Figure 32) have been explored, other interactions that have been proposed based on gene expression data or literature still need to be explored in the context of these cells. One such interaction is between GCNF and its potential suppressors Oct4 and Nanog in MAPCs. Another important phenomenon is the role of high levels of Socs3 in these cells and how this might correlate with expression of Oct4, Sox2 and Nanog. Finally, while the suppressive effect of Gata6 on Nanog has been proved, any effect in the reverse direction, i.e., Nanog suppressing Gata6 or other primitive endoderm genes is an important phenomenon that can be explored further.

Chapter 8
Reprogramming primitive endoderm precursor like cells to an ICM-like state

8.1 Introduction

Rat bone marrow-derived MAPCs have a unique gene expression signature in that they express genes associated with pluripotent stem cells, such as Oct4 and Rex1, while simultaneously expressing markers representative of the primitive endoderm lineage such as Gata6 and Sox17. Nanog is an important pluripotency marker that is absent in MAPCs and, as discussed in the previous chapters, its expression is tightly controlled by activators such as Oct4 and Sox2, as well as potential suppressors, such as Gata6. The MAPCs are very similar in gene expression, as well as morphology and surface antigen expression, to Oct4 expressing primitive endoderm precursor cells derived from E4.5 rat blastocyst. As mentioned in Chapter 2, the primitive endoderm is formed when the ICM differentiates and segregates into the epiblast and the primitive endoderm. Precursor cells to both of these populations have been identified in vivo. Such precursor cells exist transiently in a primed state having acquired increased expression of markers of one lineage and partially suppressed markers of the other. Primitive endoderm precursor cells have very low expression levels of Nanog and high levels of Gata6. Since, MAPCs represent these cells so closely, it is interesting to determine if the Nanog-Gata6 switch can be reversed in these cells to change their fate to an ICM-like cell. Upon sustained activation of Oct4 and Sox2 in MAPCs, induction of Nanog was observed but this did not lead to a fate reversal as no change in primitive endoderm gene expression was observed. This induction was done in MAPC culture conditions, though, and it is possible that alternate culture conditions supporting ICM could have allowed the cells to slowly suppress primitive endoderm.

Ever since the discovery of induced pluripotent stem cells in 2006, various cell types have been reprogrammed to ESC-like fate. Somatic cells from all the three germ lineages, endoderm, mesoderm and ectoderm, have been reprogrammed using exogenous expression of Oct4, Sox2, Klf4 and cMyc in ESC culture conditions. While these somatic cells are terminally differentiated and require extensive epigenetic and genetic changes to change their fate, non-pluripotent stem cells have also been reprogrammed using this technique. For example, neural stem cells (NSCs) expressing Sox2, Klf4 and cMyc were reprogrammed to iPSC state through exogenous induction of just Oct4 [241]. Trophoblast stem cells representing the extraembryonic trophectoderm lineage express key trophectoderm marker Cdx2 but none of the pluripotency genes [135]. These stem cells have also been reprogrammed to the ICM-like state through induced expression of the four factors [242]. There have not been any reports of reprogramming of the primitive endoderm lineage to the ICM-like state yet. MAPCs, with their primitive endoderm precursor-like phenotype, serve as an excellent model system to determine if cells primed to become extraembryonic endoderm can be reverted to an ICM state and thereby differentiated to embryonic lineages.

8.2 Reprogramming MAPCs using traditional four factor approach

The traditional reprogramming approach, as demonstrated in the initial report by the Yamanaka lab, includes infecting the starting cell type with four viruses each expressing one of the four factors, Oct4, Sox2, Klf4 and c-Myc [48, 53]. After a few days post infection, medium is changed to that of ESCs. A similar approach was taken to study reprogramming of primitive endoderm-like MAPCs to ICM-like iPSCs. Also, in order to monitor the expression of endogenous Sox2 in MAPCs, a reporter line was constructed by infecting the cells with a lentivirus that expresses GFP under the control of sox2 enhancer elements [243]. Hotta et al. developed a vector system containing GFP under the control of early transposon promoter (active in ESCs), combined with Sox2 and Oct4 binding motifs in ES cell-specific enhancers (Addgene plasmid 21317). The vector was modified to include a constitutive selection element imparting resistance to hygromycin. A stable pool of MAPCs was thus created that respond to endogenous expression of both Oct4 and Sox2 by turning on the reporter expression.

Li et al. had successfully reprogrammed rat liver progenitor cell line WB-F344 using mouse ES culture conditions where the mESC culture medium was supplemented with inhibitors of MEK, GSK3 and ALK5 signaling [244, 245]. The authors reported that while they were able to obtain colonies of iPSCs without the use of inhibitors (in mESC medium only), the colonies could not be sustained for long term. Similarly, Chang et al. reprogrammed rat neural precursor cells and rat embryonic fibroblasts by using 3 factors (Oct4, Sox2 and Klf4). The reprogrammed cells were maintained on a layer of rat embryonic fibroblasts and in medium similar to mESC medium but supplemented with inhibitors of MEK and GSK3 signaling [246]. Based on these observations, R19 cell line was infected with retroviruses expressing the four factors. Since, MAPCs express high enough levels of Oct4 and Klf4; we also tried a two-factor combination including Nanog and Sox2. In addition, due to the inverse nature of relationship between Nanog and Gata6 in early differentiation, as well as, reports suggesting that Nanog overexpression can hinder differentiation to primitive endoderm [130], we also tried a single factor reprogramming experiment with Nanog. All experiments were performed in mESC medium with and without three inhibitors (inhibitors of MEK, GSK3 and ALK5).

Figure 42 shows the morphology of cells one week for infection. Small colonies could be seen in culture that had a rather heterogeneous morphology. Staining with Alkaline phosphatase, an early marker of reprogramming showed that the colonies stain non-uniformly with AP, confirming the heterogeneous state of cells in a colony. Most colonies stained positive for endogenous Nanog expression week one post induction. This included cells that were infected with four factors but not Nanog.

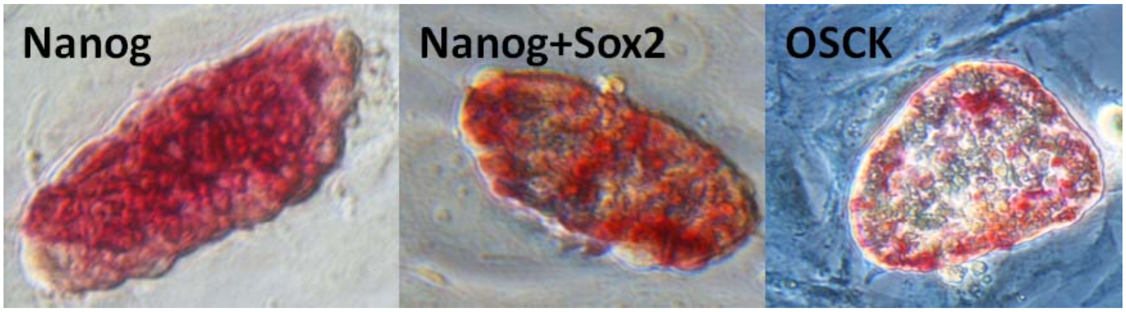


Figure 42. Alkaline phosphatase staining day 6 post infection in R19 MAPCs induced with exogenous Nanog, Nanog and Sox2 and OSCK

GFP expression could be seen in all the three cultures indicating possible endogenous expression, particularly in Nanog only cells that do not contain any exogenous Sox2.

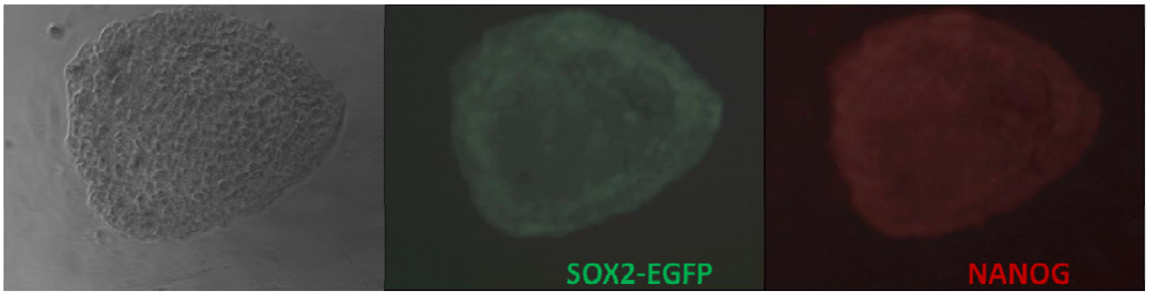


Figure 43. Sox2 enhancer driven EGFP expression and immunostain for Nanog protein in a colony

Despite this limited and early signs of some reprogramming, all cultures became progressively heterogeneous over time. More importantly, the untransformed cells (that may have not have gotten infected with all or any factors) continued growing at a fast pace. Despite the very low initial plating density, these cells always overgrew in culture and the small, partially reprogrammed colonies could not compete with these cells. Attempts at manual passaging of the small colonies also did not work. The colonies, presumably due to incomplete reprogramming, rapidly differentiated in culture.

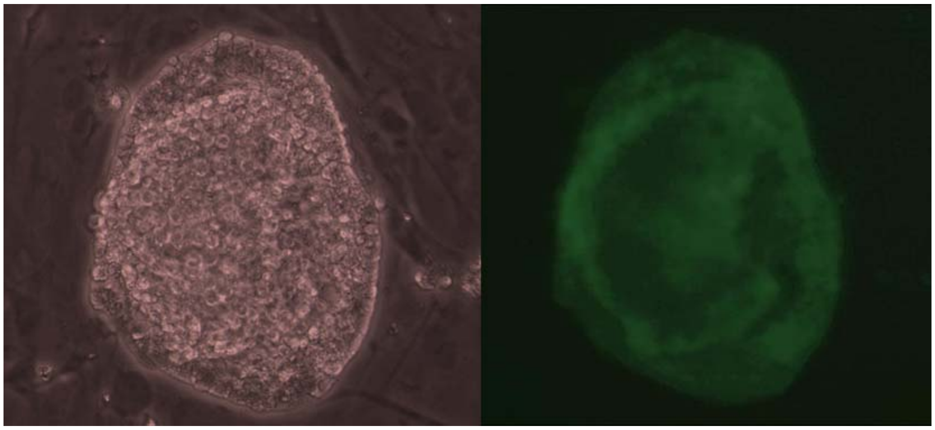


Figure 44. Rapid differentiation and concomitant EGFP loss in the center of a manually passaged colony

8.3 Reprogramming MAPCs using polycistronic vector expressing four factors

TETO-FUW-OSKM is a lentiviral vector with four factors under the control of a tetracycline operator minimal promoter [247]. The four genes are joined by 2A sequences allowing tetracycline inducible expression of all four genes from the same promoter. The polycistronic lentivrus was packaged using a second generation packaging system along with a FUW vector containing tetracycline controllable transactivator. MAPCs were infected with the polycistronic vector first and a stable pool of cells containing integrated copies of the vector was selected using selection with Zeocin. This was followed by another infection with the transactivator lentivirus followed by induction with doxycycline (Figure 45).

The cells were then induced with doxycycline and transferred to rat ESC culture medium on mouse embryonic fibroblasts. The first colony was observed around 10 days post dox induction. Figure 46 shows the morphology change and progression of a colony in culture.

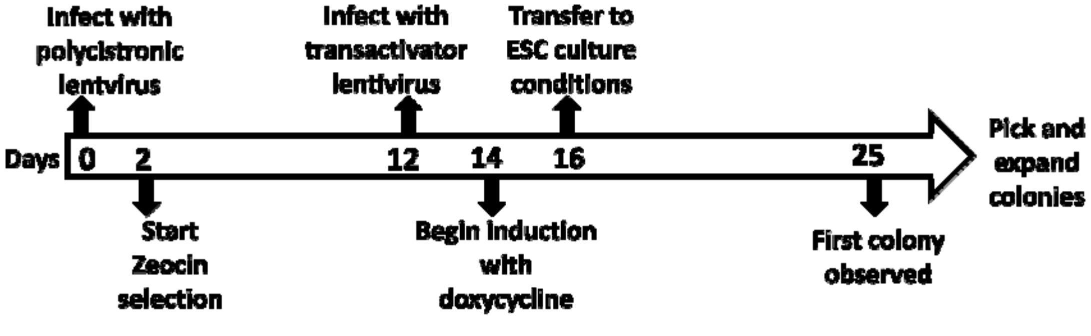
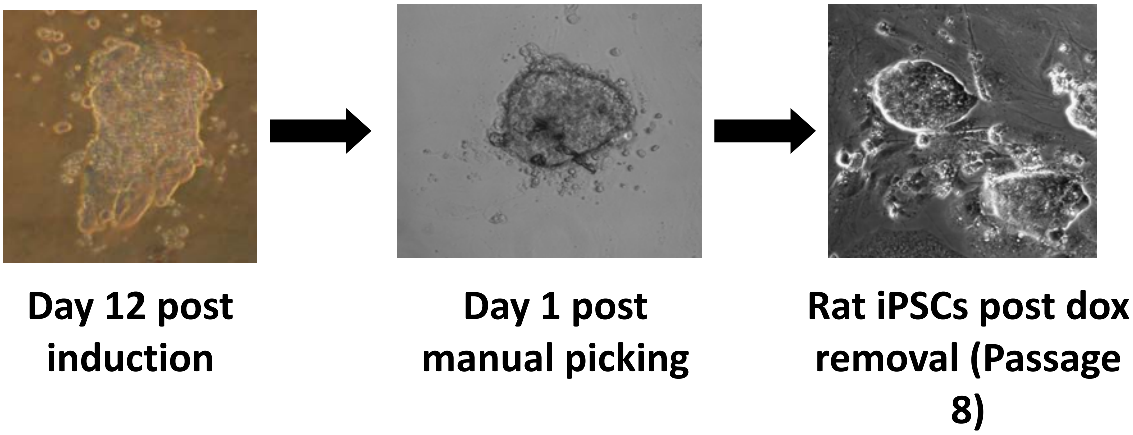

Infect with polycistronic lentvirus
Infect with transactivator lentivirus
Transfer to ESC culture conditions
Days
0
2
12
14
16
25
Pick and expand colonies
Start Zeocin selection
Begin induction with doxycycline
First colony observed

Figure 45. Timeline for reprogramming rat MAPCs using tetracycline inducible polycistronic vector system



Day 12 post induction
Day 1 post manual picking
Rat iPSCs post dox removal (Passage 8)

Figure 46. Progression of a single rat MAPCs-derived iPSC colony in culture

8.2 Gene Expression in rat MAPC-derived iPSCs

The endogenous gene expression of MAPC-derived iPSC clone SS1 was assessed using quantitative real time PCR, as shown in figure 47. At passage 4 post dox withdrawal, the clone SS1 expressed high levels of Nanog and Sox2. The Oct4 expression level was similar to the expression seen in MAPCs and was somewhat lower than the expression seen in rESCs. The expression of Gata6, Gata4 and Sox17 was very low, similar to the expression level seen in rat ESCs. This shows that the clone SS1 has acquired a pluripotent gene signature by turning on endogenous Sox2 and Nanog and suppressing the primitive endoderm. Quantitative RT-PCR conducted at later passages showed a further increase in endogenous Oct4 while the Sox2 and Nanog expression were maintained (data not shown).

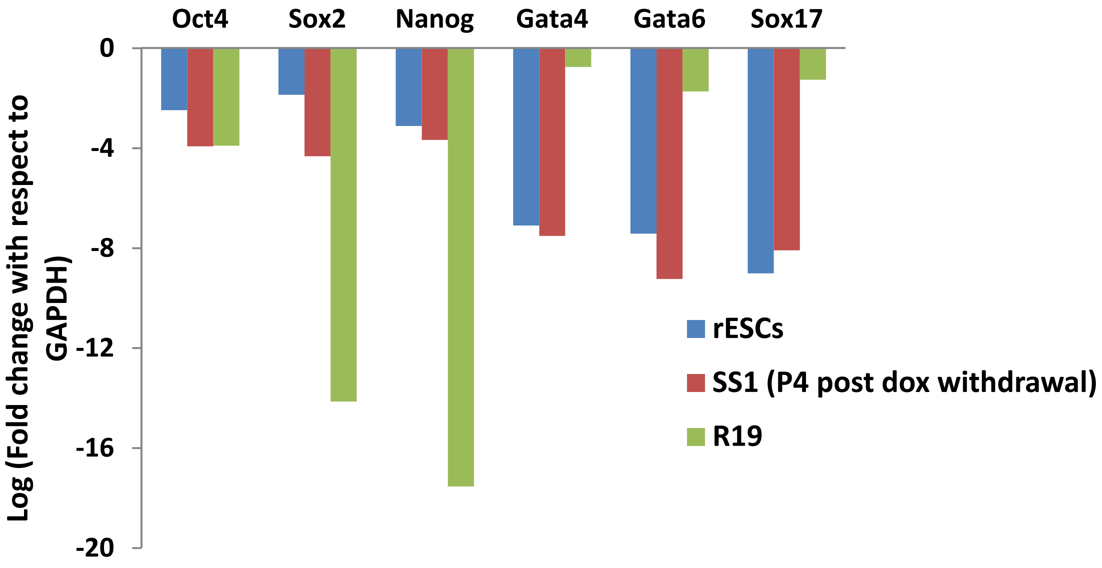


Figure 47. Expression of key pluripotency and primitive endoderm genes in rat ESCs (rESCs), rat MAPC-derived iPSC clone SS1 and R19 rat MAPCs.

8.3 MAPC-derived rat iPSCs express Nanog

Intracellular staining for endogenous expression of Nanog protein showed that the cells express high amounts of the protein and the staining was consistent with a nuclear stain (Figure 47). Nanog, being a transcription factor, stains inside the nucleus of pluripotent cells. Nanog, an important pluripotency marker, shows an inverse relationship in expression with Gata6, an early primitive endoderm marker, in early embryo. The suppression of Nanog, along with the concomitant increase in Gata6, defines the primitive endoderm precursor population in the ICM. The expression of endogenous Nanog in reprogrammed MAPCs, in addition to loss of Gata6 expression, confirms the switch from primitive endoderm to pluripotent induced pluripotent stem cells.

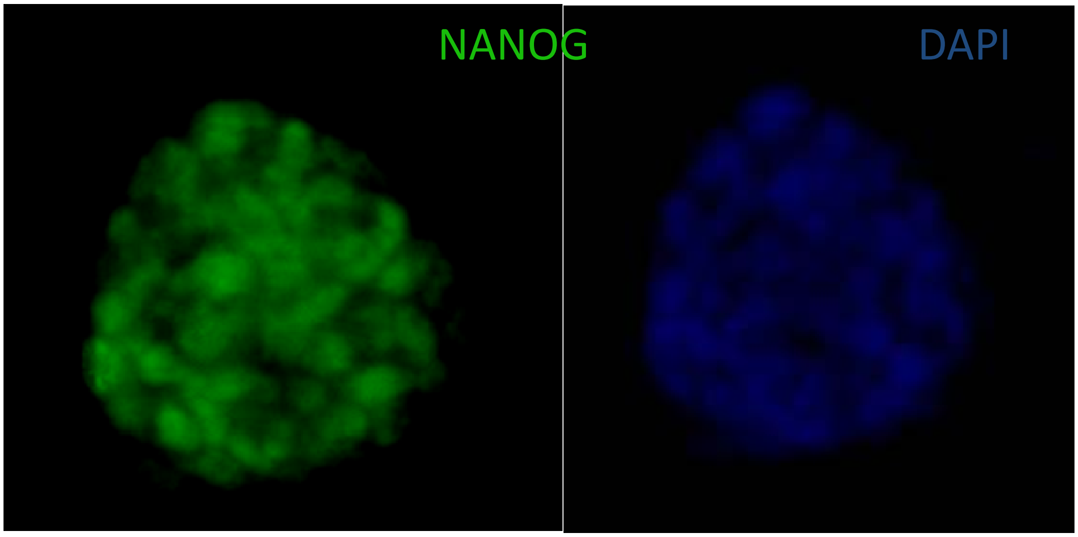


Figure 48. Immunostain showing nuclear staining for Nanog in a rat MAPC-derived iPS colony

8.4 Comparing transcriptome of Rat MAPC-derived iPSCs to rat ESCs

While quantitative RT-PCR data showed the expression of important pluripotency genes such as Sox2 and Nanog in SS1 cells, microarray analyses was performed to assess how the similarity in the global gene expression profile of SS1 with rESCs, as well as, the starting cell type of R19 MAPCs. Figure 49 shows the hierarchical clustering of these three cell types. Rat blastocyst-derived, hypoblast-like MAX8 cells were also included in the analyses. As expected, the SSE1 riPSCs are very similar in global gene expression to rat ESCs. The expression of SS1 cells was quite different from the starting cell types of MAPCs, which in turn show remarkable similarity in expression to blastocyst-derived MAX8 cells.

Figure 50 shows the expression of key pluripotency-associated genes in the four cell types. Both rESCs and SS1 rIPSCs express high levels of these key genes which are either absent or expressed at very low levels in both primitive endoderm-like MAPCs and MAX8 cells. Socs3 is the only pluripotency related gene that is expressed at higher levels in the primitive endoderm-precursor like cells. As discussed in Chapter 7, this increase in expression over rESCs could contribute to the suppression of Sox2 and Nanog in these cells since Socs3, at higher than physiologic amount, has been shown to suppress LIF and JAK/STAT3 signaling [248]. Similarly, Figure 51 shows the expression of genes associated to the extraembryonic (primitive) endoderm lineage in rat ESCS, SS1 rIPSCs, MAPCs and MAX8 cells. MAPCs and MAX8 express high levels of Gata4, Gata6, Pdgfra, Sox17 and Sparc. These characteristic markers are essentially absent in rESCs, as well as, MAPC-derived SS1 rIPSCs.

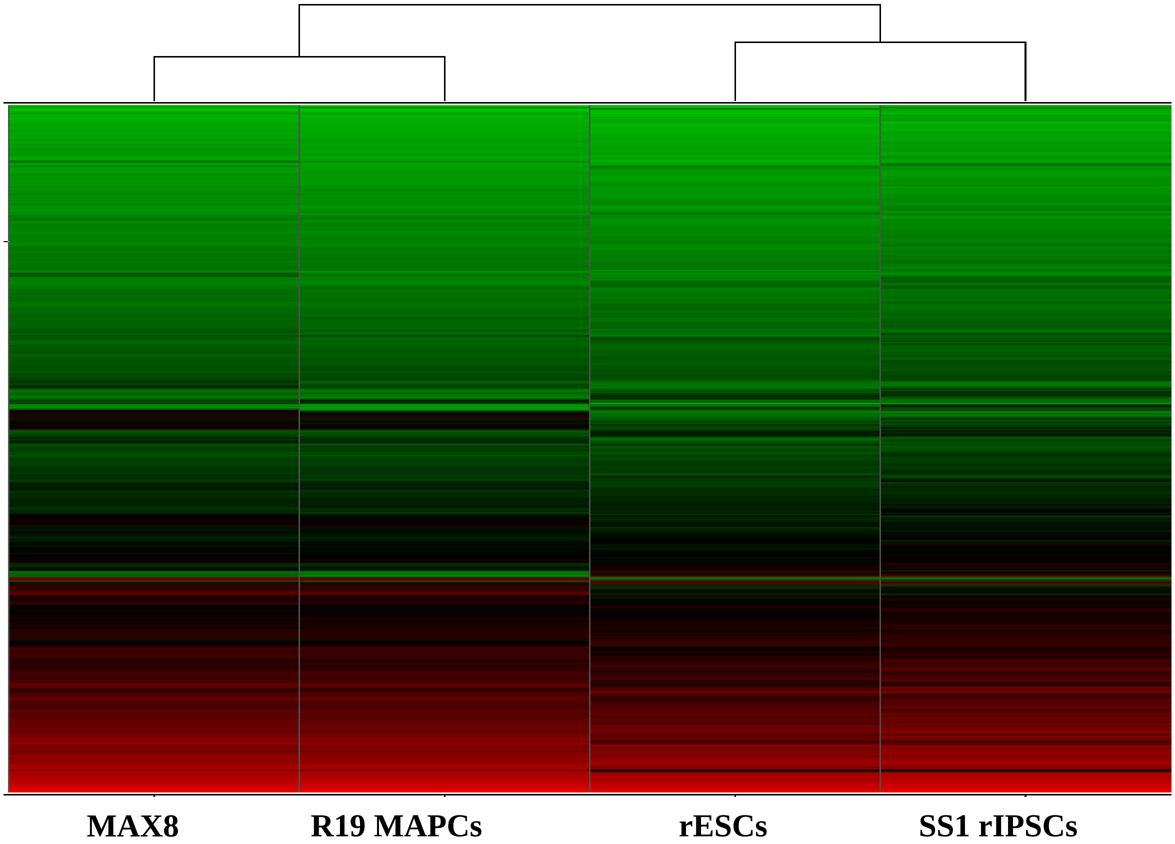


Figure 49. Hierarchical clustering of microarray data from rat blastocyst-derived hypoblast-like cells Max8, adult bone marrow-derived R19 MAPCs, rat ESCs and MAPC-derived rat iPS cell line SS1.

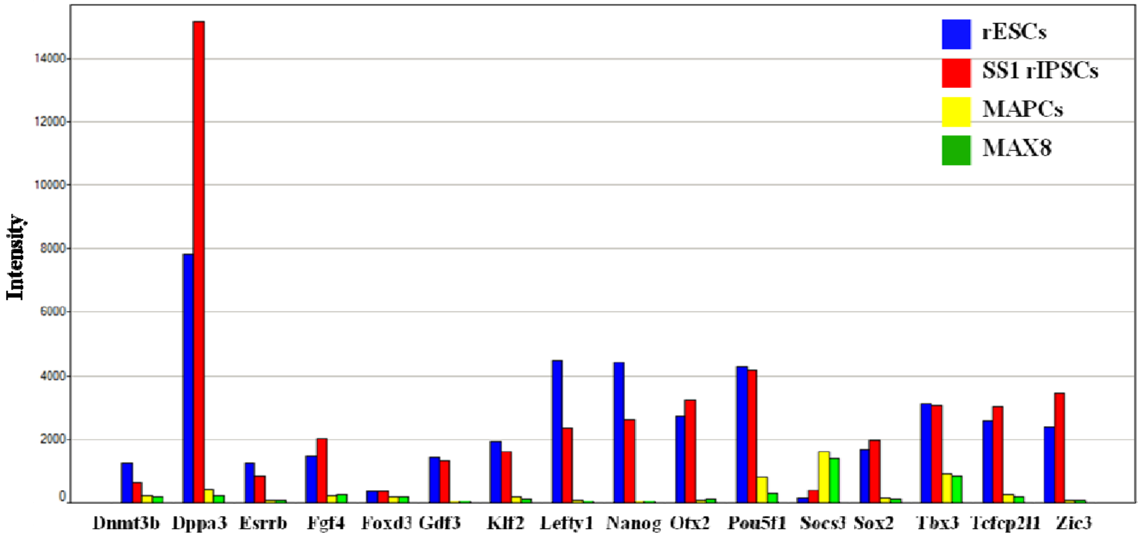


Figure 50. Expression of key Pluripotency-associated genes in rat ESCs (rESCs), SS1 rat induced pluripotent stem cells (rIPSCs), bone marrow-derived MAPCs and rat blastocyst-derived hypoblast-like MAX8 cells

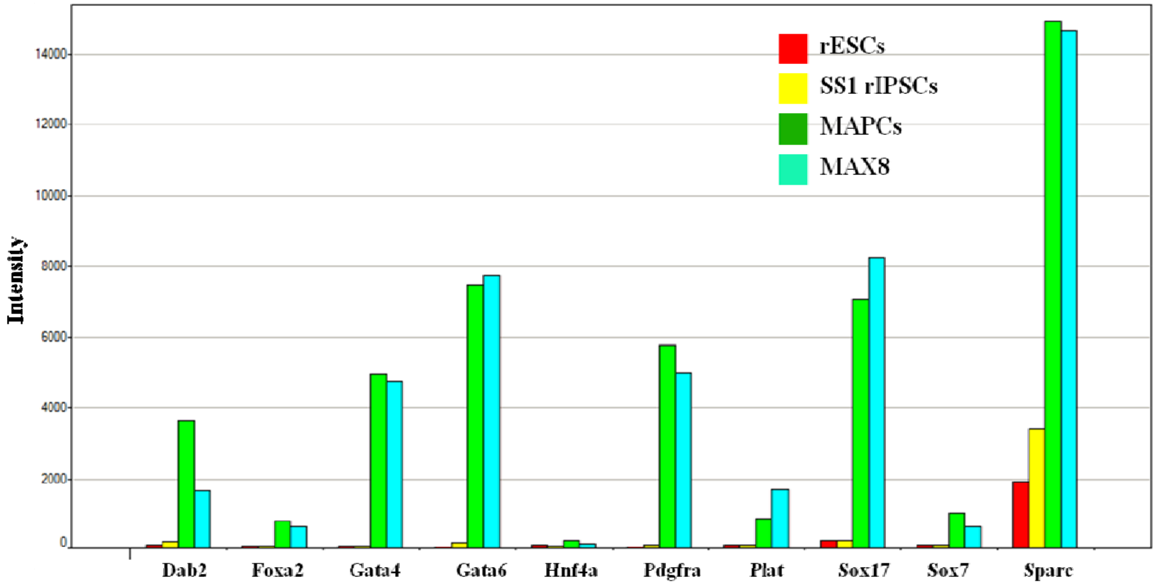


Figure 51. Expression of extraembryonic endoderm lineage-associated genes in rat ESCs (rESCs), SS1 rat induced pluripotent stem cells (rIPSCs), bone marrow-derived MAPC and rat blastocyst-derived hypoblast-like MAX8 cells

8.5 Shift in expression of Pluripotency-associated Metagene in SS1 rIPSCs

In Chapter 4, gene expression data from multiple studies involving pluripotent stem cells was used to discover a class of genes that was representative of the pluripotent stem cells and differentiated them from other samples representing partially or fully differentiated or reprogrammed cell types. This class of genes was discovered using a dimensionality reduction tool called non-negative matrix factorization (NMF) and, in the parlance of NMF, is referred as the 'metagene' representing pluripotent stem cells. This metagene, as shown in Appendix 1, consisted of most known pluripotency-related genes such as Nanog, Sox2, Zic3, Tdgf1, amongst others, but also novel genes such as Bex1, Eras and Igfbp2, whose roles in maintenance of pluripotency has not been discovered yet. Since, such a clustering technique identifies an unbiased and statistically significant group of genes; this group can be used to assess the extent of reprogramming in terms of expression change in a subset of genes that are truly representative of pluripotent, fully reprogrammed cells. An orthologous set in rat corresponding to the top 100 genes in the metagene identified from the mouse dataset was determined. Figure 52 shows the expression of this orthologous set (85 genes in total) in R19 MAPCs, with respect to rat ESCs. As indicated by the solid lines which represent a twofold variation in expression in either direction, about half of the genes in this metagene have reasonably similar expression (less than two fold variation) between rESCs and rMAPCs. This is because MAPCs do express high levels of many genes associated with stem cells, particularly those involved in growth, such as Klf4 and nMyc. But, as shown below, MAPCs show reduced to absent expression of a large number of genes in this set too. This is expected because while MAPCs have been shown to be multipotent in their differentiation capacity, they are certainly distinct from pluripotent ESCs and IPSCs, as discussed in previous chapters.

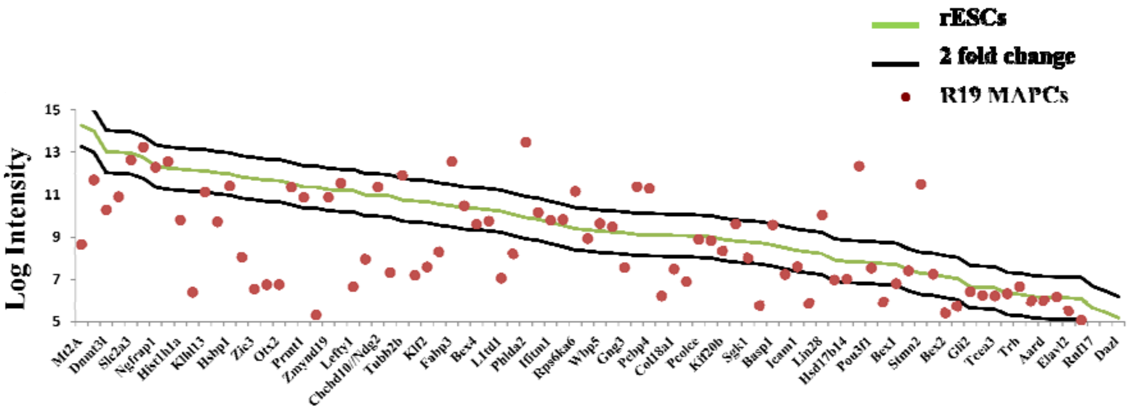


Figure 52. Expression of pluripotency-associated metagene cluster in R19 MAPCs with respect to pluripotent rESCs. The solid black lines indicate a twofold change margin

Figure 53 shows the similar representation for MAPC-derived SS1 rIPSCs. The figure clearly shows how the expression of genes in the pluripotency-associated metagene group has changed in SS1 cells, in comparison to the parent MAPCs. SS1 cells are remarkably similar to rESCs in expression of these genes. Majority of genes lie within the two fold margin and out of the ones that lie outside this margin, majority actually show higher expression in SS1 than rESCs.

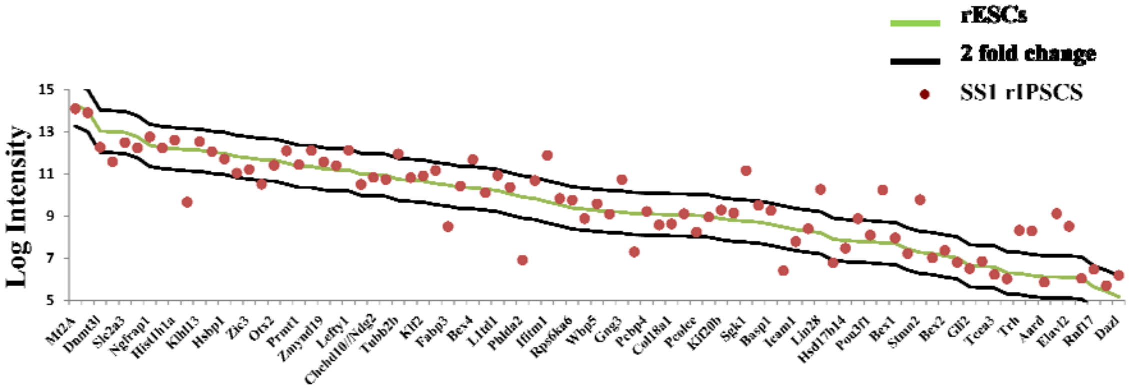


Figure 53. Expression of pluripotency-associated metagene cluster in SS1 rIPSCs with respect to pluripotent rESCs. The solid black lines indicate a twofold margin

8.6 Discussion

The primitive endoderm lineage arises from the pluripotent cells of the ICM. The other group of cells in the ICM (that do not form the primitive endoderm) give rise to the pluripotent epiblast. Recent research has shown that the morphologically indistinguishable cells of the ICM do not have identical gene expression profile [4]. They can actually be divided into two groups on the basis of their transcriptome- the epiblast precursor and the primitive endoderm precursor. Another significant finding of this report was that the primitive endoderm precursor population still had high expression of the pluripotency marker Oct4 although the Nanog expression had decreased to low levels and the cells had turned on the expression of key primitive endoderm marker Gata6. The epiblast precursor, on the other hand, maintained high Nanog (while maintain their Oct4 too) and showed little to no expression of Gata6. Other studies have also shown the presence of Oct4 and Gata6 expressing cells in the ICM [34, 134]. Just how do subsets of cells which are still expressing the master regulator gene Oct4, turn off Nanog and become primitive endoderm while their close cousins maintain pluripotency is an intriguing question. The answer to this question, when discovered, will provide very important cues to the establishment and maintenance of pluripotency in early embryo, as well as, in pluripotent stem cells.

Verfaillie lab first reported the isolation of multipotent cells from the bone marrow of an adult rat that expressed Oct4 and Gata6 but no Nanog or Sox2. These cells, termed MAPCs, were called multipotent for their ability to differentiate to cell types of multiple lineages. Recently, strikingly similar cells have been isolated from the rat embryo [3]. Verfaillie lab has also recently shown isolation of MAPC-like cells from the rat ICM using rat MAPC culture conditions. The transcriptome of MAPCs has been compared to pluripotent rat ESCs and embryo-derived MAPC-like cells. MAPCs are remarkably similar, if not identical, to embryo derived primitive endoderm-precursor cells. The gene expression is very similar to that of the primitive endoderm-precursor cells in the ICM identified by Kurimoto et al. This analysis confirms that MAPCs express key markers of both the primitive endoderm and the embryonic epiblast lineage. As mentioned above, the transcriptional switch that that results into the separation of primitive endoderm from the pluripotent epiblast is crucial to a better understanding of how pluripotency is maintained in stem cells. The transcriptional network in MAPCs was analyzed by observing the effect of knockdown of key primitive endoderm gene Gata6, as well as, by probing the effect of growth factor signaling in the cells. It was discovered that there is a direct effect of Gata6 suppression on expression of Nanog. Also, the dynamics of the gene network were probed by inducing exogenous expression of some key genes like

Sox2 and Nanog and observing the effect on endogenous expression. Endogenous Nanog could be induced in MAPCs through sustained expression of Oct4 and Sox2.

Whether these primitive endoderm precursor cells can be switched back to the pluripotent ICM-like state is an important question too. iPSC studies usually show reprogramming of a somatic cell of an embryonic lineage to pluripotent, ESC-like state. Using the traditional reprogramming cocktail, MAPCs could be reprogrammed to an ESC-like state. While through our approach, it has been successfully demonstrated that even cell resembling an extrembryonic lineage can be reprogrammed to an ICM-like state, this discovery has raised other interesting questions that should be addressed in future studies. For example, the traditional reprogramming cocktail consisting of Oct4, Sox2, Klf4 and c-Myc was used in this study but it has been mentioned earlier that MAPCs express very high levels of Klf4 and n-Myc. The latter has been shown to be able to replace c-Myc in the four factor cocktail. This suggests that it is likely that all the four factors are not mandatory for this reprogramming. In addition, MAPCs also express Oct4, albeit about 4 fold lower than rat ESCs. While this might lead one to think that Oct4 may also be redundant in reprogramming, careful analysis need to be performed to arrive at a definite result. It is worthwhile to point out here that the experiments performed in chapter 7 using exogenous induction of Oct4 and Sox2, seemed to suggest that induction of only Sox2 does not lead to Nanog induction on day 13. To see an increase in endogenous Nanog, either Oct4 alone or in combination with Sox2 needed to be induced. This again reinforces the need to understand the significance of four fold lower expression of Oct4 in these cells in comparison to rat ESCs.

Chapter 9
Conclusion and Future Directions

Stem cells have captured the interest and imagination of scientific community for many decades. Not only do these cells offer fascinating clinical opportunities, a lot stands to be gained from developing a better understanding of the gene expression that defines these cells and is responsible for keeping them in a 'naïve' state. The research discussed in this thesis is aimed at discovering a transcriptional signature for mouse and human embryonic and induced pluripotent stem cells. It further compares the transcriptome of the pluripotent stem cells with that of a unique multipotent stem cell called MAPCs. The uniqueness of these bone marrow-derived stem cells lies in their striking similarity with the nascent hypoblast in embryonic development.

9.1 Meta-analysis of pluripotent stem cell data

High throughput transcriptome data captures genome wide variations in gene expression and is an extremely useful source for developing insights into gene regulation. A meta-analysis combining data from multiple studies offers advantages of an unbiased analysis which allows identification of a global, not 'lab-specific' signature. Public databases such as Gene Expression Omnibus (GEO) are useful repositories for data collection. A meta-analysis which processes large scale data from studies involving myriad experimental conditions can also enable identification of underlying transcriptional networks through analysis of gene coregulation. By combining data from multiple different studies involving human and mouse pluripotent cells, an expression signature for these cells was discovered. Non-negative matrix factorization (NMF) is an extremely useful dimensionality reduction technique that allowed detection of natural trends and classes in the data. It also allowed identification of the genes that most significantly differentiate them from other cell types in the meta-analysis. In addition to this, a ranking order for these genes was identified which provides an estimate of their relative importance in classifying pluripotent stem cells from other cell types such as differentiation or reprogramming samples. Using a reverse engineering tool called ARACNe, potentially important signaling centers or 'hubs' were also predicted.

Another application of this gene expression analysis is to combine the results of this analysis with other types of genetic and genomic data such as physical interaction data. This approach can be extremely useful for predicting novel, high-confidence interactions in pluripotent stem cells, as well as, developing hypotheses based on the results of this approach, as discussed in next section.

9.2 Functional Interaction Maps allow identification of conserved subnetworks

Functional interaction maps for human and mouse genomes integrate diverse genetic and genomic data from many sources. This includes data involving protein-protein physical interaction, phenotype/disease, phylogenetic profiles etc. We developed an algorithm to overlay gene expression data from our meta-analysis onto these functional maps to discover important transcriptional modules. More importantly, this approach could be used to discover conserved functional interactions between the human and mouse pluripotent stem cells. These human and mouse pluripotent stem cells, though similar in many, are not identical and possibly represent slightly different stages in mammalian development.

Chapters 4 and 5 were committed to uncovering a gene expression signature representative of pluripotent stem cells in both human and mouse, and further extracting more useful information from this analysis by combining it with known physical interaction data. While this represents a global transcriptome level analysis of pluripotent stem cells, it does not answer specific questions about the regulation of their 'core' pluripotency gene network. The last three chapters of this thesis are dedicated to understanding the latter through the use of a unique cell type as a model system. These special primitive endoderm precursor-like cells seem to be the in-vitro counterparts of a cell type very closely associated to the pluripotent inner cell mass in the embryo. The analysis of transcriptional regulation in these cells (expressing many key pluripotency genes) helped answer many important questions about gene regulation in pluripotent stem cells *in vitro*, as well as, *in vivo*.

9.3 Primitive Endoderm Precursor-like cells from the bone marrow and the embryo

The Verfaillie lab first reported the derivation of multipotent stem cells (MAPCs) from the rat bone marrow that showed significant potential to differentiate to multiple cell types of the three germ layers. These cells bear a striking similarity with rat embryo-derived cells that represent nascent hypoblast lineage. Using the culture conditions for MAPCs, the Verfaillie lab has been able to derive similar cells from the rat embryo.

MAPCs, as well as these rat embryo-derived hypoblastic cells, have a very interesting gene expression signature in that they express genes characteristic of pluripotent stem cells, such as Oct4 and Zfp42 (Rex1), while also expressing markers of hypoblast (primitive endoderm) lineage such as Gata6 and Pdfgra. Through single cell cDNA amplification analysis of the ICM, existence of highly similar cells has been shown in the ICM of the mouse embryo. These significant developments have made such primitive endoderm precursor-like cells a very important tool for understanding the regulation of pluripotency in early embryonic development. Mouse embryonic and induced pluripotent stem cells are believed to represent the cells of the ICM that give rise to the epiblast and embryonic germ lineages. The primitive endoderm precursor-like cells, on the other hand, are similar to those cells in the ICM that have been predisposed to give rise to the extraembryonic endoderm lineage. Due to this unique profile of the MAPCs, we have explored the gene expression profile of these cells in order to add to our understanding of the pluripotent gene network. This system has allowed us to look more closely at the regulation of an important pluripotency marker Nanog, and how it's down regulation can cause pluripotent stem cells to differentiate to the extraembryonic lineage, despite continued expression of Oct4.

A transcriptional network model for MAPCs was developed based upon experimental results probing growth factors requirements, microarray data analysis and available literature. Components of this network were further explored through overexpression and knockdown approaches. The effect of endogenous Oct4 and Sox2 on Nanog was determined. Through promoter methylation, the non-methylated state of Nanog promoter was also verified. Finally, the suppression of Nanog by primitive endoderm gene Gata6 was proven. While some interactions in the gene regulatory network were tested and confirmed in this work, more aspects of the network remain to be verified. One such aspect was to determine if MAPC gene expression can be switched to a pluripotent stem cell-like state through one or more genes in the network.

9.4 Reprogramming MAPCs to ICM-like state

Through induction of the four transcription factors, Oct4, Sox2, Klf4 and c-Myc, it was shown that these extraembryonic endoderm-like cells can in fact be reprogrammed to an ICM-like state. A protocol was developed to induce the reprogramming cocktail in MAPCs by using a single polycistronic vector expressing all the four genes under the control of a single promoter. Successful reprogramming could be achieved because the protocol allows for pre-selection of a stable pool of cells with the transgene expressing 4 factors integrated into the genome. This stable pool was doxycycline inducible. Microarray analysis confirmed that the gene expression profile of MAPCs can be switched completely to become highly similar to pluripotent rat ESCs using this approach. We also showed that using the pluripotent gene expression signature that was derived from the meta-analysis in the first part of this thesis, a clear transition in expression pattern can be seen as the cells get reprogrammed. This simultaneously confirms the validity of our pluripotency signature in another species (rat), as well as, the reprogramming of MAPCs to a true pluripotent state.

9.5 Future directions

The meta-analysis approach towards deciphering the gene expression signature of pluripotent stem cells can provide useful information, which when combined with other sources of genomic and genetic information such as functional interaction maps, can be used to ask important biological questions. One such question addressed in this thesis is that of conserved pockets of gene expression called subnetworks that are at play in pluripotent stem cells of two species, human and mouse. These insights can be further explored experimentally to confirm their relevance to the phenotype. Also, tools such as ARACNe are able to detect underlying coregulation and predict *de novo* interactions. This is a unique way to uncover new interactions and those interactions that are predicted with a high degree of confidence should be explored experimentally.

The differentiation of cells of the ICM into pluripotent epiblast and primitive endoderm is being widely studied in order to understand this second important differentiation event during embryonic development. Many elegant studies probing this phenomenon have suggested a genetic switch between pluripotency marker Nanog and primitive endoderm marker Gata6. As discussed extensively in this thesis, bone marrow derived MAPCs (and embryo derived MAX8) seem to represent the *in vitro* counterparts of the nascent hypoblast *in vivo*. Rat ESCs similarly appear to be the in vitro counterpart of the ICM (or at least those cells of the ICM that are predisposed towards forming the epiblast versus the primitive endoderm). In this thesis, the transcriptome of the rat ESCs has been compared with that of the MAPCs. A transcriptional network has been developed based off of this comparison, along with cues from experiments involving growth factor requirements for MAPCs. One component of this network, i.e., the Nanog-Gata6 relationship has been explored in this thesis. There is plenty of scope to explore other important elements of this transcriptional circuit, such as the effect of a very high level of Socs3 expression in MAPCs and MAX8 cells. Socs3, or suppressor of cytokine signaling, is an important signaling component of the Jak/Stat3 signaling. Similarly, the overexpression studies have suggested that transient overexpression of pluripotency genes such as Oct4 and Sox2 is able to turn on the expression of endogenous Nanog and Zic3. The long term effects of such an overexpression remain to be studies. It has been shown that the induction of exogenous four factors, OSKM, is able to change the fate of MAPCs to pluripotent iPSCs. This experiment has opened the doors for further experimentation that elucidates which of these four factors are critical and which are redundant. MAPCs express high levels of Klf4, n-Myc and Oct4 (albeit at a lower level than rESCs). This suggests that it might be possible to eliminate at least two of the four factors. Such an analysis will also shed more light onto the transcriptional circuit mentioned above. Since this transcriptional circuit is very intricately tied to the transcriptional machinery responsible for the pluripotency of ESCs, this will further help understand the core gene network that lies at the heart of pluripotency.

Bibliography

1. Jiang Y, Jahagirdar BN, Reinhardt RL, Schwartz RE, Keene CD, Ortiz-Gonzalez XR, Reyes M, Lenvik T, Lund T, Blackstad M *et al*: **Pluripotency of mesenchymal stem cells derived from adult marrow**. *Nature* 2002, **418**(6893):41-49.
2. Subramanian K, Gerearts M, Pauwelyn KA, Park Y, Owens DJ, Muijtjens M, Ulloa-Montoya F, Jiang Y, Verfaillie CM, Hu WS: **Isolation procedure and characterization of multipotent adult progenitor cells from rat bone marrow**. *Methods in molecular biology (Clifton, NJ* 2010, **636**:55-78.
3. Debeb BG, Galat V, Epple-Farmer J, Iannaccone S, Woodward WA, Bader M, Iannaccone P, Binas B: **Isolation of Oct4-expressing extraembryonic endoderm precursor cell lines**. *PloS one* 2009, **4**(9):e7216.
4. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, Ueda HR, Saitou M: **An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis**. *Nucleic acids research* 2006, **34**(5):e42.
5. Gilbert S: **Early Mammalian Development**. In. Edited by Developmental Biology te: Sinauer Associates; 2000.
6. Murray L, Chen B, Galy A, Chen S, Tushinski R, Uchida N, Negrin R, Tricot G, Jagannath S, Vesole D *et al*: **Enrichment of human hematopoietic stem cell activity in the CD34+Thy-1+Lin- subpopulation from mobilized peripheral blood**. *Blood* 1995, **85**(2):368-378.
7. Pierelli L, Scambia G, Bonanno G, Rutella S, Puggioni P, Battaglia A, Mozzetti S, Marone M, Menichella G, Rumi C *et al*: **CD34+/CD105+ cells are enriched in primitive circulating progenitors residing in the G0 phase of the cell cycle and contain all bone marrow and cord blood CD34+/CD38low/- precursors**. *British journal of haematology* 2000, **108**(3):610-620.
8. Friedenstein AJ, Chailakhyan RK, Latsinik NV, Panasyuk AF, Keiliss-Borok IV: **Stromal cells responsible for transferring the microenvironment of the hemopoietic tissues. Cloning in vitro and retransplantation in vivo**. *Transplantation* 1974, **17**(4):331-340.
9. da Silva Meirelles L, Chagastelles PC, Nardi NB: **Mesenchymal stem cells reside in virtually all post-natal organs and tissues**. *Journal of cell science* 2006, **119**(Pt 11):2204-2213.
10. Zuk PA, Zhu M, Mizuno H, Huang J, Futrell JW, Katz AJ, Benhaim P, Lorenz HP, Hedrick MH: **Multilineage cells from human adipose tissue: implications for cell-based therapies**. *Tissue engineering* 2001, **7**(2):211-228.
11. Petersen BE, Bowen WC, Patrene KD, Mars WM, Sullivan AK, Murase N, Boggs SS, Greenberger JS, Goff JP: **Bone marrow as a potential source of hepatic oval cells**. *Science (New York, NY* 1999, **284**(5417):1168-1170.

12. Pittenger MF, Mackay AM, Beck SC, Jaiswal RK, Douglas R, Mosca JD, Moorman MA, Simonetti DW, Craig S, Marshak DR: **Multilineage potential of adult human mesenchymal stem cells**. *Science (New York, NY* 1999, **284**(5411):143-147.
13. Kopen GC, Prockop DJ, Phinney DG: **Marrow stromal cells migrate throughout forebrain and cerebellum, and they differentiate into astrocytes after injection into neonatal mouse brains**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(19):10711-10716.
14. Sharma S, Raju R, Sui S, Hu WS: **Stem cell culture engineering - process scale up and beyond**. *Biotechnology journal* 2011.
15. Goldman SA: **Disease targets and strategies for the therapeutic modulation of endogenous neural stem and progenitor cells**. *Clinical pharmacology and therapeutics* 2007, **82**(4):453-460.
16. Benraiss A, Chmielnicki E, Lerner K, Roh D, Goldman SA: **Adenoviral brain-derived neurotrophic factor induces both neostriatal and olfactory neuronal recruitment from endogenous progenitor cells in the adult forebrain**. *J Neurosci* 2001, **21**(17):6718-6731.
17. Curtis MA, Penney EB, Pearson AG, van Roon-Mom WM, Butterworth NJ, Dragunow M, Connor B, Faull RL: **Increased cell proliferation and neurogenesis in the adult human Huntington's disease brain**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(15):9023-9027.
18. Evans MJ, Kaufman MH: **Establishment in culture of pluripotential cells from mouse embryos**. *Nature* 1981, **292**(5819):154-156.
19. Smith AG, Hooper ML: **Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells**. *Developmental biology* 1987, **121**(1):1-9.
20. Smith AG, Heath JK, Donaldson DD, Wong GG, Moreau J, Stahl M, Rogers D: **Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides**. *Nature* 1988, **336**(6200):688-690.
21. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM: **Embryonic stem cell lines derived from human blastocysts**. *Science (New York, NY* 1998, **282**(5391):1145-1147.
22. Ying QL, Nichols J, Chambers I, Smith A: **BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3**. *Cell* 2003, **115**(3):281-292.
23. Buehr M, Nichols J, Stenhouse F, Mountford P, Greenhalgh CJ, Kantachuvesiri S, Brooker G, Mullins J, Smith AG: **Rapid loss of Oct-4 and pluripotency in cultured rodent blastocysts and derivative cell lines**. *Biology of reproduction* 2003, **68**(1):222-229.
24. Vassilieva S, Guan K, Pich U, Wobus AM: **Establishment of SSEA-1- and Oct-4-expressing rat embryonic stem-like cell lines and effects of cytokines of the IL-6 family on clonal growth**. *Experimental cell research* 2000, **258**(2):361-373.

25. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A: **The ground state of embryonic stem cell self-renewal**. *Nature* 2008, **453**(7194):519-523.

26. Buehr M, Meek S, Blair K, Yang J, Ure J, Silva J, McLay R, Hall J, Ying QL, Smith A: **Capture of authentic embryonic stem cells from rat blastocysts**. *Cell* 2008, **135**(7):1287-1298.

27. Li P, Tong C, Mehrian-Shai R, Jia L, Wu N, Yan Y, Maxson RE, Schulze EN, Song H, Hsieh CL *et al*: **Germline competent embryonic stem cells derived from rat blastocysts**. *Cell* 2008, **135**(7):1299-1310.

28. Nichols J, Smith A, Buehr M: **Rat and mouse epiblasts differ in their capacity to generate extraembryonic endoderm**. *Reproduction, fertility, and development* 1998, **10**(7-8):517-525.

29. Damjanov I, Sell S: **Yolk sac carcinoma grown from rat egg cylinders**. *Journal of the National Cancer Institute* 1977, **58**(5):1523-1525.

30. Brons IG, Smithers LE, Trotter MW, Rugg-Gunn P, Sun B, Chuva de Sousa Lopes SM, Howlett SK, Clarkson A, Ahrlund-Richter L, Pedersen RA *et al*: **Derivation of pluripotent epiblast stem cells from mammalian embryos**. *Nature* 2007, **448**(7150):191-195.

31. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, Gardner RL, McKay RD: **New cell lines from mouse epiblast share defining features with human embryonic stem cells**. *Nature* 2007, **448**(7150):196-199.

32. Greber B, Wu G, Bernemann C, Joo JY, Han DW, Ko K, Tapia N, Sabour D, Sterneckert J, Tesar P *et al*: **Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells**. *Cell stem cell* 2010, **6**(3):215-226.

33. Nichols J, Smith A: **Naive and primed pluripotent states**. *Cell stem cell* 2009, **4**(6):487-492.

34. Rossant J: **Stem cells and early lineage development**. *Cell* 2008, **132**(4):527-531.

35. Hanna J, Markoulaki S, Mitalipova M, Cheng AW, Cassady JP, Staerk J, Carey BW, Lengner CJ, Foreman R, Love J *et al*: **Metastable pluripotent states in NOD-mouse-derived ESCs**. *Cell stem cell* 2009, **4**(6):513-524.

36. Guo G, Smith A: **A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency**. *Development (Cambridge, England)* 2010, **137**(19):3185-3192.

37. Gu P, Goodwin B, Chung AC, Xu X, Wheeler DA, Price RR, Galardi C, Peng L, Latour AM, Koller BH *et al*: **Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development**. *Molecular and cellular biology* 2005, **25**(9):3492-3505.

38. ten Berge D, Kurek D, Blauwkamp T, Koole W, Maas A, Eroglu E, Siu RK, Nusse R: **Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells**. *Nature cell biology* 2011, **13**(9):1070-1075.

39. Briggs R, King TJ: **Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs**. *Proceedings of the National Academy of Sciences of the United States of America* 1952, **38**(5):455-463.

40. King TJ, Briggs R: **Changes in the Nuclei of Differentiating Gastrula Cells, as Demonstrated by Nuclear Transplantation**. *Proceedings of the National Academy of Sciences of the United States of America* 1955, **41**(5):321-325.

41. Gurdon JB, Laskey RA, Reeves OR: **The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs**. *Journal of embryology and experimental morphology* 1975, **34**(1):93-112.

42. Campbell KH, McWhir J, Ritchie WA, Wilmut I: **Sheep cloned by nuclear transfer from a cultured cell line**. *Nature* 1996, **380**(6569):64-66.

43. Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KH: **Viable offspring derived from fetal and adult mammalian cells**. *Nature* 1997, **385**(6619):810-813.

44. Wilmut I, Beaujean N, de Sousa PA, Dinnyes A, King TJ, Paterson LA, Wells DN, Young LE: **Somatic cell nuclear transfer**. *Nature* 2002, **419**(6907):583-586.

45. Wakayama T, Perry AC, Zuccotti M, Johnson KR, Yanagimachi R: **Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei**. *Nature* 1998, **394**(6691):369-374.

46. Lanza RP, Cibelli JB, Faber D, Sweeney RW, Henderson B, Nevala W, West MD, Wettstein PJ: **Cloned cattle can be healthy and normal**. *Science (New York, NY* 2001, **294**(5548):1893-1894.

47. De Sousa PA, Dobrinsky JR, Zhu J, Archibald AL, Ainslie A, Bosma W, Bowering J, Bracken J, Ferrier PM, Fletcher J *et al*: **Somatic cell nuclear transfer in the pig: control of pronuclear formation and integration with improved methods for activation and maintenance of pregnancy**. *Biology of reproduction* 2002, **66**(3):642-650.

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

49. Xie H, Ye M, Feng R, Graf T: **Stepwise reprogramming of B cells into macrophages**. *Cell* 2004, **117**(5):663-676.

50. Hanna J, Wernig M, Markoulaki S, Sun CW, Meissner A, Cassady JP, Beard C, Brambrink T, Wu LC, Townes TM *et al*: **Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin**. *Science (New York, NY* 2007, **318**(5858):1920-1923.

51. Meissner A, Wernig M, Jaenisch R: **Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells**. *Nature biotechnology* 2007, **25**(10):1177-1181.

52. Okita K, Ichisaka T, Yamanaka S: **Generation of germline-competent induced pluripotent stem cells**. *Nature* 2007, **448**(7151):313-317.

53. 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):861-872.

54. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R: **In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state**. *Nature* 2007, **448**(7151):318-324.

55. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R *et al*: **Induced pluripotent stem cell lines derived from human somatic cells**. *Science (New York, NY* 2007, **318**(5858):1917-1920.

56. Waddington CH: **The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology**: Allen & Unwin; 1957.

57. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG *et al*: **Core transcriptional regulatory circuitry in human embryonic stem cells**. *Cell* 2005, **122**(6):947-956.

58. Chew JL, Loh YH, Zhang W, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P *et al*: **Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells**. *Molecular and cellular biology* 2005, **25**(14):6031-6046.

59. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J *et al*: **The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells**. *Nature genetics* 2006, **38**(4):431-440.

60. Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, Robson P: **Transcriptional regulation of nanog by OCT4 and SOX2**. *The Journal of biological chemistry* 2005, **280**(26):24731-24737.

61. Okamoto K, Okazawa H, Okuda A, Sakai M, Muramatsu M, Hamada H: **A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells**. *Cell* 1990, **60**(3):461-472.

62. Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Scholer H, Smith A: **Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4**. *Cell* 1998, **95**(3):379-391.

63. Niwa H, Miyazaki J, Smith AG: **Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells**. *Nature genetics* 2000, **24**(4):372-376.

64. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J *et al*: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells**. *Cell* 2008, **133**(6):1106-1117.

65. Pardo M, Lang B, Yu L, Prosser H, Bradley A, Babu MM, Choudhary J: **An expanded Oct4 interaction network: implications for stem cell biology, development, and disease**. *Cell stem cell* 2010, **6**(4):382-395.

66. van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot RA: **An Oct4-centered protein interaction network in embryonic stem cells**. *Cell stem cell* 2010, **6**(4):369-381.

67. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R: **Multipotent cell lineages in early mouse development depend on SOX2 function**. *Genes & development* 2003, **17**(1):126-140.

68. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P: **Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst**. *Developmental cell* 2010, **18**(4):675-685.

69. Yuan H, Corbi N, Basilico C, Dailey L: **Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3**. *Genes & development* 1995, **9**(21):2635-2645.

70. Nakatake Y, Fukui N, Iwamatsu Y, Masui S, Takahashi K, Yagi R, Yagi K, Miyazaki J, Matoba R, Ko MS *et al*: **Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells**. *Molecular and cellular biology* 2006, **26**(20):7772-7782.

71. Kuroda T, Tada M, Kubota H, Kimura H, Hatano SY, Suemori H, Nakatsuji N, Tada T: **Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression**. *Molecular and cellular biology* 2005, **25**(6):2475-2485.

72. Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, Okochi H, Okuda A, Matoba R, Sharov AA *et al*: **Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells**. *Nature cell biology* 2007, **9**(6):625-635.

73. Lin SC, Wani MA, Whitsett JA, Wells JM: **Klf5 regulates lineage formation in the pre-implantation mouse embryo**. *Development (Cambridge, England)* 2010, **137**(23):3953-3963.

74. Chan KK, Zhang J, Chia NY, Chan YS, Sim HS, Tan KS, Oh SK, Ng HH, Choo AB: **KLF4 and PBX1 directly regulate NANOG expression in human embryonic stem cells**. *Stem cells (Dayton, Ohio)* 2009, **27**(9):2114-2125.

75. Jiang J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH: **A core Klf circuitry regulates self-renewal of embryonic stem cells**. *Nature cell biology* 2008, **10**(3):353-360.

76. Guo G, Yang J, Nichols J, Hall JS, Eyres I, Mansfield W, Smith A: **Klf4 reverts developmentally programmed restriction of ground state pluripotency**. *Development (Cambridge, England)* 2009, **136**(7):1063-1069.

77. Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S: **Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts**. *Nature biotechnology* 2008, **26**(1):101-106.

78. Wernig M, Meissner A, Cassady JP, Jaenisch R: **c-Myc is dispensable for direct reprogramming of mouse fibroblasts**. *Cell stem cell* 2008, **2**(1):10-12.

79. Nakagawa M, Takizawa N, Narita M, Ichisaka T, Yamanaka S: **Promotion of direct reprogramming by transformation-deficient Myc**. *Proceedings of the*

National Academy of Sciences of the United States of America 2010, **107**(32):14152-14157.

80. Feng B, Jiang J, Kraus P, Ng JH, Heng JC, Chan YS, Yaw LP, Zhang W, Loh YH, Han J *et al*: **Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb**. *Nature cell biology* 2009, **11**(2):197-203.
81. Heng JC, Feng B, Han J, Jiang J, Kraus P, Ng JH, Orlov YL, Huss M, Yang L, Lufkin T *et al*: **The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells**. *Cell stem cell* 2010, **6**(2):167-174.
82. Zhao Y, Yin X, Qin H, Zhu F, Liu H, Yang W, Zhang Q, Xiang C, Hou P, Song Z *et al*: **Two supporting factors greatly improve the efficiency of human iPSC generation**. *Cell stem cell* 2008, **3**(5):475-479.
83. Nishimoto M, Fukushima A, Okuda A, Muramatsu M: **The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2**. *Molecular and cellular biology* 1999, **19**(8):5453-5465.
84. Hong H, Takahashi K, Ichisaka T, Aoi T, Kanagawa O, Nakagawa M, Okita K, Yamanaka S: **Suppression of induced pluripotent stem cell generation by the p53-p21 pathway**. *Nature* 2009, **460**(7259):1132-1135.
85. Kawamura T, Suzuki J, Wang YV, Menendez S, Morera LB, Raya A, Wahl GM, Belmonte JC: **Linking the p53 tumour suppressor pathway to somatic cell reprogramming**. *Nature* 2009, **460**(7259):1140-1144.
86. Li H, Collado M, Villasante A, Strati K, Ortega S, Canamero M, Blasco MA, Serrano M: **The Ink4/Arf locus is a barrier for iPS cell reprogramming**. *Nature* 2009, **460**(7259):1136-1139.
87. Marion RM, Strati K, Li H, Murga M, Blanco R, Ortega S, Fernandez-Capetillo O, Serrano M, Blasco MA: **A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity**. *Nature* 2009, **460**(7259):1149-1153.
88. Utikal J, Polo JM, Stadtfeld M, Maherali N, Kulalert W, Walsh RM, Khalil A, Rheinwald JG, Hochedlinger K: **Immortalization eliminates a roadblock during cellular reprogramming into iPS cells**. *Nature* 2009, **460**(7259):1145-1148.
89. Yang CS, Li Z, Rana TM: **microRNAs modulate iPS cell generation**. *RNA (New York, NY* 2011, **17**(8):1451-1460.
90. Choi YJ, Lin CP, Ho JJ, He X, Okada N, Bu P, Zhong Y, Kim SY, Bennett MJ, Chen C *et al*: **miR-34 miRNAs provide a barrier for somatic cell reprogramming**. *Nature cell biology* 2011, **13**(11):1353-1360.
91. Judson RL, Babiarz JE, Venere M, Blelloch R: **Embryonic stem cell-specific microRNAs promote induced pluripotency**. *Nature biotechnology* 2009, **27**(5):459-461.
92. Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, Zhang Y, Yang W, Gruber PJ, Epstein JA *et al*: **Highly efficient miRNA-mediated**

reprogramming of mouse and human somatic cells to pluripotency. *Cell stem cell* 2011, **8**(4):376-388.

93. Card DA, Hebbar PB, Li L, Trotter KW, Komatsu Y, Mishina Y, Archer TK: **Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells**. *Molecular and cellular biology* 2008, **28**(20):6426-6438.
94. Chen S, Do JT, Zhang Q, Yao S, Yan F, Peters EC, Scholer HR, Schultz PG, Ding S: **Self-renewal of embryonic stem cells by a small molecule**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(46):17266-17271.
95. Huangfu D, Maehr R, Guo W, Eijkelenboom A, Snitow M, Chen AE, Melton DA: **Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds**. *Nature biotechnology* 2008, **26**(7):795-797.
96. Shi Y, Desponts C, Do JT, Hahm HS, Scholer HR, Ding S: **Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds**. *Cell stem cell* 2008, **3**(5):568-574.
97. Shi Y, Do JT, Desponts C, Hahm HS, Scholer HR, Ding S: **A combined chemical and genetic approach for the generation of induced pluripotent stem cells**. *Cell stem cell* 2008, **2**(6):525-528.
98. Stadtfeld M, Maherali N, Breault DT, Hochedlinger K: **Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse**. *Cell stem cell* 2008, **2**(3):230-240.
99. Brambrink T, Foreman R, Welstead GG, Lengner CJ, Wernig M, Suh H, Jaenisch R: **Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells**. *Cell stem cell* 2008, **2**(2):151-159.
100. Yamanaka S: **Elite and stochastic models for induced pluripotent stem cell generation**. *Nature* 2009, **460**(7251):49-52.
101. Huangfu D, Osafune K, Maehr R, Guo W, Eijkelenboom A, Chen S, Muhlestein W, Melton DA: **Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2**. *Nature biotechnology* 2008, **26**(11):1269-1275.
102. Hanna J, Markoulaki S, Schorderet P, Carey BW, Beard C, Wernig M, Creyghton MP, Steine EJ, Cassady JP, Foreman R *et al*: **Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency**. *Cell* 2008, **133**(2):250-264.
103. Varas F, Stadtfeld M, de Andres-Aguayo L, Maherali N, di Tullio A, Pantano L, Notredame C, Hochedlinger K, Graf T: **Fibroblast-derived induced pluripotent stem cells show no common retroviral vector insertions**. *Stem cells (Dayton, Ohio)* 2009, **27**(2):300-306.
104. Stadtfeld M, Nagaya M, Utikal J, Weir G, Hochedlinger K: **Induced pluripotent stem cells generated without viral integration**. *Science (New York, NY* 2008, **322**(5903):945-949.

105. Hanna J, Saha K, Pando B, van Zon J, Lengner CJ, Creyghton MP, van Oudenaarden A, Jaenisch R: **Direct cell reprogramming is a stochastic process amenable to acceleration**. *Nature* 2009, **462**(7273):595-601.

106. Chin MH, Mason MJ, Xie W, Volinia S, Singer M, Peterson C, Ambartsumyan G, Aimiuwu O, Richter L, Zhang J *et al*: **Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures**. *Cell stem cell* 2009, **5**(1):111-123.

107. Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, Jaenisch R, Young RA: **Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells**. *Cell stem cell* 2010, **7**(2):249-257.

108. Newman AM, Cooper JB: **Lab-specific gene expression signatures in pluripotent stem cells**. *Cell stem cell* 2010, **7**(2):258-262.

109. Armstrong L, Tilgner K, Saretzki G, Atkinson SP, Stojkovic M, Moreno R, Przyborski S, Lako M: **Human induced pluripotent stem cell lines show stress defense mechanisms and mitochondrial regulation similar to those of human embryonic stem cells**. *Stem cells (Dayton, Ohio)* 2010, **28**(4):661-673.

110. Saretzki G, Armstrong L, Leake A, Lako M, von Zglinicki T: **Stress defense in murine embryonic stem cells is superior to that of various differentiated murine cells**. *Stem cells (Dayton, Ohio)* 2004, **22**(6):962-971.

111. Saretzki G, Walter T, Atkinson S, Passos JF, Bareth B, Keith WN, Stewart R, Hoare S, Stojkovic M, Armstrong L *et al*: **Downregulation of multiple stress defense mechanisms during differentiation of human embryonic stem cells**. *Stem cells (Dayton, Ohio)* 2008, **26**(2):455-464.

112. Polo JM, Liu S, Figueroa ME, Kulalert W, Eminli S, Tan KY, Apostolou E, Stadtfeld M, Li Y, Shioda T *et al*: **Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells**. *Nature biotechnology* 2010, **28**(8):848-855.

113. Kim K, Doi A, Wen B, Ng K, Zhao R, Cahan P, Kim J, Aryee MJ, Ji H, Ehrlich LI *et al*: **Epigenetic memory in induced pluripotent stem cells**. *Nature* 2010, **467**(7313):285-290.

114. Bar-Nur O, Russ HA, Efrat S, Benvenisty N: **Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet Beta cells**. *Cell stem cell* 2011, **9**(1):17-23.

115. 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):643-655.

116. Mizugishi K, Aruga J, Nakata K, Mikoshiba K: **Molecular properties of Zic proteins as transcriptional regulators and their relationship to GLI proteins**. *The Journal of biological chemistry* 2001, **276**(3):2180-2188.

117. Lim LS, Loh YH, Zhang W, Li Y, Chen X, Wang Y, Bakre M, Ng HH, Stanton LW: **Zic3 is required for maintenance of pluripotency in embryonic stem cells**. *Molecular biology of the cell* 2007, **18**(4):1348-1358.

118. Okuda A, Fukushima A, Nishimoto M, Orimo A, Yamagishi T, Nabeshima Y, Kuro-o M, Nabeshima Y, Boon K, Keaveney M *et al*: **UTF1, a novel transcriptional coactivator expressed in pluripotent embryonic stem cells and extra-embryonic cells**. *The EMBO journal* 1998, **17**(7):2019-2032.

119. Cooney AJ, Hummelke GC, Herman T, Chen F, Jackson KJ: **Germ cell nuclear factor is a response element-specific repressor of transcription**. *Biochemical and biophysical research communications* 1998, **245**(1):94-100.

120. Ouyang Z, Zhou Q, Wong WH: **ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(51):21521-21526.

121. Kidder BL, Yang J, Palmer S: **Stat3 and c-Myc genome-wide promoter occupancy in embryonic stem cells**. *PloS one* 2008, **3**(12):e3932.

122. Mathur D, Danford TW, Boyer LA, Young RA, Gifford DK, Jaenisch R: **Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET**. *Genome biology* 2008, **9**(8):R126.

123. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nature genetics* 2005, **37**(4):382-390.

124. Ross JJ, Hong Z, Willenbring B, Zeng L, Isenberg B, Lee EH, Reyes M, Keirstead SA, Weir EK, Tranquillo RT *et al*: **Cytokine-induced differentiation of multipotent adult progenitor cells into functional smooth muscle cells**. *The Journal of clinical investigation* 2006, **116**(12):3139-3149.

125. Roelandt P, Pauwelyn KA, Sancho-Bru P, Subramanian K, Ordovas L, Vanuytsel K, Geraerts M, Firpo M, De Vos R, Fevery J *et al*: **Human embryonic and rat adult stem cells with primitive endoderm-like phenotype can be fated to definitive endoderm, and finally hepatocyte-like cells**. *PloS one* 2010, **5**(8):e12101.

126. Ulloa-Montoya F, Kidder BL, Pauwelyn KA, Chase LG, Luttun A, Crabbe A, Geraerts M, Sharov AA, Piao Y, Ko MS *et al*: **Comparative transcriptome analysis of embryonic and adult stem cells with extended and limited differentiation capacity**. *Genome biology* 2007, **8**(8):R163.

127. Kunath T, Arnaud D, Uy GD, Okamoto I, Chureau C, Yamanaka Y, Heard E, Gardner RL, Avner P, Rossant J: **Imprinted X-inactivation in extra-embryonic endoderm cell lines from mouse blastocysts**. *Development (Cambridge, England)* 2005, **132**(7):1649-1661.

128. Lengner CJ, Camargo FD, Hochedlinger K, Welstead GG, Zaidi S, Gokhale S, Scholer HR, Tomilin A, Jaenisch R: **Oct4 expression is not required for mouse somatic stem cell self-renewal**. *Cell stem cell* 2007, **1**(4):403-415.

129. Ralston A, Rossant J: **Genetic regulation of stem cell origins in the mouse embryo**. *Clinical genetics* 2005, **68**(2):106-112.

130. 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):631-642.

131. Fujikura J, Yamato E, Yonemura S, Hosoda K, Masui S, Nakao K, Miyazaki Ji J, Niwa H: **Differentiation of embryonic stem cells is induced by GATA factors**. *Genes & development* 2002, **16**(7):784-789.
132. Arman E, Haffner-Krausz R, Chen Y, Heath JK, Lonai P: **Targeted disruption of fibroblast growth factor (FGF) receptor 2 suggests a role for FGF signaling in pregastrulation mammalian development**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(9):5082-5087.
133. Feldman B, Poueymirou W, Papaioannou VE, DeChiara TM, Goldfarb M: **Requirement of FGF-4 for postimplantation mouse development**. *Science (New York, NY* 1995, **267**(5195):246-249.
134. Chazaud C, Yamanaka Y, Pawson T, Rossant J: **Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway**. *Developmental cell* 2006, **10**(5):615-624.
135. Tanaka S, Kunath T, Hadjantonakis AK, Nagy A, Rossant J: **Promotion of trophoblast stem cell proliferation by FGF4**. *Science (New York, NY* 1998, **282**(5396):2072-2075.
136. Kanungo J, Potapova I, Malbon CC, Wang H: **MEKK4 mediates differentiation in response to retinoic acid via activation of c-Jun N-terminal kinase in rat embryonal carcinoma P19 cells**. *The Journal of biological chemistry* 2000, **275**(31):24032-24039.
137. Fowler KJ, Mitrangas K, Dziadek M: **In vitro production of Reichert's membrane by mouse embryo-derived parietal endoderm cell lines**. *Experimental cell research* 1990, **191**(2):194-203.
138. Notarianni E, Flechon J: **Parietal endoderm cell line from a rat blastocyst**. *Placenta* 2001, **22**(1):111-123.
139. Galat V, Binas B, Iannaccone S, Postovit LM, Debeb BG, Iannaccone P: **Developmental potential of rat extraembryonic stem cells**. *Stem cells and development* 2009, **18**(9):1309-1318.
140. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al*: **Expression monitoring by hybridization to high-density oligonucleotide arrays**. *Nature biotechnology* 1996, **14**(13):1675-1680.
141. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science (New York, NY* 1995, **270**(5235):467-470.
142. Bridge PD, Sawilowsky SS: **Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and and Wilcoxon Rank-Sum test in small samples applied research**. *Journal of clinical epidemiology* 1999, **52**(3):229-235.
143. Dudoit S, Shaffer PJ, Boldrick JC: **Multiple hypothesis testing in microarray experiments**. *Statistical Science* 2003, **18**(1):71-103.

144. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(9):5116-5121.

145. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures**. *Bioinformatics (Oxford, England)* 2003, **19**(3):368-375.

146. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.

147. Jolliffe IT: **Principal component analysis**, Second edn: Springer; 2002.

148. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data**. *Bioinformatics (Oxford, England)* 2001, **17**(9):763-774.

149. Liu W, Yuan K, Ye D: **Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis**. *Journal of biomedical informatics* 2008, **41**(4):602-606.

150. Lee DD, Seung HS: **Learning the parts of objects by non-negative matrix factorization**. *Nature* 1999, **401**(6755):788-791.

151. Brunet JP, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(12):4164-4169.

152. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. *BMC bioinformatics* 2006, **7 Suppl 1**:S7.

153. Morita S, Kojima T, Kitamura T: **Plat-E: an efficient and stable system for transient packaging of retroviruses**. *Gene therapy* 2000, **7**(12):1063-1066.

154. Treff NR, Vincent RK, Budde ML, Browning VL, Magliocca JF, Kapur V, Odorico JS: **Differentiation of embryonic stem cells conditionally expressing neurogenin 3**. *Stem cells (Dayton, Ohio)* 2006, **24**(11):2529-2537.

155. Nikolova-Krstevski V, Bhasin M, Otu HH, Libermann T, Oettgen P: **Gene expression analysis of embryonic stem cells expressing VE-cadherin (CD144) during endothelial differentiation**. *BMC genomics* 2008, **9**:240.

156. Kim JB, Zaehres H, Wu G, Gentile L, Ko K, Sebastiano V, Arauzo-Bravo MJ, Ruau D, Han DW, Zenke M *et al*: **Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors**. *Nature* 2008, **454**(7204):646-650.

157. Sampath P, Pritchard DK, Pabon L, Reinecke H, Schwartz SM, Morris DR, Murry CE: **A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation**. *Cell stem cell* 2008, **2**(5):448-460.

158. Orford K, Kharchenko P, Lai W, Dao MC, Worhunsky DJ, Ferro A, Janzen V, Park PJ, Scadden DT: **Differential H3K4 methylation identifies**

developmentally poised hematopoietic genes. *Developmental cell* 2008, **14**(5):798-809.

159. Endoh M, Endo TA, Endoh T, Fujimura Y, Ohara O, Toyoda T, Otte AP, Okano M, Brockdorff N, Vidal M *et al*: **Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity**. *Development (Cambridge, England)* 2008, **135**(8):1513-1524.

160. Langton S, Gudas LJ: **CYP26A1 knockout embryonic stem cells exhibit reduced differentiation and growth arrest in response to retinoic acid**. *Developmental biology* 2008, **315**(2):331-354.

161. Sinkkonen L, Hugenschmidt T, Berninger P, Gaidatzis D, Mohn F, Artus-Revel CG, Zavolan M, Svoboda P, Filipowicz W: **MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells**. *Nature structural & molecular biology* 2008, **15**(3):259-267.

162. Lorincz MT, Zawistowski VA: **Expanded CAG repeats in the murine Huntington's disease gene increases neuronal differentiation of embryonic and neural stem cells**. *Molecular and cellular neurosciences* 2009, **40**(1):1-13.

163. Christoforou N, Miller RA, Hill CM, Jie CC, McCallion AS, Gearhart JD: **Mouse ES cell-derived cardiac precursor cells are multipotent and facilitate identification of novel cardiac genes**. *The Journal of clinical investigation* 2008, **118**(3):894-903.

164. Nord AS, Vranizan K, Tingley W, Zambon AC, Hanspers K, Fong LG, Hu Y, Bacchetti P, Ferrin TE, Babbitt PC *et al*: **Modeling insertional mutagenesis using gene length and expression in murine embryonic stem cells**. *PloS one* 2007, **2**(7):e617.

165. Barrera LO, Li Z, Smith AD, Arden KC, Cavenee WK, Zhang MQ, Green RD, Ren B: **Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs**. *Genome research* 2008, **18**(1):46-59.

166. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A: **Dissecting direct reprogramming through integrative genomic analysis**. *Nature* 2008, **454**(7200):49-55.

167. Lu SJ, Hipp JA, Feng Q, Hipp JD, Lanza R, Atala A: **GeneChip analysis of human embryonic stem cell differentiation into hemangioblasts: an in silico dissection of mixed phenotypes**. *Genome biology* 2007, **8**(11):R240.

168. Lowry WE, Richter L, Yachechko R, Pyle AD, Tchieu J, Sridharan R, Clark AT, Plath K: **Generation of human induced pluripotent stem cells from dermal fibroblasts**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(8):2883-2888.

169. Baker DE, Harrison NJ, Maltby E, Smith K, Moore HD, Shaw PJ, Heath PR, Holden H, Andrews PW: **Adaptation to culture of human embryonic stem cells and oncogenesis in vivo**. *Nature biotechnology* 2007, **25**(2):207-215.

170. Park IH, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, Daley GQ: **Reprogramming of human somatic cells to pluripotency with defined factors**. *Nature* 2008, **451**(7175):141-146.

171. Masaki H, Ishikawa T, Takahashi S, Okumura M, Sakai N, Haga M, Kominami K, Migita H, McDonald F, Shimada F *et al*: **Heterogeneity of pluripotent marker gene expression in colonies generated in human iPS cell induction culture**. *Stem cell research* 2007, **1**(2):105-115.

172. Avery K, Avery S, Shepherd J, Heath PR, Moore H: **Sphingosine-1-phosphate mediates transcriptional regulation of key targets associated with survival, proliferation, and pluripotency in human embryonic stem cells**. *Stem cells and development* 2008, **17**(6):1195-1205.

173. Maherali N, Ahfeldt T, Rigamonti A, Utikal J, Cowan C, Hochedlinger K: **A high-efficiency system for the generation and study of human induced pluripotent stem cells**. *Cell stem cell* 2008, **3**(3):340-345.

174. Kocabas AM, Crosby J, Ross PJ, Otu HH, Beyhan Z, Can H, Tam WL, Rosa GJ, Halgren RG, Lim B *et al*: **The transcriptome of human oocytes**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(38):14027-14032.

175. Seguin CA, Draper JS, Nagy A, Rossant J: **Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells**. *Cell stem cell* 2008, **3**(2):182-195.

176. Conrad S, Renninger M, Hennenlotter J, Wiesner T, Just L, Bonin M, Aicher W, Buhring HJ, Mattheus U, Mack A *et al*: **Generation of pluripotent stem cells from adult human testis**. *Nature* 2008, **456**(7220):344-349.

177. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data**. *Genome research* 2003, **13**(7):1706-1718.

178. Mahley RW: **Apolipoprotein E: cholesterol transport protein with expanding role in cell biology**. *Science (New York, NY* 1988, **240**(4852):622-630.

179. Vogel T, Guo NH, Guy R, Drezlich N, Krutzsch HC, Blake DA, Panet A, Roberts DD: **Apolipoprotein E: a potent inhibitor of endothelial and tumor cell proliferation**. *Journal of cellular biochemistry* 1994, **54**(3):299-308.

180. Bachner D, Schroder D, Betat N, Ahrens M, Gross G: **Apolipoprotein E (ApoE), a Bmp-2 (bone morphogenetic protein) upregulated gene in mesenchymal progenitors (C3H10T1/2), is highly expressed in murine embryonic development**. *BioFactors (Oxford, England)* 1999, **9**(1):11-17.

181. Babin PJ, Thisse C, Durliat M, Andre M, Akimenko MA, Thisse B: **Both apolipoprotein E and A-I genes are present in a nonmammalian vertebrate and are highly expressed during embryonic development**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(16):8622-8627.

182. Piedrahita JA, Zhang SH, Hagaman JR, Oliver PM, Maeda N: **Generation of mice carrying a mutant apolipoprotein E gene inactivated by gene targeting in embryonic stem cells**. *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(10):4471-4475.

183. Quertermous EE, Hidai H, Blanar MA, Quertermous T: **Cloning and characterization of a basic helix-loop-helix protein expressed in early mesoderm and the developing somites**. *Proceedings of the National Academy of Sciences of the United States of America* 1994, **91**(15):7066-7070.

184. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks**. *Bioinformatics (Oxford, England)* 2002, **18 Suppl 1**:S233-240.

185. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic acids research* 2004, **32**(Database issue):D449-451.

186. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update**. *Nucleic acids research* 2005, **33**(Database issue):D418-424.

187. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM *et al*: **The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology**. *Nucleic acids research* 2005, **33**(Database issue):D471-475.

188. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics (Oxford, England)* 2005, **21**(16):3439-3440.

189. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs**. *Nucleic acids research* 2005, **33**(Database issue):D476-480.

190. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, Troyanskaya OG: **A genomewide functional network for the laboratory mouse**. *PLoS computational biology* 2008, **4**(9):e1000165.

191. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG: **Discovery of biological networks from diverse functional genomic data**. *Genome biology* 2005, **6**(13):R114.

192. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, Troyanskaya OG: **Exploring the human genome with functional maps**. *Genome research* 2009, **19**(6):1093-1106.

193. Rappolee DA, Basilico C, Patel Y, Werb Z: **Expression and function of FGF-4 in peri-implantation development in mouse embryos**. *Development (Cambridge, England)* 1994, **120**(8):2259-2269.

194. Kunath T, Saba-El-Leil MK, Almousailleakh M, Wray J, Meloche S, Smith A: **FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment**. *Development (Cambridge, England)* 2007, **134**(16):2895-2902.

195. Xu RH, Sampsell-Barron TL, Gu F, Root S, Peck RM, Pan G, Yu J, Antosiewicz-Bourget J, Tian S, Stewart R *et al*: **NANOG is a direct target of**

TGFbeta/activin-mediated SMAD signaling in human ESCs. *Cell stem cell* 2008, **3**(2):196-206.

196. Xu RH, Peck RM, Li DS, Feng X, Ludwig T, Thomson JA: **Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells**. *Nature methods* 2005, **2**(3):185-190.

197. Faast R, White J, Cartwright P, Crocker L, Sarcevic B, Dalton S: **Cdk6-cyclin D3 activity in murine ES cells is resistant to inhibition by p16(INK4a)**. *Oncogene* 2004, **23**(2):491-502.

198. Neganova I, Zhang X, Atkinson S, Lako M: **Expression and functional analysis of G1 to S regulatory components reveals an important role for CDK2 in cell cycle regulation in human embryonic stem cells**. *Oncogene* 2009, **28**(1):20-30.

199. Fujii-Yamamoto H, Kim JM, Arai K, Masai H: **Cell cycle and developmental regulations of replication factors in mouse embryonic stem cells**. *The Journal of biological chemistry* 2005, **280**(13):12976-12987.

200. Tarasov KV, Tarasova YS, Tam WL, Riordon DR, Elliott ST, Kania G, Li J, Yamanaka S, Crider DG, Testa G *et al*: **B-MYB is essential for normal cell cycle progression and chromosomal stability of embryonic stem cells**. *PloS one* 2008, **3**(6):e2478.

201. Tanaka Y, Patestos NP, Maekawa T, Ishii S: **B-myb is required for inner cell mass formation at an early stage of development**. *The Journal of biological chemistry* 1999, **274**(40):28067-28070.

202. Kong M, Barnes EA, Ollendorff V, Donoghue DJ: **Cyclin F regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction**. *The EMBO journal* 2000, **19**(6):1378-1388.

203. Heo JS, Lee MY, Han HJ: **Sonic hedgehog stimulates mouse embryonic stem cell proliferation by cooperation of Ca2+/protein kinase C and epidermal growth factor receptor as well as Gli1 activation**. *Stem cells (Dayton, Ohio)* 2007, **25**(12):3069-3080.

204. Yang X, Meng X, Su X, Mauchley DC, Ao L, Cleveland JC, Jr., Fullerton DA: **Bone morphogenic protein 2 induces Runx2 and osteopontin expression in human aortic valve interstitial cells: role of Smad1 and extracellular signal-regulated kinase 1/2**. *The Journal of thoracic and cardiovascular surgery* 2009, **138**(4):1008-1015.

205. Wang T, Tamakoshi T, Uezato T, Shu F, Kanzaki-Kato N, Fu Y, Koseki H, Yoshida N, Sugiyama T, Miura N: **Forkhead transcription factor Foxf2 (LUN)-deficient mice exhibit abnormal development of secondary palate**. *Developmental biology* 2003, **259**(1):83-94.

206. Chatzi C, van den Brink CE, van der Saag PT, McCaig CD, Shen S: **Expression of a mutant retinoic acid receptor beta alters lineage differentiation in mouse embryonic stem cells**. *Stem cells and development*, **19**(7):951-960.

207. Hannenhalli S, Kaestner KH: **The evolution of Fox genes and their role in development and disease**. *Nature reviews* 2009, **10**(4):233-240.

208. Miyashita T, Hanashita T, Toriyama M, Takagi R, Akashika T, Higashikubo N: **Gene cloning and biochemical characterization of the BMP-2 of Pinctada fucata**. *Bioscience, biotechnology, and biochemistry* 2008, **72**(1):37-47.

209. Mullor JL, Dahmane N, Sun T, Ruiz i Altaba A: **Wnt signals are targets and mediators of Gli function**. *Curr Biol* 2001, **11**(10):769-773.

210. Parsons SJ, Parsons JT: **Src family kinases, key regulators of signal transduction**. *Oncogene* 2004, **23**(48):7906-7909.

211. Anneren C, Cowan CA, Melton DA: **The Src family of tyrosine kinases is important for embryonic stem cell self-renewal**. *The Journal of biological chemistry* 2004, **279**(30):31590-31598.

212. Sun Y, Li H, Liu Y, Mattson MP, Rao MS, Zhan M: **Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation**. *PloS one* 2008, **3**(10):e3406.

213. Perlingeiro RC: **Endoglin is required for hemangioblast and early hematopoietic development**. *Development (Cambridge, England)* 2007, **134**(16):3041-3048.

214. Wallin R, Cain D, Hutson SM, Sane DC, Loeser R: **Modulation of the binding of matrix Gla protein (MGP) to bone morphogenetic protein-2 (BMP-2)**. *Thrombosis and haemostasis* 2000, **84**(6):1039-1044.

215. Niwa H, Burdon T, Chambers I, Smith A: **Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3**. *Genes & development* 1998, **12**(13):2048-2060.

216. Yang J, van Oosten AL, Theunissen TW, Guo G, Silva JC, Smith A: **Stat3 activation is limiting for reprogramming to ground state pluripotency**. *Cell stem cell* 2010, **7**(3):319-328.

217. Cai KQ, Capo-Chichi CD, Rula ME, Yang DH, Xu XX: **Dynamic GATA6 expression in primitive endoderm formation and maturation in early mouse embryogenesis**. *Dev Dyn* 2008, **237**(10):2820-2829.

218. Koutsourakis M, Langeveld A, Patient R, Beddington R, Grosveld F: **The transcription factor GATA6 is essential for early extraembryonic development**. *Development (Cambridge, England)* 1999, **126**(4):723-732.

219. Williams RL, Hilton DJ, Pease S, Willson TA, Stewart CL, Gearing DP, Wagner EF, Metcalf D, Nicola NA, Gough NM: **Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells**. *Nature* 1988, **336**(6200):684-687.

220. Auernhammer CJ, Melmed S: **Leukemia-inhibitory factor-neuroimmune modulator of endocrine function**. *Endocrine reviews* 2000, **21**(3):313-345.

221. Stahl N, Farruggella TJ, Boulton TG, Zhong Z, Darnell JE, Jr., Yancopoulos GD: **Choice of STATs and other substrates specified by modular tyrosine-based motifs in cytokine receptors**. *Science (New York, NY* 1995, **267**(5202):1349-1353.

222. Yoshida-Koide U, Matsuda T, Saikawa K, Nakanuma Y, Yokota T, Asashima M, Koide H: **Involvement of Ras in extraembryonic endoderm differentiation of**

embryonic stem cells. *Biochemical and biophysical research communications* 2004, **313**(3):475-481.

223. Boeuf H, Hauss C, Graeve FD, Baran N, Kedinger C: **Leukemia inhibitory factor-dependent transcriptional activation in embryonic stem cells**. *The Journal of cell biology* 1997, **138**(6):1207-1217.

224. Niwa H, Ogawa K, Shimosato D, Adachi K: **A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells**. *Nature* 2009, **460**(7251):118-122.

225. Alexander WS, Starr R, Metcalf D, Nicholson SE, Farley A, Elefanty AG, Brysha M, Kile BT, Richardson R, Baca M *et al*: **Suppressors of cytokine signaling (SOCS): negative regulators of signal transduction**. *Journal of leukocyte biology* 1999, **66**(4):588-592.

226. Nicholson SE, Willson TA, Farley A, Starr R, Zhang JG, Baca M, Alexander WS, Metcalf D, Hilton DJ, Nicola NA: **Mutational analyses of the SOCS proteins suggest a dual domain requirement but distinct mechanisms for inhibition of LIF and IL-6 signal transduction**. *The EMBO journal* 1999, **18**(2):375-385.

227. Starr R, Hilton DJ: **Negative regulation of the JAK/STAT pathway**. *Bioessays* 1999, **21**(1):47-52.

228. Forrai A, Boyle K, Hart AH, Hartley L, Rakar S, Willson TA, Simpson KM, Roberts AW, Alexander WS, Voss AK *et al*: **Absence of suppressor of cytokine signalling 3 reduces self-renewal and promotes differentiation in murine embryonic stem cells**. *Stem cells (Dayton, Ohio)* 2006, **24**(3):604-614.

229. Duval D, Reinhardt B, Kedinger C, Boeuf H: **Role of suppressors of cytokine signaling (Socs) in leukemia inhibitory factor (LIF) -dependent embryonic stem cell survival**. *Faseb J* 2000, **14**(11):1577-1584.

230. Bousquet C, Susini C, Melmed S: **Inhibitory roles for SHP-1 and SOCS-3 following pituitary proopiomelanocortin induction by leukemia inhibitory factor**. *The Journal of clinical investigation* 1999, **104**(9):1277-1285.

231. Boyle K, Zhang JG, Nicholson SE, Trounson E, Babon JJ, McManus EJ, Nicola NA, Robb L: **Deletion of the SOCS box of suppressor of cytokine signaling 3 (SOCS3) in embryonic stem cells reveals SOCS box-dependent regulation of JAK but not STAT phosphorylation**. *Cellular signalling* 2009, **21**(3):394-404.

232. Dawson MA, Bannister AJ, Gottgens B, Foster SD, Bartke T, Green AR, Kouzarides T: **JAK2 phosphorylates histone H3Y41 and excludes HP1alpha from chromatin**. *Nature* 2009, **461**(7265):819-822.

233. Griffiths DS, Li J, Dawson MA, Trotter MW, Cheng YH, Smith AM, Mansfield W, Liu P, Kouzarides T, Nichols J *et al*: **LIF-independent JAK signalling to chromatin in embryonic stem cells uncovered from an adult stem cell disease**. *Nature cell biology* 2010, **13**(1):13-21.

234. Plusa B, Piliszek A, Frankenberg S, Artus J, Hadjantonakis AK: **Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst**. *Development (Cambridge, England)* 2008, **135**(18):3081-3091.

235. Artus J, Panthier JJ, Hadjantonakis AK: **A role for PDGF signaling in expansion of the extra-embryonic endoderm lineage of the mouse blastocyst**. *Development (Cambridge, England)* 2010, **137**(20):3361-3372.

236. Gu P, LeMenuet D, Chung AC, Mancini M, Wheeler DA, Cooney AJ: **Orphan nuclear receptor GCNF is required for the repression of pluripotency genes during retinoic acid-induced embryonic stem cell differentiation**. *Molecular and cellular biology* 2005, **25**(19):8507-8519.

237. Herman JG, Latif F, Weng Y, Lerman MI, Zbar B, Liu S, Samid D, Duan DS, Gnarra JR, Linehan WM *et al*: **Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma**. *Proceedings of the National Academy of Sciences of the United States of America* 1994, **91**(21):9700-9704.

238. Li E, Beard C, Jaenisch R: **Role for DNA methylation in genomic imprinting**. *Nature* 1993, **366**(6453):362-365.

239. Lagarkova MA, Volchkov PY, Lyakisheva AV, Philonenko ES, Kiselev SL: **Diverse epigenetic profile of novel human embryonic stem cell lines**. *Cell cycle (Georgetown, Tex* 2006, **5**(4):416-420.

240. Bibikova M, Laurent LC, Ren B, Loring JF, Fan JB: **Unraveling epigenetic regulation in embryonic stem cells**. *Cell stem cell* 2008, **2**(2):123-134.

241. Kim JB, Sebastiano V, Wu G, Arauzo-Bravo MJ, Sasse P, Gentile L, Ko K, Ruau D, Ehrich M, van den Boom D *et al*: **Oct4-induced pluripotency in adult neural stem cells**. *Cell* 2009, **136**(3):411-419.

242. Kuckenberg P, Peitz M, Kubaczka C, Becker A, Egert A, Wardelmann E, Zimmer A, Brustle O, Schorle H: **Lineage conversion of murine extraembryonic trophoblast stem cells to pluripotent stem cells**. *Molecular and cellular biology* 2011, **31**(8):1748-1756.

243. Hotta A, Cheung AY, Farra N, Vijayaragavan K, Seguin CA, Draper JS, Pasceri P, Maksakova IA, Mager DL, Rossant J *et al*: **Isolation of human iPS cells using EOS lentiviral vectors to select for pluripotency**. *Nature methods* 2009, **6**(5):370-376.

244. Li W, Ding S: **Generation of novel rat and human pluripotent stem cells by reprogramming and chemical approaches**. *Methods in molecular biology (Clifton, NJ* 2010, **636**:293-300.

245. Li W, Wei W, Zhu S, Zhu J, Shi Y, Lin T, Hao E, Hayek A, Deng H, Ding S: **Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors**. *Cell stem cell* 2009, **4**(1):16-19.

246. Chang MY, Kim D, Kim CH, Kang HC, Yang E, Moon JI, Ko S, Park J, Park KS, Lee KA *et al*: **Direct reprogramming of rat neural precursor cells and fibroblasts into pluripotent stem cells**. *PloS one* 2010, **5**(3):e9838.

247. Carey BW, Markoulaki S, Hanna J, Saha K, Gao Q, Mitalipova M, Jaenisch R: **Reprogramming of murine and human somatic cells using a single polycistronic vector**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(1):157-162.

248. Diao Y, Wang X, Wu Z: **SOCS1, SOCS3, and PIAS1 promote myogenic differentiation by inhibiting the leukemia inhibitory factor-induced JAK1/STAT1/STAT3 pathway**. *Molecular and cellular biology* 2009, **29**(18):5084-5093.

Appendix

	Probe Set ID	Gene Symbol	W1
1	1448595_a_at	Bex1	2.131
2	1456511_x_at	Eras	1.801
3	1417837_at	Phlda2	1.697
4	1428209_at	Bex4	1.678
5	1437052_s_at	Slc2a3	1.670
6	1450989_at	Tdgf1	1.643
7	1437752_at	Lin28	1.620
8	1423424_at	Zic3	1.614
9	1429388_at	Nanog	1.608
10	1418362_at	Zfp42	1.601
11	1417760_at	Nr0b1	1.593
12	1456242_at	EG653016	1.566
13	1416552_at	Dppa5a	1.559
14	1420085_at	Fgf4	1.558
15	1454159_a_at	Igfbp2	1.554
16	1437015_x_at	Pla2g1b	1.547
17	1448269_a_at	Klhl13	1.544
18	1416967_at	Sox2	1.543
19	1423429_at	Rhox5	1.542
20	1422943_a_at	Hspb1	1.541
21	1448752_at	Car2	1.522
22	1453219_a_at	Tdrd12	1.508
23	1419418_a_at	Morc1	1.500
24	1423281_at	Stmn2	1.498
25	1425035_s_at	Dnmt3l	1.495
26	1451230_a_at	Wbp5	1.492
27	1438237_at	Rex2	1.490
28	1438018_at	Hook1	1.481
29	1423212_at	Phc1	1.477
30	1417638_at	Lefty1	1.475
31	1418756_at	Trh	1.463
32	1429759_at	Rps6ka6	1.459
33	1422058_at	Nodal	1.459
34	1460226_at	Trap1a	1.458
35	1435783_at	B230112C05Rik	1.455
36	1417945_at	Pou5f1	1.453
37	1417155_at	Mycn	1.449
38	1424254_at	Ifitm1	1.441
39	1423582_at	Dmrt1	1.425
40	1421882_a_at	Elavl2	1.423
41	1449064_at	Tdh	1.420

42	1436837_at	Mael	1.415
43	1433845_x_at	Dusp9	1.399
44	1432466_a_at	Apoe	1.392
45	1438009_at	Hist1h2ab/// RP23-480B19.10	1.391
46	1453223_s_at	Dppa2	1.388
47	1436227_at	Lefty2	1.387
48	1427242_at	Ddx4	1.386
49	1457314_at	L1td1	1.384
50	1455235_x_at	Ldhb	1.383
51	1429343_at	Rbmxl2	1.375
52	1424531_a_at	Tcea3	1.372
53	1456515_s_at	Tcfl5	1.370
54	1436990_s_at	Ndg2	1.365
55	1448991_a_at	Ina	1.363
56	1424067_at	Icam1	1.362
57	1427238_at	Fbxo15	1.359
58	1452679_at	Tubb2b	1.342
59	1452730_at	Rps4y2	1.342
60	1452787_a_at	Prmt1	1.336
61	1417388_at	Bex2	1.335
62	1427170_at	Psma8	1.327
63	1438820_at	Rnf17	1.327
64	1428572_at	Basp1	1.321
65	1453599_at	Trim71	1.320
66	1428301_at	ENSMUSG00000063277	1.317
67	1425926_a_at	Otx2	1.316
68	1451285_at	Fus	1.315
69	1448890_at	Klf2	1.314
70	1417392_a_at	Slc7a7	1.312
71	1459211_at	Gli2	1.310
72	1454215_at	2410007B07Rik	1.308
73	1428842_a_at	Ngfrap1	1.305
74	1450626_at	Manba	1.302
75	1424152_at	Sall4	1.297
76	1452004_at	Calca	1.295
77	1454114_a_at	Nhedc1	1.295
78	1426955_at	Col18a1	1.295
79	1433658_x_at	Pcbp4	1.294
80	1416041_at	Sgk1	1.290
81	1439065_x_at	OTTMUSG00000010173	1.288
82	1423516_a_at	Nid2	1.286
83	1449592_at	Tcf15	1.284
84	1426722_at	Slc38a2	1.284
85	1452731_x_at	B930046C15Rik	1.283
86	1416899_at	Utf1	1.283
87	1417428_at	Gng3	1.281
88	1443844_at	1700123J19Rik	1.280

89	1449799_s_at	Pkp2	1.280
90	1415996_at	Txnip	1.280
91	1448690_at	Kcnk1	1.273
92	1457026_at	Liph	1.272
93	1437165_a_at	Pcolce	1.266
94	1419542_at	Dazl	1.264
95	1419153_at	2810417H13Rik	1.262
96	1434279_at	---	1.261
97	1429366_at	Lrrc34	1.261
98	1428942_at	Mt2	1.260
99	1445226_at	BC023969	1.259

Table 4. Top 100 genes in the metagene associated with pluripotent mouse stem cells as predicted by NMF

	PROBE SET ID	GENE SYMBOL	W1
1	231381_at	HESRG	1.727
2	220184_at	NANOG	1.719
3	1559280_a_at	---	1.643
4	214240_at	GAL	1.642
5	216405_at	---	1.640
6	206286_s_at	TDGF1	1.573
7	214022_s_at	IFITM1	1.548
8	216379_x_at	CD24	1.548
9	214023_x_at	TUBB2B	1.541
10	231061_at	---	1.524
11	204285_s_at	PMAIP1	1.523
12	204141_at	TUBB2A	1.502
13	207714_s_at	SERPINH1	1.495
14	222392_x_at	PERP	1.495
15	206268_at	LEFTY1	1.486
16	210905_x_at	POU5F1P4	1.479
17	222154_s_at	LOC26010	1.456
18	202790_at	CLDN7	1.456
19	219955_at	L1TD1	1.454
20	217728_at	S100A6	1.454
21	231628_s_at	---	1.450
22	232985_s_at	DPPA4	1.447
23	231896_s_at	DENR	1.440
24	204409_s_at	EIF1AY	1.431
25	201650_at	KRT19	1.428
26	206424_at	CYP26A1	1.426

27	214532_x_at	POU5F1	1.421
28	203706_s_at	FZD7	1.420
29	201909_at	RPS4Y1	1.419
30	203449_s_at	TERF1	1.418
31	218162_at	OLFML3	1.418
32	203404_at	ARMCX2	1.417
33	225660_at	SEMA6A	1.417
34	205547_s_at	TAGLN	1.415
35	209101_at	CTGF	1.412
36	202499_s_at	SLC2A3	1.410
37	212533_at	WEE1	1.407
38	203313_s_at	TGIF1	1.406
39	243161_x_at	ZFP42	1.397
40	210265_x_at	POU5F1P3	1.396
41	203632_s_at	GPRC5B	1.391
42	223062_s_at	PSAT1	1.383
43	203381_s_at	APOE	1.379
44	201596_x_at	KRT18	1.374
45	212203_x_at	IFITM3	1.370
46	1553874_a_at	ZSCAN10	1.369
47	205000_at	DDX3Y	1.363
48	235343_at	---	1.360
49	224597_at	LOC647979	1.360
50	204342_at	SLC25A24	1.354
51	210466_s_at	SERBP1	1.352
52	204009_s_at	KRAS	1.351
53	212097_at	CAV1	1.351
54	230788_at	GCNT2	1.348
55	208700_s_at	TKT	1.348
56	226350_at	CHML	1.347
57	203109_at	UBE2M	1.347
58	201058_s_at	MYL9	1.345
59	203453_at	SCNN1A	1.337
60	204083_s_at	TPM2	1.335
61	200704_at	LITAF	1.331
62	231407_s_at	FOXH1	1.329
63	212077_at	CALD1	1.329
64	222685_at	FAM29A	1.329
65	226587_at	---	1.328
66	210852_s_at	AASS	1.325
67	209008_x_at	KRT8	1.324
68	214363_s_at	MATR3	1.323
69	209875_s_at	SPP1	1.321

70	201578_at	PODXL	1.318
71	209262_s_at	NR2F6	1.317
72	204237_at	GULP1	1.314
73	237193_s_at	---	1.313
74	206291_at	NTS	1.312
75	213564_x_at	LDHB	1.312
76	200754_x_at	SFRS2	1.311
77	224582_s_at	---	1.311
78	224581_s_at	---	1.310
79	200824_at	GSTP1	1.308
80	1568609_s_at	LOC727820	1.307
81	237810_at	CLDN6	1.306
82	224654_at	DDX21	1.306
83	218856_at	TNFRSF21	1.306
84	209191_at	TUBB6	1.299
85	224579_at	SLC38A1	1.297
86	201215_at	PLS3	1.291
87	208712_at	CCND1	1.291
88	200853_at	H2AFZ	1.291
89	209357_at	CITED2	1.290
90	226265_at	QSER1	1.286
91	218156_s_at	TSR1	1.285
92	226884_at	LRRN1	1.285
93	200844_s_at	PRDX6	1.284
94	209757_s_at	MYCN	1.282
95	209267_s_at	SLC39A8	1.281
96	224932_at	C22orf16	1.281
97	200697_at	HK1	1.281
98	200807_s_at	HSPD1	1.280
99	201426_s_at	VIM	1.279
100	213506_at	F2RL1	1.275

Table 5. Top 100 genes in the metagene associated with human pluripotent stem cells as predicted by NMF